Bernoulli **5**(3), 1999, 381–412

Observed information in semi-parametric models

SUSAN A. MURPHY[1] and AAD W. VAN DER VAART[2]

[1]*Department of Statistics, 440 Mason Hall, University of Michigan, Ann Arbor, MI 48109-1027, USA. e-mail*: *samurphy@umich.edu*
[2]*Department of Mathematics, Free University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, Netherlands. e-mail*: *aad@cs.vu.nl*

We discuss the estimation of the asymptotic covariance matrix of semi-parametric maximum likelihood estimators by the observed profile information. We show that a discretized version of the second derivative of the profile likelihood function yields consistent estimators of minus the efficient information matrix.

Keywords: least favourable submodel; profile likelihood; standard errors

1. Introduction

In many semi-parametric models, 'regular' parameters can be estimated by (semi-parametric) maximum likelihood estimators. The asymptotic theory for such estimators has been developed for a number of models of practical interest, and is similar to the asymptotic theory for maximum likelihood estimators in classical parametric models. In particular, the maximum likelihood estimators are asymptotically normal, where the inverse of the 'efficient Fisher information matrix' gives the asymptotic covariance matrix. The latter matrix is the Fisher information matrix corrected for the presence of an infinite-dimensional nuisance parameter. See, for example, Bickel *et al.* (1993) for an extensive review of information bounds. See Gill (1989), Chang (1990), Gu and Zhang (1993), Qin (1993), van der Laan (1993), Qin and Lawless (1994), van der Vaart (1994a; 1994b; 1994c; 1996), Murphy (1995), Gill *et al*. (1995), Huang (1996), Parner (1998), Qin and Wong (1996) and Mammen and van de Geer (1997) for results on the asymptotics of particular maximum likelihood estimators.

It is natural to use the asymptotic normality of the estimator in order to form confidence intervals and test statistics. This requires an estimator of the standard error or equivalently of the Fisher information matrix. In some specific cases the efficient Fisher information matrix is of closed form. For example, under the assumption that the observation time is independent of the covariates, Huang (1996) gives an explicit estimator of the asymptotic variance in a proportional hazards model applied to current status data. Sometimes the 'efficient score' or 'efficient influence function' is explicit. Then since the efficient Fisher information matrix is the covariance of the efficient score function, one may estimate the

1350–7265

efficient score function and use the average over the sample of the squared estimated efficient score function to estimate the asymptotic variance. A similar procedure may be carried out if the efficient influence function is explicit. This is done in a mixture model by Gaydos and Lindsay (1996) and also by Huang (1996) when the independence assumption does not hold. In the latter case, the efficient score function is a function of the ratio of conditional means. Huang uses nonparametric smoothing to estimate each of the conditional means.

However, in general, the asymptotic covariance is not given by a closed formula, or even as an expectation of a known function – see van der Laan (1993), van der Vaart (1994b; 1994c), Murphy (1995) and Huang and Wellner (1995) for some examples. One possible option is to consider a discretized (for instance, at observed data points) version of the efficient information matrix. Then, to calculate the asymptotic covariance matrix, one must invert the matrix of high dimension. This is true, for instance, in the semi-parametric frailty model considered by Murphy (1995), where estimators for the standard error of the estimated frailty variance are found by inverting a matrix, which is of the same dimension as the data. In some models, the special structure of the model leads to other estimators (Parner 1996). In this paper, we consider a general method for the estimation of the asymptotic covariance based on using the 'observed profile information'. This is a natural generalization of a commonly used estimation method in parametric models.

A popular estimator for the asymptotic covariance of a maximum likelihood estimator in classical parametric models is the inverse of the 'observed information matrix'. The latter matrix is defined as

$$-\frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}_{\hat{\theta}}(X_i), \tag{1.1}$$

and is equal to $-(1/n)$ times the second derivative of the log-likelihood function, evaluated at the maximum likelihood estimator. As is well known, this estimator is asymptotically consistent for the inverse of the asymptotic variance under some regularity conditions. In practice, one might replace the analytic derivative in (1.1) by a discretized derivative, which can be computed directly from the likelihood function.

In a semi-parametric model the full parameter is partitioned into a parameter of interest and an infinite-dimensional nuisance parameter. The observed information matrix for the full parameter would be a linear operator, and its inverse may not exist in the models where a part of the nuisance parameter is not estimable at $\sqrt{n}$-rate. Thus, estimating the asymptotic covariance matrix of the maximum likelihood estimator of the parameter of interest by inverting this linear operator appears impractical. Instead, we propose to replace the likelihood function by the profile likelihood function, and use the 'observed profile information'.

More precisely, suppose that we observe a sample $X_1, \ldots, X_n$ from a distribution depending on a parameter $\psi = (\theta, \eta)$, ranging over a set $\Psi = \Theta \times H$. The parameter of interest is $\theta \in \Theta \subset \mathbb{R}^p$. Given a 'likelihood' $\mathrm{lik}(\theta, \eta)(x)$ for one observation x, define

$$\mathbb{M}_n(\theta) = \sup_{\eta} \frac{1}{n} \sum_{i=1}^{n} \log \mathrm{lik}(\theta, \eta)(X_i). \tag{1.2}$$

This is the *profile likelihood function* for estimating the parameter θ. The maximum likelihood estimator $\hat{\theta}$ is the maximum point of the map $\theta \mapsto \mathbb{M}_n(\theta)$. As an estimator for the asymptotic covariance matrix of $\hat{\theta}$ one could propose minus the inverse of the second derivative of $\theta \mapsto \mathbb{M}_n(\theta)$ evaluated at $\hat{\theta}$.

We can explain heuristically why this method might provide a consistent estimator of the inverse of the asymptotic covariance matrix as follows. If $\hat{\eta}_\theta$ achieves the supremum in (1.2), then the map $\theta \mapsto (\theta, \hat{\eta}_\theta)$ ought to be an estimator of a least favourable submodel for the estimation of θ (see Severini and Wong 1992). By definition, differentiation of the likelihood along the least favourable submodel (if the derivative exists) yields the efficient score function for θ. The efficient information matrix is the covariance matrix of the efficient score function, and, as usual, the expectation of minus the second derivative along this submodel should yield the same matrix.

The observed profile information is already used as an estimator in practice. For a simplistic example, consider estimation of the regression coefficient θ in Cox's proportional hazards model (with right censoring). Relative to a convenient choice of the likelihood, the estimator $\hat{\eta}_\theta$ of the cumulative baseline hazard function is an explicit function of the data and θ, and the profile likelihood function can be computed explicitly. In fact, this is Cox's partial likelihood (see Cox 1975; Andersen *et al.* 1993, pp. 481–482). The usual estimator of the inverse of the asymptotic variance, minus the second derivative of the partial likelihood, is precisely the observed profile information.

Severini and Wong (1992) and Severini and Staniswalis (1994) consider a particular class of semi-parametric models, and use a 'generalized' observed profile information to estimate the covariance matrix of $\hat{\theta}$. Their estimator of the nuisance parameter for a fixed θ is not a maximum likelihood estimator, but a weighted maximum likelihood estimator. However, considered as a function of θ, this estimator is an estimator of the least favourable submodel and is differentiable in θ. As a result, the likelihood evaluated at θ behaves as a profile likelihood for θ.

It is not clear from the definition of the profile likelihood $\theta \mapsto \mathbb{M}_n(\theta)$ that a second derivative matrix exists. If it does, then it may not be easily computable in models in which the estimator of the nuisance parameter is not explicit. To overcome these problems, discretized versions of the observed profile information are proposed by Nielsen *et al.* (1992), Huang and Wellner (1995) and Murphy *et al.* (1997). The main purpose of this paper is to prove the asymptotic consistency of such a discretized version. More precisely, under suitable conditions, we show that, for every $h_n \xrightarrow{\mathrm{P}} 0$ such that $(\sqrt{n}h_n)^{-1} = O_P(1)$,

$$-2\frac{\mathbb{M}_n(\hat{\theta} + h_n v_n) - \mathbb{M}_n(\hat{\theta})}{h_n^2} \xrightarrow{\mathrm{P}} v^{\mathrm{T}} \tilde{I}_0 v, \tag{1.3}$$

for every sequence of 'directions' $v_n \xrightarrow{\mathrm{P}} v \in \mathbb{R}^p$, where $\tilde{I}_0$ is the efficient information matrix for estimating θ, evaluated at the 'true' parameter $\psi_0 = (\theta_0, \eta_0)$. Note that as $h_n \to 0$ and for fixed n we obtain minus the second derivative of $\theta \mapsto \mathbb{M}_n(\theta)$ (if this exists) at $\theta = \hat{\theta}$, since its first derivative at this point vanishes by the definition of $\hat{\theta}$. The result (1.3) establishes the

consistency of most discretization schemes for calculating the second derivative matrix. For instance, with e_i the ith unit vector in $\mathbb{R}^p$,

$$-\frac{\mathbb{M}_n(\hat{\theta}+h_ne_i+h_ne_j)-\mathbb{M}_n(\hat{\theta}+h_ne_i)-\mathbb{M}_n(\hat{\theta}+h_ne_j)+\mathbb{M}_n(\hat{\theta})}{h_n^2} \xrightarrow{\mathrm{P}} (\tilde{I}_0)_{i,j}.$$

We check our conditions for a number of examples, using the theory of empirical processes. We believe that the approach works also for most of the other examples of semi-parametric likelihood estimators that have been treated in the literature so far. The proof is based on 'sandwiching' the profile likelihood, using approximately least favourable submodels. This is a similar device to that employed by Murphy and van der Vaart (1997) on semi-parametric likelihood ratio statistics.

The definition of a semi-parametric likelihood estimator requires the definition of a likelihood function for the model. In some models this is just a suitable version of the density of the observations, as in classical parametric models. In other models we use an empirical likelihood, which is a density (of the absolutely continuous part) with respect to counting measure, even though counting measure may not dominate the model. Combinations of these two extremes, as well as modifications, may be useful as well. For the theory it is sufficient that the function of the parameter and the observation that is designated to be 'the likelihood' satisfies certain regularity conditions. In the fourth example, 'the likelihood' is actually a penalized likelihood.

The paper is organized as follows. In Section 2 we formulate and prove the main result. One condition of the main theorem concerns a rate of convergence. In Section 3 we give two general approaches to establish this type of rate of convergence. In Sections 4–7 we verify the conditions for four non-trivial examples.

The symbols $\mathbb{P}_n$ and $\mathbb{G}_n$ are used for the empirical distribution and the empirical process of the observations, respectively. Furthermore, we use operator notation for evaluating expectations. Thus, for every measurable function f and probability measure P,

$$\mathbb{P}_nf=\frac{1}{n}\sum_{i=1}^n f(X_i), \qquad Pf=\int f\,\mathrm{d}P, \qquad \mathbb{G}_nf=\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i)-P_0f),$$

where P_0 is the true underlying measure of the observations. A distance function on the nuisance parameter space, H, is denoted by $d(\eta, \eta')$.

2. Main result

The maximum likelihood estimator for (θ, η) is the parameter $(\hat{\theta}, \hat{\eta})$ that maximizes the log-likelihood $(\theta, \eta) \mapsto \mathbb{P}_n \log \mathrm{lik}(\theta, \eta)$ defined in (1.2). The estimator $\hat{\theta}$ maximizes the profile likelihood $\theta \mapsto \mathbb{M}_n(\theta)$. We shall assume that this has already been shown to be asymptotically normal, and that

$$\sqrt{n}(\hat{\theta}_n-\theta_0)=\tilde{I}_0^{-1}\mathbb{G}_n\tilde{\ell}_0+o_P(1). \tag{2.1}$$

We refer to $\tilde{\ell}_0$ as the 'efficient score function', and to $\tilde{I}_0$ as the 'efficient Fisher information

matrix'. This is assumed to be the covariance matrix of $\tilde{\dot{\ell}}_0(X)$ under P_0 and to be non-singular.

For a fixed θ, denote by $\hat{\eta}_\theta$ a random element at which the supremum in the definition of $\mathbb{M}_n(\theta)$ is (nearly) achieved, and set $\hat{\psi}_\theta = (\theta, \hat{\eta}_\theta)$. Then $(\hat{\theta}, \hat{\eta}_{\hat{\theta}})$ is the maximum likelihood estimator of (θ, η).

Our assumptions all relate to the existence of approximately least favourable p-dimensional submodels. We assume that, for each $\psi = (\theta, \eta)$, there exists a map, which we denote by $t \mapsto \boldsymbol{\eta}_t(\psi)$, from a fixed neighbourhood of θ to the parameter set for η, such that the map $t \mapsto \dot{\ell}(t, \psi)(x)$ defined by

$$\dot{\ell}(t, \psi)(x) = \log \mathrm{lik}(t, \boldsymbol{\eta}_t(\psi))(x)$$

is twice continuously differentiable, for all x. We denote the derivatives by $\dot{\dot{\ell}}(t, \psi)(x)$ and $\ddot{\ell}(t, \psi)(x)$, respectively. The p-dimensional submodel with parameters $(t, \boldsymbol{\eta}_t(\psi))$ should pass through $\psi = (\theta, \eta)$ at $t = \theta$:

$$\boldsymbol{\eta}_\theta(\theta, \eta) = \eta, \qquad \text{every } (\theta, \eta). \tag{2.2}$$

The second important structural requirement that should lead to the construction of this submodel is that it be least favourable at (θ_0, η_0) for estimating θ in the sense that

$$\dot{\dot{\ell}}(\theta_0, \psi_0)(x) = \tilde{\dot{\ell}}_0. \tag{2.3}$$

More precisely, we need this equality together with some regularity conditions. Similar conditions are used by Murphy and van der Vaart (1997) to prove the validity of the likelihood ratio test. Assume that for any random sequences such that $\tilde{\theta} \xrightarrow{\mathrm{P}} \theta_0$ and $\overline{\psi} \xrightarrow{\mathrm{P}} \psi_0$,

$$\mathbb{G}_n \dot{\dot{\ell}}(\tilde{\theta}, \overline{\psi}) = \mathbb{G}_n \tilde{\dot{\ell}}_0 + o_P(1), \tag{2.4}$$

$$\mathbb{P}_n \ddot{\ell}(\tilde{\theta}, \overline{\psi}) \xrightarrow{\mathrm{P}} -\tilde{I}_0, \tag{2.5}$$

$$P_0 \dot{\dot{\ell}}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) = -\tilde{I}_0(\tilde{\theta} - \theta_0) + o_P(\|\tilde{\theta} - \theta_0\| + n^{-1/2}). \tag{2.6}$$

Here the assumption $\overline{\psi} \xrightarrow{\mathrm{P}} \psi_0$ implicitly assumes a topology on the set of nuisance parameters η. In applications of the following theorem this topology should be chosen such that $\hat{\eta}_{\tilde{\theta}} \xrightarrow{\mathrm{P}} \eta_0$ for every $\tilde{\theta} \xrightarrow{\mathrm{P}} \theta_0$.

Theorem 2.1. *Suppose that (2.1)–(2.2) and (2.4)–(2.6) are satisfied and that* $\hat{\eta}_{\tilde{\theta}} \xrightarrow{\mathrm{P}} \eta_0$ *for every* $\tilde{\theta} \xrightarrow{\mathrm{P}} \theta_0$. *Then (1.3) is valid for every random sequence* $h_n \xrightarrow{\mathrm{P}} 0$ *such that* $(\sqrt{n} h_n)^{-1} = O_P(1)$.

Proof. For $\overline{\theta} = \hat{\theta} + h_n v_n$, we have, by (2.2),

$$\mathbb{M}_n(\overline{\theta}) - \mathbb{M}_n(\hat{\theta}) = \mathbb{P}_n \log \mathrm{lik}(\overline{\theta}, \hat{\eta}_{\overline{\theta}}) - \mathbb{P}_n \log \mathrm{lik}(\hat{\theta}, \hat{\eta}_{\hat{\theta}})$$

$$\begin{cases} \geqslant \mathbb{P}_n \log \mathrm{lik}(\overline{\theta}, \boldsymbol{\eta}_{\overline{\theta}}(\hat{\psi}_{\hat{\theta}})) - \mathbb{P}_n \log \mathrm{lik}(\hat{\theta}, \boldsymbol{\eta}_{\hat{\theta}}(\hat{\psi}_{\hat{\theta}})) \\ \leqslant \mathbb{P}_n \log \mathrm{lik}(\overline{\theta}, \boldsymbol{\eta}_{\overline{\theta}}(\hat{\psi}_{\overline{\theta}})) - \mathbb{P}_n \log \mathrm{lik}(\hat{\theta}, \boldsymbol{\eta}_{\hat{\theta}}(\hat{\psi}_{\overline{\theta}})). \end{cases}$$

Both the upper and the lower bound are differences $\mathbb{P}_n \dot{\ell}(\overline{\theta}, \psi) - \mathbb{P}_n \dot{\ell}(\hat{\theta}, \psi)$, with $\psi = \hat{\psi}_{\overline{\theta}}$ and $\psi = \hat{\psi}_{\hat{\theta}}$, respectively. We apply a two-term Taylor expansion to these differences, leaving ψ fixed.

For the lower bound, we expand around $\hat{\theta}$ and obtain that this is equal to

$$h_n v_n^{\mathrm{T}} \mathbb{P}_n \dot{\ell}(\hat{\theta}, \hat{\psi}_{\hat{\theta}}) + \tfrac{1}{2} h_n^2 v_n^{\mathrm{T}} \mathbb{P}_n \ddot{\ell}(\tilde{\theta}, \hat{\psi}_{\hat{\theta}}) v_n,$$

for $\tilde{\theta}$ a convex combination of $\overline{\theta}$ and $\hat{\theta}$. The first term is zero because the map $t \mapsto \mathbb{P}_n \log \mathrm{lik}(t, \boldsymbol{\eta}_t(\hat{\psi}_{\hat{\theta}}))$ is maximized at $t = \hat{\theta}$, since $\hat{\psi}_{\hat{\theta}} = (\hat{\theta}, \hat{\eta})$, whence $\boldsymbol{\eta}_{\hat{\theta}}(\hat{\psi}_{\hat{\theta}}) = \hat{\eta}$, by (2.2). The second term is $-\frac{1}{2} h_n^2 (v_n^{\mathrm{T}} \tilde{I}_0 v_n + o_P(1))$ by assumption (2.5).

For the upper bound, we expand around $\overline{\theta}$ and obtain that this is equal to

$$h_n v_n^{\mathrm{T}} \mathbb{P}_n \dot{\ell}(\overline{\theta}, \hat{\psi}_{\overline{\theta}}) - \tfrac{1}{2} h_n^2 v_n^{\mathrm{T}} \mathbb{P}_n \ddot{\ell}(\tilde{\theta}, \hat{\psi}_{\overline{\theta}}) v_n,$$

for $\tilde{\theta}$ a convex combination of $\overline{\theta}$ and $\hat{\theta}$. The second term is $\frac{1}{2} h_n^2 (v_n^{\mathrm{T}} \tilde{I}_0 v_n + o_P(1))$ by assumption (2.5). The first term is equal to

$$\frac{h_n}{\sqrt{n}} v_n^{\mathrm{T}} \mathbb{G}_n \dot{\ell}(\overline{\theta}, \hat{\psi}_{\overline{\theta}}) + h_n v_n^{\mathrm{T}} P_0 \dot{\ell}(\overline{\theta}, \hat{\psi}_{\overline{\theta}})$$

$$= \frac{h_n}{\sqrt{n}} (v_n^{\mathrm{T}} \tilde{I}_0 \sqrt{n}(\hat{\theta} - \theta_0) + o_P(1)) - h_n [v_n^{\mathrm{T}} \tilde{I}_0 (\overline{\theta} - \theta_0) + o_P(\|\overline{\theta} - \theta_0\| + n^{-1/2})],$$

by (2.1) and (2.4), and (2.6), respectively. This reduces to $-h_n^2 (v_n^{\mathrm{T}} \tilde{I}_0 v_n + o_P(1))$ by the assumptions on h_n. □

Conditions (2.4) and (2.5) are regularity conditions on the least favourable submodel. They can be verified using the theory of empirical processes. See, for example, Lemma 2.2 below. These conditions can be slightly relaxed. To obtain the best result in one of our examples, we shall need to relax (2.4)–(2.5) to the conditions that for every $\tilde{\theta} \xrightarrow{\mathrm{P}} \theta_0$ and $\overline{\theta} \xrightarrow{\mathrm{P}} \theta_0$,

$$\mathbb{G}_n \dot{\ell}(\tilde{\theta}, \hat{\psi}_{\overline{\theta}}) = \mathbb{G}_n \tilde{\ell}_0 + o_P(1 + \sqrt{n}\|\overline{\theta} - \theta_0\|). \tag{2.4′}$$

$$\mathbb{P}_n \ddot{\ell}(\tilde{\theta}, \hat{\psi}_{\overline{\theta}}) \xrightarrow{\mathrm{P}} -\tilde{I}_0, \tag{2.5′}$$

The theorem goes through under this latter pair of conditions.

Condition (2.6) is more involved. There are several reasons why it ought to be valid. First, by its definition, $\hat{\psi}_\theta$ maximizes the log-likelihood for a fixed value of the parameter θ. It should be close to the maximizer of the Kullback–Leibler information $P_0 \log \mathrm{lik}(\psi)$ for a fixed parameter θ. As shown by Severini and Wong (1992), the latter maximizers should yield a least favourable submodel $\theta \mapsto \psi_\theta$ for the estimation of θ. In other words, the score function at θ_0 of the model $\theta \mapsto \mathrm{lik}(\hat{\psi}_\theta)$ should be close to the efficient score function $\tilde{\ell}_0$. Thus, we may expect

$$P_0 \dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) = (P_0 - P_{\hat{\psi}_{\tilde{\theta}}}) \dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}})$$

$$= -P_0 (\tilde{\theta} - \theta_0)^{\mathrm{T}} \dot{\tilde{\ell}}_0 \dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) + o_P(\|\tilde{\theta} - \theta_0\|).$$

This would yield (2.6), because by our construction $\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}})$ approaches $\tilde{\ell}_0$. This is probably the best intuitive justification of the condition. However, it is hard to make it precise. For instance, it appears already hard to show that the path $\theta \mapsto \log \mathrm{lik}(\hat{\psi}_\theta)$ would be differentiable.

The second intuitive justification of (2.6) is as follows. Since $\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}})$ is constructed to converge to $\tilde{\ell}_0$, we may expect

$$\begin{aligned} P_0\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) &= (P_0 - P_{\hat{\psi}_{\tilde{\theta}}})\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) \\ &= (P_0 - P_{\hat{\psi}_{\tilde{\theta}}})\tilde{\ell}_0 + o_P(\|\tilde{\theta} - \theta_0\|) \\ &= -P_0[(\tilde{\theta} - \theta_0)^{\mathrm{T}}\ell_0 + A_0(\hat{\eta}_{\tilde{\theta}} - \eta_0)]\tilde{\ell}_0 + o_P(\|\tilde{\theta} - \theta_0\|), \end{aligned}$$

where ℓ_0 and A_0 are the derivatives of the log-likelihood with respect to θ and η, respectively. Since the efficient score function $\tilde{\ell}_0$ is obtained by subtracting from the score ℓ_0 for θ its projection onto the score space for the parameter η (the range of A_0), the factor involving $A_0(\hat{\eta}_{\tilde{\theta}} - \eta_0)$ can be cancelled and the inner product of ℓ_0 and $\tilde{\ell}_0$ yields the matrix $\tilde{I}_0$.

The third approach is the least insightful one, but is the easiest one to implement in some examples. We start by proving that $P_0\dot{\ell}(\theta_0, \hat{\psi}_{\tilde{\theta}}) = o_P(\|\tilde{\theta} - \theta_0\| + n^{-1/2})$. This requires special properties of the model and/or a rate of convergence on the nuisance parameter, or, alternatively, an approach as in the preceding paragraphs. Then we may expect

$$\begin{aligned} P_0\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) &= P_0(\dot{\ell}(\tilde{\theta}, \psi_{\tilde{\theta}}) - \dot{\ell}(\theta_0, \hat{\psi}_{\tilde{\theta}})) + o_P(\|\tilde{\theta} - \theta_0\| + n^{-1/2}) \\ &= P_0\ddot{\ell}(\theta_0, \hat{\psi}_{\tilde{\theta}})(\tilde{\theta} - \theta_0) + o_P(\|\tilde{\theta} - \theta_0\| + n^{-1/2}) \\ &= -\tilde{I}_0(\tilde{\theta} - \theta_0) + o_P(\|\tilde{\theta} - \theta_0\| + n^{-1/2}). \end{aligned}$$

Here the last step follows by the usual identity relating the second derivative of the log-likelihood to the square of the first derivative, and is the population version of (2.5).

We summarize this last method, together with conditions to verify (2.4) and (2.5), in the following lemma. See, for example, van der Vaart and Wellner (1996) for the definitions and examples of Glivenko–Cantelli and Donsker classes. The lemma assumes implicitly that $\exp \ell(t, \psi)(x)$ is a probability density with respect to some dominating measure, up to a factor that does not depend on t, in order to verify equation (2.8).

Lemma 2.2. *Suppose that there exists a neighbourhood V of (θ_0, ψ_0) such that the class of functions $\{\dot{\ell}(t, \psi): (t, \psi) \in V\}$ is P_0-Donsker with square-integrable envelope function, and such that the class of functions $\{\ddot{\ell}(t, \psi): (t, \psi) \in V\}$ is P_0-Glivenko–Cantelli and is bounded in $L_1(P_0)$. Furthermore, suppose that the functions $(t, \psi) \mapsto \dot{\ell}(t, \psi)(x)$ and $(t, \psi) \mapsto \ddot{\ell}(t, \psi)(x)$ are continuous at (θ_0, ψ_0) for P_0-almost every x, and suppose that $\dot{\ell}(\theta_0, \psi_0) = \tilde{\ell}_0$. Then (2.4) and (2.5) are satisfied. Furthermore, if $\hat{\psi}_{\tilde{\theta}} \xrightarrow{\mathrm{P}} \psi_0$, then (2.6) is equivalent to*

$$P_0\dot{\ell}(\theta_0, \hat{\psi}_{\tilde{\theta}}) = o_P(\|\tilde{\theta} - \theta_0\| + n^{-1/2}). \tag{2.7}$$

Proof. Since $\dot{\ell}(t, \psi) \to \tilde{\ell}_0$ as $(t, \psi) \to (\theta_0, \psi_0)$, and the functions $\dot{\ell}(t, \psi)$ are dominated by a square-integrable function, we have by dominated convergence

$$P_0(\dot{\ell}(\tilde{\theta}, \overline{\psi}) - \tilde{\ell})^2 \xrightarrow{\mathrm{P}} 0.$$

Together with the assumption that the functions $\dot{\ell}(t, \psi)$ belong to a Donsker class, this yields (2.4). See, for example, Lemma 3.3.5 in van der Vaart and Wellner (1996).

Similarly, using the Glivenko–Cantelli assumption, we have

$$P_0\ddot{\ell}(\tilde{\theta}, \overline{\psi}) \xrightarrow{\mathrm{P}} P_0\ddot{\ell}(\theta_0, \psi_0),$$

$$\mathbb{P}_n\ddot{\ell}(\tilde{\theta}, \overline{\psi}) \xrightarrow{\mathrm{P}} P_0\ddot{\ell}(\theta_0, \psi_0).$$

Since $t \mapsto \exp \ell(t, \psi)$ is proportional to a smooth one-dimensional submodel, its derivatives satisfy the usual identity

$$P_0\ddot{\ell}(\theta_0, \psi_0) = -P_0\dot{\ell}^2(\theta_0, \psi_0) = -\tilde{I}_0. \tag{2.8}$$

This completes the proof of (2.5).

For the proof of (2.6) we have, by Taylor's theorem, for $\overline{\theta}$ a point between $\tilde{\theta}$ and θ_0,

$$P_0\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) = P_0\dot{\ell}(\theta_0, \hat{\psi}_{\tilde{\theta}}) + P_0\ddot{\ell}(\overline{\theta}, \hat{\psi}_{\tilde{\theta}})(\tilde{\theta} - \theta_0).$$

The expectation in the second term on the right converges in probability to $-\tilde{I}_0$. □

3. Rates of convergence

The verification of (2.6) or (2.7) may require a rate of convergence of the 'estimators' $\hat{\eta}_{\tilde{\theta}}$. In this section we present two theorems that yield such a rate. Both theorems extend general results on M-estimators to M-estimators with estimated nuisance parameters, and are also of independent interest.

In our first theorem, consider estimators $\hat{\eta}_{\tilde{\theta}}$ such that

$$\mathbb{P}_n\kappa_{\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}, h} = 0,$$

for a collection of measurable functions $x \mapsto \kappa_{\theta,\eta,h}(x)$ indexed by the parameter (θ, η) and an arbitrary index $h \in \mathscr{H}$. In examples, these functions often take the form $A_{\theta,\eta}h$ or $A_{\theta,\eta}h - P_{\theta,\eta}A_{\theta,\eta}h$ for a 'score operator' $A_{\theta,\eta}$. Define

$$W_{n2}(\theta, \eta)h = \mathbb{P}_n\kappa_{\theta,\eta,h},$$

$$W_2(\theta, \eta)h = P_{\theta_0,\eta_0}\kappa_{\theta,\eta,h}.$$

(The index 2 is superfluous here, but makes the notation consistent with proofs of asymptotic normality of the maximum likelihood estimators, and our examples.) We assume that the maps $h \mapsto W_{n2}(\theta, \eta)h$ and $h \mapsto W_2(\theta, \eta)h$ are uniformly bounded, so that W_{n2} and W_2 can be viewed as maps from the parameter set $\Theta \times H$ into $\ell^\infty(\mathscr{H})$. The parameter set H for η is

viewed as a subset of a Banach space $\mathbb{L}$ with norm d. We impose the following regularity conditions. For some $\delta > 0$,

$$\{\kappa_{\theta,\eta,h}\colon \|\theta - \theta_0\| < \delta,\ d(\eta, \eta_0) < \delta,\ h \in \mathscr{H}\} \text{ is } P_{\theta_0,\eta_0}\text{-Donsker}, \tag{3.1}$$

$$\sup_{h\in\mathscr{H}} P_{\theta_0,\eta_0}(\kappa_{\theta,\eta,h} - \kappa_{\theta_0,\eta_0,h})^2 \to 0, \qquad \theta \to \theta_0,\ \eta \to \eta_0. \tag{3.2}$$

Theorem 3.1. *Suppose that* $W_2\colon \Theta \times H \subset \mathbb{R}^p \times \mathbb{L} \mapsto \ell^\infty(\mathscr{H})$ *is Fréchet-differentiable at* (θ_0, η_0) *with derivative* $\dot{W}_2\colon \mathbb{R}^p \times \operatorname{lin} H \mapsto \ell^\infty(\mathscr{H})$ *such that the map* $\dot{W}_2(0, \cdot)\colon \operatorname{lin} H \mapsto \ell^\infty(\mathscr{H})$ *is invertible with an inverse that is continuous on its range. Furthermore, assume that (3.1) holds, that* $W_2(\theta_0, \eta_0) = 0$, *that* $\tilde{\theta} \xrightarrow{P} \theta_0$ *and that* $\hat{\eta}_{\tilde{\theta}} \xrightarrow{P} \eta_0$. *Then* $d(\hat{\eta}_{\tilde{\theta}}, \eta_0) = O_P^*(n^{-1/2} + \|\tilde{\theta} - \theta_0\|)$ *and when (3.2) also holds,*

$$\dot{W}_2(0, \hat{\eta}_{\tilde{\theta}} - \eta_0) = -(W_{n2} - W_2)(\theta_0, \eta_0) - \dot{W}_2(\tilde{\theta} - \theta_0, 0) + o_P^*(\|\tilde{\theta} - \theta_0\| + n^{-1/2}).$$

Proof. By the definition of $\hat{\eta}_\theta$,

$$\begin{aligned} W_2(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) - W_2(\theta_0, \eta_0) &= W_2(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) - W_{n2}(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) \\ &= -(W_{n2} - W_2)(\theta_0, \eta_0) + o_P^*(n^{-1/2}), \end{aligned} \tag{3.3}$$

by (3.1) and (3.2) – see, for example, Lemma 3.3.5 in van der Vaart and Wellner (1996). By the differentiability of W_2,

$$\begin{aligned} \dot{W}_2(\tilde{\theta} - \theta_0, \hat{\eta}_{\tilde{\theta}} - \eta_0) &= W_2(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) - W_2(\theta_0, \eta_0) + o_P^*(\|\tilde{\theta} - \theta_0\| + d(\hat{\eta}_{\tilde{\theta}}, \eta_0)), \\ &= -(W_{n2} - W_2)(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) + o_P^*(\|\tilde{\theta} - \theta_0\| + d(\hat{\eta}_{\tilde{\theta}}, \eta_0)) \end{aligned} \tag{3.4}$$

by the first line in (3.3). Since $\dot{W}_2$ is linear, the left-hand side is equal to $\dot{W}_2(0, \hat{\eta}_{\tilde{\theta}} - \eta_0) + \dot{W}_2(\tilde{\theta} - \theta_0, 0)$. The first term on the right in (3.4) is of the order $O_P(n^{-1/2})$ by (3.1). In view of the continuous invertibility of $\dot{W}_2$, it follows that $d(\hat{\eta}_{\tilde{\theta}}, \eta_0)$ is of the order $O_P(n^{-1/2} + \|\tilde{\theta} - \theta_0\|)$, thus verifying the first assertion of the theorem. Reinsert this on the right-hand side of the preceding display and use the second line of (3.3) to find the second assertion. □

The preceding theorem is a variation on the theorem used by van der Vaart (1994b; 1994c) and Murphy (1995), among others, to prove the asymptotic normality of the maximum likelihood estimator $(\hat{\theta}, \hat{\eta})$. Actually, its conditions are implied by the conditions imposed in these papers, so that, at least in these cases, the estimator $\hat{\eta}_{\tilde{\theta}}$ behaves well whenever $(\hat{\theta}_n, \hat{\eta}_n)$ behaves well and $\tilde{\theta}$ behaves well. Of course, not using the maximum likelihood estimator for $\tilde{\theta}$ may cause the estimator $\hat{\eta}_{\tilde{\theta}}$ for η to be inefficient.

In our second theorem, consider estimators $\hat{\eta}_\theta$ contained in a set H_n that, for a given θ, satisfy

$$\mathbb{P}_n m_{\theta,\hat{\eta}_\theta} \geqslant \mathbb{P}_n m_{\theta,\eta_0}$$

for given measurable functions $x \mapsto m_{\theta,\eta}(x)$. This is valid, for example, for $\hat{\eta}_\theta$ equal to the maximizer of the function $\eta \mapsto \mathbb{P}_n m_{\theta,\eta}$ over H_n, if this set contains η_0.

Assume that the following conditions are satisfied for every $\theta \in \Theta_n$, every $\eta \in H_n$ and every $\delta > 0$. The symbols $\gtrsim$ and $\lesssim$ mean greater than, or smaller than, up to a constant that may depend on the true parameter or the model, but not on any other parameter values.

$$P_0(m_{\theta,\eta} - m_{\theta,\eta_0}) \lesssim -d_\theta^2(\eta, \eta_0) + \|\theta - \theta_0\|^2, \tag{3.5}$$

$$E^* \sup_{\theta \in \Theta_n, \eta \in H_n, \|\theta - \theta_0\| < \delta, d_\theta(\eta, \eta_0) < \delta} |\mathbb{G}_n(m_{\theta,\eta} - m_{\theta,\eta_0})| \lesssim \phi_n(\delta). \tag{3.6}$$

Here $d_\theta^2(\eta, \eta_0)$ may be thought of as the square of a distance, but the following theorem is true for arbitrary functions $\eta \mapsto \mathrm{d}_\theta^2(\eta, \eta_0)$. (Contrary to what the notation suggests, this function may even take negative values. In the latter case, set $d_\theta(\eta, \eta_0) = (d_\theta^2(\eta, \eta_0) \vee 0)^{1/2}$.) In particular, it may be set equal to the infimum over θ of minus the left-hand side of (3.5), thus rendering this to be trivially satisfied. Usually d_θ does not depend on θ but in this form the following theorem is flexible enough to apply to penalized minimum contrast estimators, where the smoothing parameter can be included in θ. See Section 7.

Theorem 3.2. *Suppose that (3.6) is valid for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ and sets $\Theta_n \times H_n$ such that $\mathrm{P}(\tilde{\theta} \in \Theta_n, \hat{\eta}_{\tilde{\theta}} \in H_n) \to 1$. Then $d_{\tilde{\theta}}(\hat{\eta}_{\tilde{\theta}}, \eta_0) \lesssim O_P^*(\delta_n + \|\tilde{\theta} - \theta_0\|)$ for any sequence of positive numbers δ_n such that $\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n^2$ for every n.*

Proof. For each $n \in \mathbb{N}$, $j \in \mathbb{Z}$ and $M > 0$ define a set

$$S_{n,j,M} = \{(\theta, \eta) \in \Theta_n \times H_n : 2^{j-1}\delta_n < d_\theta(\eta, \eta_0) \lesssim 2^j\delta_n, \|\theta - \theta_0\| \lesssim 2^{-M} d_\theta(\eta, \eta_0)\}.$$

Then the intersection of the events $\tilde{\theta} \in \Theta_n$, $\hat{\eta}_{\tilde{\theta}} \in H_n$ and $d_{\tilde{\theta}}(\hat{\eta}_{\tilde{\theta}}, \eta_0) \gtrsim 2^M(\delta_n + \|\tilde{\theta} - \theta_0\|)$ is contained in the union of the events $\{(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) \in S_{n,j,M}\}$ over $j \gtrsim M$. By the definition of $\hat{\eta}_{\tilde{\theta}}$, the variable $\sup_{(\theta,\eta) \in S_{n,j,M}} \mathbb{P}_n(m_{\theta,\eta} - m_{\theta,\eta_0})$ is non-negative on the event $\{(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) \in S_{n,j,M}\}$. Conclude that, for every $\delta > 0$,

$$\begin{aligned} P^*(d_{\tilde{\theta}}(\hat{\eta}_{\tilde{\theta}}, \eta_0) \gtrsim 2^M(\delta_n + \|\tilde{\theta} - \theta_0\|), \tilde{\theta} \in \Theta_n, \hat{\eta}_{\tilde{\theta}} \in H_n) \\ \lesssim \sum_{j \gtrsim M} \mathrm{P}^*\Big(\sup_{(\theta,\eta) \in S_{j,n,M}} \mathbb{P}_n(m_{\theta,\eta} - m_{\theta,\eta_0}) \gtrsim 0\Big). \end{aligned}$$

For every j involved in the sum, we have, for every $(\theta, \eta) \in S_{j,n,M}$ and every sufficiently large M,

$$\begin{aligned} P_0(m_{\theta,\eta} - m_{\theta,\eta_0}) &\lesssim -d_\theta^2(\eta, \eta_0) + \|\theta - \theta_0\|^2 \\ &\lesssim -(1 - 2^{-2M}) d_\theta^2(\eta, \eta_0) \lesssim -2^{2j-2}\delta_n^2. \end{aligned}$$

Thus, using Markov's inequality, we see that the series is bounded by

$$\sum_{j \geqslant M} \mathrm{P}^*\Big(\sup_{(\theta,\eta)\in S_{j,n,M}} |\mathbb{G}_n(m_{\theta,\eta} - m_{\theta,\eta_0})| \geqslant \sqrt{n} 2^{2j-2}\delta_n^2\Big) \leqslant \sum_{j \geqslant M} \frac{\phi_n(2^{j+1}\delta_n)}{\sqrt{n}\delta_n^2 2^{2j}}$$

$$\leqslant \sum_{j \geqslant M} 2^{j\alpha - 2j},$$

in view of the definition of δ_n, and the fact that $\phi_n(c\delta) \leqslant c^\alpha \phi_n(\delta)$ for every $c > 1$ by the assumption on ϕ_n. This expression converges to zero for every $M = M_n \to \infty$. □

For $d_\theta = d$ not depending on θ condition (3.5) is implied by the conditions

$$P_0(m_{\theta,\eta_0} - m_{\theta_0,\eta_0}) \geqslant -\|\theta - \theta_0\|^2, \tag{3.7}$$

$$P_0(m_{\theta,\eta} - m_{\theta_0,\eta_0}) \leqslant -d^2(\eta, \eta_0). \tag{3.8}$$

The two conditions are the natural requirement that the criterion function $(\theta, \eta) \mapsto P_0 m_{\theta,\eta}$ behaves quadratically (relative to a distance) around the point of maximum (θ_0, η_0). There is more chance that this is true in a neighbourhood of (θ_0, η_0). Thus, it is useful to note that the theorem remains true if the conditions (3.6), (3.7) and (3.8) hold only for (θ, η) in this neighbourhood and every sufficiently small δ, provided that $(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}})$ are known to be consistent. We shall use this observation in our examples without much comment.

Condition (3.5) concerns the modulus of continuity of the empirical process and is more technical. A simple method to verify this condition is given by the following lemma. Let $\mathscr{M}_\delta$ be the set of all functions $x \mapsto m_{\theta,\eta}(x) - m_{\theta,\eta_0}(x)$ with $d_\theta(\eta, \eta_0) < \delta$ and $\|\theta - \theta_0\| < \delta$ and write $J(\delta, \mathscr{M}_\delta, L_2(P_0))$ for its entropy-with-bracketing integral

$$J(\delta, \mathscr{M}_\delta, L_2(P_0)) = \int_0^\delta \sqrt{1 + \log \mathrm{N}_{[]}(\varepsilon, \mathscr{M}_\delta, L_2(P_0))}\, \mathrm{d}\varepsilon.$$

Lemma 3.3. *Suppose that the functions $(x, \theta, \eta) \mapsto m_{\theta,\eta}(x)$ are uniformly bounded for (θ, η) ranging over a neighbourhood of (θ_0, η_0) and that*

$$P_0(m_{\theta,\eta} - m_{\theta,\eta_0})^2 \leqslant d_\theta^2(\eta, \eta_0) + \|\theta - \theta_0\|^2. \tag{3.9}$$

Then condition (3.6) is satisfied for any functions ϕ_n such that

$$\phi_n(\delta) \geqslant J(\delta, \mathscr{M}_\delta, L_2(P_0))\left(1 + \frac{J(\delta, \mathscr{M}_\delta, L_2(P_0))}{\delta^2\sqrt{n}}\right). \tag{3.10}$$

Consequently, in the conclusion of Theorem 3.2 we may use $J(\delta, \mathscr{M}_\delta, L_2(P_0))$ instead of $\phi_n(\delta)$.

Proof. The first assertion is an immediate consequence of Lemma 3.4.2 in van der Vaart and Wellner (1996).

For the second assertion, let ϕ_n be equal to the right-hand side of (3.10), and note that the equations $\phi_n(\delta) \leqslant \sqrt{n}\delta^2$ and $J(\delta) \leqslant \sqrt{n}\delta^2$ are equivalent. □

4. Cox's regression model for current status data

In current status data, n subjects are examined each at a random observation time and at this time it is observed whether the survival time has occurred or not. The survival time, T, is assumed independent of the observation time, Y, given the covariate, Z. Suppose that the hazard function of T given $Z = z$ is given by Cox's regression model: the hazard at time y is $e^{\theta^T z}\lambda(y)$. Then the cumulative hazard at time y of T given $Z = z$ is of the form $e^{\theta^T z}\int_0^y \lambda(s)\,ds\, e^{\theta z}\Lambda(y)$. The unknown parameters are θ, a vector of regression coefficients, in a known compact subset of $\mathbb{R}^p$, and $\Lambda \in \mathbf{\Lambda}$, the set of non-decreasing, cadlag functions from the positive real line to $[0, M]$, for a known M. We observe n i.i.d. copies of $X = (Y, \delta, Z)$, where $\delta = 1$ if $T \leqslant Y$ and zero otherwise.

The density of X is given by

$$p_{\theta,\Lambda}(X) = (1 - \exp(-e^{\theta^T Z}\Lambda(Y)))^{\delta}(\exp(-e^{\theta^T Z}\Lambda(Y)))^{1-\delta} f^{Y,Z}(Y, Z),$$

where $f^{Y,Z}$ is the joint density of (Y, Z). Since we are interested in inference for (θ, Λ) only, we take the likelihood $\text{lik}(\theta, \Lambda, X)$ equal to this expression, but with the term $f^{Y,Z}(Y, Z)$ omitted.

We make the following assumptions. The observation times Y are in an interval $[\sigma, \tau]$ and possess a Lebesgue density which is continuous and positive on $[\sigma, \tau]$. The true parameter θ_0 is an interior point of the parameter set, and the true parameter Λ_0 satisfies $\Lambda_0(\sigma -) > 0$ and $\Lambda_0(\tau) < M$, and is continuously differentiable on $[\sigma, \tau]$. The covariate vector Z is bounded and $E\{\text{cov}(Z|Y)\} > 0$. Finally, we assume that the function h_0 given by (4.1) has a version which is differentiable with a bounded derivative on $[\sigma, \tau]$.

Under these assumptions the maximum likelihood estimator of (θ, Λ) exists, $\hat{\theta}$ is asymptotically efficient in the sense of (2.1) and $\|\hat{\Lambda} - \Lambda_0\|_2 = O_P(n^{-1/3})$. Here $\|\cdot\|$ is the L_2-norm on $[\sigma, \tau]$. See Huang (1996) and Murphy and van der Vaart (1997).

In this model the score function for θ takes the form

$$\dot{\ell}_{\theta,\Lambda}(x) = z\Lambda(y)Q(x; \theta, \Lambda),$$

for the function $Q(x; \theta, \Lambda)$ given by

$$Q(x; \theta, \Lambda) = e^{\theta^T z}\left[\delta \frac{e^{-e^{\theta^T z}\Lambda(y)}}{1 - e^{-e^{\theta^T z}\Lambda(y)}} - (1 - \delta)\right].$$

Inserting a submodel $t \mapsto \Lambda_t$ such that $h(y) = -\partial/\partial t_{|t=0}\Lambda_t(y)$ exists for every y into the log-likelihood and differentiating at $t = 0$ we obtain a score function for Λ of the form

$$A_{\theta,\Lambda}h(x) = h(y)Q(x; \theta, \Lambda). \tag{4.1}$$

For every non-decreasing, non-negative function h the submodel $\Lambda_t = \Lambda + th$ is well defined if t is positive and yields a (one-sided) derivative h at $t = 0$. Thus (4.1) gives a (one-sided) score for Λ at least for all h of this type. The linear span of these functions contains $\dot{\ell}_{\theta,\Lambda}h$ for all bounded functions h of bounded variation. The efficient score function for θ is defined as $\tilde{\dot{\ell}}_0 = \dot{\ell}_{\theta,\Lambda} - A_{\theta,\Lambda}h_0$ for the vector of functions h_0 minimizing the distance $P_{\theta\Lambda}\|\dot{\ell}_{\theta,\Lambda} - A_{\theta,\Lambda}h\|^2$. In view of the similar structure of the scores for θ and Λ, this is a

weighted least-squares problem with weight function $Q(x; \theta, \Lambda)$. The solution at the true parameters is given by the vector-valued function

$$h_0(Y) = \Lambda_0(Y)h_{00}(Y) = \Lambda_0(Y)\frac{\mathrm{E}_{\theta_0\Lambda_0}(ZQ^2(X; \theta_0, \Lambda_0)|Y)}{\mathrm{E}_{\theta_0\Lambda_0}(Q^2(X; \theta_0, \Lambda_0)|Y)}. \tag{4.2}$$

As the formula shows (and as follows from the nature of the minimization problem), the vector of functions $h_0(y)$ is unique only up to null sets for the distribution of Y. However, it is an assumption that (under the true parameters) there exists a version of the conditional expectation that is differentiable with bounded derivative.

Thus we define, for t a vector in $\mathbb{R}^p$,

$$\mathbf{\Lambda}_t(\theta, \Lambda) = \Lambda + (\theta - t)^{\mathrm{T}}\phi(\Lambda)(h_{00} \circ \Lambda_0^{-1} \circ \Lambda)$$

$$\ell(t, \theta, \Lambda) = \log \mathrm{lik}(t, \mathbf{\Lambda}_t(\theta, \Lambda)).$$

Here $\phi: [0, M] \mapsto [0, \infty)$ is a fixed function such that $\phi(y) = y$ on the interval $[\Lambda_0(\sigma), \Lambda_0(\tau)]$, such that the function $y \mapsto \phi(y)/y$ is Lipschitz and such that $\phi(y) \leqslant c(y \wedge (M - y))$ for a sufficiently large constant c specified below (and depending on (θ_0, Λ_0) only). (By our assumption that $[\Lambda_0(\sigma), \Lambda_0(\tau)] \subset (0, M)$ such a function exists.) The function $\mathbf{\Lambda}_t(\theta, \Lambda)$ is essentially Λ plus a perturbation in the least favourable direction, but its definition is somewhat complicated in order to ensure that $\mathbf{\Lambda}_t(\theta, \Lambda)$ really defines a cumulative hazard function within our parameter space, at least for t that are sufficiently close to θ. First, the construction using $h_{00} \circ \Lambda_0^{-1} \circ \Lambda$, rather than h_{00}, (taken from Huang 1996) ensures that the perturbation that is added to Λ is absolutely continuous with respect to Λ; otherwise $\mathbf{\Lambda}_t(\theta, \Lambda)$ would not be a non-decreasing function. Second, the function ϕ 'truncates' the values of the perturbed hazard function to $[0, M]$.

A precise proof that $\mathbf{\Lambda}_t(\theta, \Lambda)$ is a parameter is as follows. Since the function ϕ is bounded and Lipschitz and, by assumption, $h_{00} \circ \Lambda_0^{-1}$ is bounded and Lipschitz, so is their product and hence, for $u \leqslant v$ and $\|\theta - t\| < \varepsilon$,

$$\mathbf{\Lambda}_t(\theta, \Lambda)(v) - \mathbf{\Lambda}_t(\theta, \Lambda)(u) \geqslant (\Lambda(v) - \Lambda(u))(1 - \varepsilon\|\phi h_{00} \circ \Lambda_0^{-1}\|_{\mathrm{Lipschitz}}).$$

For sufficiently small ε the right-hand side is non-negative. Next, for $\|\theta - t\| < \varepsilon$,

$$\mathbf{\Lambda}_t(\theta, \Lambda) \leqslant \Lambda + \varepsilon\phi(\Lambda)\|h_{00}\|_\infty.$$

This is certainly bounded above by M (on $[0, \tau]$) if $\phi(y) \leqslant (M - y)/(\varepsilon\|h_{00}\|_\infty)$ for all $0 \leqslant y \leqslant M$. Finally, $\mathbf{\Lambda}_t(\theta, \Lambda)$ can be seen to be non-negative on $[\sigma, \tau]$ by the condition that $\phi(y) \leqslant cy$.

It is proved below that

$$\|\hat{\Lambda}_{\tilde{\theta}} - \Lambda_0\|_2 = O_P(\|\tilde{\theta} - \theta_0\| + n^{-1/3}). \tag{4.3}$$

Thus, we shall use the L_2-norm on the nuisance parameter set.

Differentiating $\ell(t, \theta, \Lambda)$ with respect to t yields

$$\dot{\ell}(x; t, \theta, \Lambda) = \left[z - \frac{\phi(\Lambda)(y)}{\mathbf{\Lambda}_t(\theta, \Lambda)(y)} h_{00} \circ \Lambda_0^{-1} \circ \Lambda(y)\right]\mathbf{\Lambda}_t(\theta, \Lambda)(y)Q(x; t, \mathbf{\Lambda}_t(\theta, \Lambda)).$$

For $(t, \theta, \Lambda) = (\theta_0, \theta_0, \Lambda_0)$ this reduces to $\tilde{\dot{\ell}}_0$, since $\Lambda_0(\tau) < M$ by assumption, thus verifying equation (2.3). Murphy and van der Vaart (1997) verify the conditions of Lemma 2.2 when θ is a scalar; the verification for a vector θ is similar.

All that remains for the application of Theorem 2.1 is a verification of equation (2.7). Abbreviating $\dot{\ell}(\cdot; \theta_0, \theta_0, \Lambda)$ to $\dot{\ell}(\Lambda)$, we have

$$P_0\dot{\ell}(\theta_0, \theta_0, \hat{\Lambda}_{\tilde{\theta}}) = (P_0 - P_{\theta_0,\hat{\Lambda}_{\tilde{\theta}}})\dot{\ell}(\Lambda_0) + (P_0 - P_{\theta_0,\hat{\Lambda}_{\tilde{\theta}}})(\dot{\ell}(\hat{\Lambda}_{\tilde{\theta}}) - \dot{\ell}(\Lambda_0)). \tag{4.4}$$

Since $\dot{\ell}(\Lambda_0)$ is the efficient score function for θ and hence is orthogonal to every Λ-score, the first term on the right can be rewritten as

$$P_0\dot{\ell}(\Lambda_0)[(p_0 - p_{\theta_0,\hat{\Lambda}_{\tilde{\theta}}})/p_0 - \ell_\Lambda(\theta_0, \Lambda_0)(\Lambda_0 - \hat{\Lambda}_{\tilde{\theta}})]. \tag{4.5}$$

Here the term in square brackets is exactly the linear approximation in $\Lambda_0 - \hat{\Lambda}_{\tilde{\theta}}$ of the first. Taking the Taylor expansion one term further shows that the term in square brackets is bounded by a multiple of $(\Lambda_0 - \hat{\Lambda}_{\tilde{\theta}})^2$ and hence (4.5) is bounded by a multiple of $P_0(\Lambda_0 - \hat{\Lambda}_{\tilde{\theta}})^2$, which is negligible to the right order by (4.3). The second term in (4.4) can be bounded similarly, since both $\Lambda \mapsto p_{\theta_0,\Lambda}$ and $\Lambda \mapsto \dot{\ell}(\theta_0, \theta_0, \Lambda)$ are uniformly Lipschitz functions. This verifies (2.7) with a $o_P(n^{-2/3})$ remainder term, but with $(\theta_0, \hat{\Lambda}_{\tilde{\theta}})$ in place of $\hat{\psi}_{\tilde{\theta}} = (\tilde{\theta}, \hat{\Lambda}_{\tilde{\theta}})$. The difference of these two expressions can be seen to be $o_P(\|\tilde{\theta} - \theta_0\|)$, and (2.7) follows. (Note that $P_0\partial/\partial\theta\dot{\ell}(\theta_0, \theta, \eta_0)$ evaluated at $\theta = \theta_0$ vanishes, by the usual manipulations with (efficient) score functions.)

Finally, we prove (4.3). Since $\hat{\Lambda}_\theta$ maximizes the log-likelihood for fixed θ, and since $x \mapsto \log x$ is concave,

$$0 \leqslant \mathbb{P}_n \log \frac{p_{\theta,\hat{\Lambda}_\theta}}{p_{\theta,\Lambda_0}} = \mathbb{P}_n \left(\log \frac{p_{\theta,\hat{\Lambda}_\theta}}{p_{\theta_0,\Lambda_0}} - \log \frac{p_{\theta,\Lambda_0}}{p_{\theta_0,\Lambda_0}} \right)$$

$$\leqslant 2\mathbb{P}_n \log \frac{p_{\theta,\hat{\Lambda}_\theta} + p_{\theta_0,\Lambda_0}}{2p_{\theta_0,\Lambda_0}} - \mathbb{P}_n \log \frac{p_{\theta,\Lambda_0}}{p_{\theta_0,\Lambda_0}}.$$

With this in mind, we may apply Theorem 3.2 with $\eta = \Lambda$ and

$$m_{\theta,\Lambda} = \begin{cases} \log \dfrac{p_{\theta,\Lambda_0}}{p_{\theta_0,\Lambda_0}} & \text{if } \Lambda = \Lambda_0 \\ 2 \log \dfrac{p_{\theta,\Lambda} + p_{\theta_0,\Lambda_0}}{2p_{\theta_0,\Lambda_0}} & \text{otherwise.} \end{cases}$$

This choice of $m_{\theta,\Lambda}$ has the advantage over the more obvious choice $\log(p_{\theta,\Lambda}/p_{\theta,\Lambda_0})$ that the functions $m_{\theta,\Lambda}$ are uniformly bounded, thus permitting the application of Lemma 3.3. (Note that, by our assumptions, $\text{lik}(\theta, \Lambda_0)(x)$ is bounded away from 0 and ∞, uniformly in x and θ.)

Equation (3.8) holds for θ in a neighbourhood of θ_0 and every Λ, with d equal to the L_2-norm, by Lemma 8.5 of Murphy and van der Vaart (1997) and the well-known relation $P\log(q/p) \leqslant -h^2(p, q)$, relating Kullback–Leibler divergence and squared Hellinger distance – see, for example, the proof of Lemma 5.35 in van der Vaart (1998). A Taylor series argument in θ suffices to verify equation (3.7). To verify (3.6) we use Lemma 3.3. Arguments as the proof of Lemma 3.1 of Huang (1996) and Lemma 8.4 of Murphy and van

der Vaart (1997) show that $J(\delta) \leqslant \delta^{1/2}$. A Taylor series argument can be used to verify (3.9). Thus, Theorem 3.2 shows that (4.3) is satisfied.

5. Proportional odds model for right-censored data

In the proportional odds model, the survival function is parameterized such that the ratios of the odds of survival for subjects with different covariates are constant with time: the conditional survival function $S_Z(u)$ of the event time, T, given the covariates Z, satisfies

$$-\text{logit}(S_Z(u)) = \log \eta(u) + Z^{\mathrm{T}}\theta,$$

where $\text{logit}(x) = \log(x/(1-x))$. The unknown parameters are θ, a vector of regression coefficients ranging over a known compact subset of $\mathbb{R}^p$, and η, a non-decreasing, cadlag function from the positive real line to the positive real line, with $\eta(0) = 0$. We observe n i.i.d. copies of $X = (Y, \delta, Z)$, where $Y = T \wedge C$ is the minimum of T and a censoring time C which, given a vector of covariates Z, are independent. The censoring indicator δ is 1 if $T \leqslant C$ and 0 otherwise.

For $\mathrm{d}\eta$ a density of η with respect to some dominating measure, the density of X is

$$p_{\theta,\eta}(x) = \left(\frac{\mathrm{e}^{-z^{\mathrm{T}}\theta}(1 - F_C(y - |z))}{(\eta(y) + \mathrm{e}^{-z^{\mathrm{T}}\theta})(\eta(y-) + \mathrm{e}^{-z^{\mathrm{T}}\theta})} \mathrm{d}\eta(y) \right)^{\delta} \left(\frac{\mathrm{e}^{-z^{\mathrm{T}}\theta}}{\eta(y) + \mathrm{e}^{-z^{\mathrm{T}}\theta}} f_C(y|z) \right)^{1-\delta} f_Z(z),$$

where F_Z is the marginal distribution of Z, F_C is the conditional distribution of C given Z, and lower-case letters denote the respective densities. This density is not suitable for use as a likelihood. Instead, we use the empirical likelihood, which is obtained by replacing the densities f_C, $\mathrm{d}\eta$ and f_Z by the point probabilities $F_C\{Y\}$, $\eta\{Y\}$ and $F_Z\{Y\}$. Since we are interested in inference about (θ, η) only, we drop the terms involving F_C and F_Z, and define the likelihood to be

$$\text{lik}(\theta, \eta)(x) = \left(\frac{\mathrm{e}^{-z^{\mathrm{T}}\theta}\eta\{y\}}{(\eta(y) + \mathrm{e}^{-z^{\mathrm{T}}\theta})(\eta(y-) + \mathrm{e}^{-z^{\mathrm{T}}\theta})} \right)^{\delta} \left(\frac{\mathrm{e}^{-z^{\mathrm{T}}\theta}}{\eta(y) + \mathrm{e}^{-z^{\mathrm{T}}\theta}} \right)^{1-\delta}.$$

Murphy *et al.* (1997) show that the maximum likelihood estimator of (θ, η) exists, is consistent and is asymptotically normal and efficient under the following assumptions. First, for a finite number τ, both $\mathrm{P}(C \geqslant \tau) = \mathrm{P}(C = \tau) > 0$ and $\mathrm{P}(T > \tau) > 0$. Thus, the study ends at a time τ such that, on average, a positive fraction of individuals is still at risk. Second, $\mathrm{P}(T \leqslant C | Z) > 0$ almost surely; so, for any possible covariate pattern, the chance of observing a true event is positive. Finally, it is assumed that the support of Z is bounded, that the true regression coefficient, θ_0, belongs to the interior of the parameter space and that the covariance matrix of Z is positive definite.

The maximum likelihood estimator of η, $\hat{\eta}$, is a non-decreasing step function with support points at the observed event time. Consistency of $\hat{\eta}$ is relative to the supremum norm $\|\eta\|_\infty = \sup_{y \in [0,\tau]} |\eta(y)|$.

In order to define an approximately least favourable submodel, we calculate the score functions for θ and η. The score function for θ is given by

$$\dot{\ell}_{\theta,\eta}(x) = -z\left(1 - \frac{e^{-z^T\theta}}{\eta(y) + e^{-z^T\theta}} - \frac{\delta e^{-z^T\theta}}{\eta(y-) + e^{-z^T\theta}}\right).$$

The score operator for η in the direction of h (an arbitrary bounded function) is

$$A_{\theta,\eta}h(x) = \delta h(y) - \frac{\int_0^y h\,d\eta}{\eta(y) + e^{-z^T\theta}} - \frac{\delta\int_0^{y-} h\,d\eta}{\eta(y-) + e^{-z^T\theta}}.$$

This score operator is a linear operator from $L_2(\eta)$ to $L_2(P_{\theta,\eta})$. Let $A^*_{\theta,\eta}$ denote its adjoint. After some calculation we obtain

$$A^*_{\theta,\eta}A_{\theta,\eta}h(u) = h(u)P_{\theta,\eta}\left[\frac{I\{y \geqslant u\}}{\eta(y) + e^{-z^T\theta}} + \frac{\delta I\{y > u\}}{\eta(y-) + e^{-z^T\theta}}\right]$$

$$- P_{\theta,\eta}\left[\frac{I\{y \geqslant u\}\int_0^y h\,d\eta}{(\eta(y) + e^{-z^T\theta})^2} + \frac{\delta I\{y > u\}\int_0^{y-} h\,d\eta}{(\eta(y-) + e^{-z^T\theta_0})^2}\right],$$

$$A^*_{\theta,\eta}\dot{\ell}_{\theta,\eta} = P_{\theta,\eta}\left[\left(\frac{I\{y \geqslant u\}}{(\eta(y) + e^{-z^T\theta})^2} + \frac{\delta I\{y > u\}}{(\eta(y-) + e^{-z^T\theta})^2}\right)e^{-z^T\theta}z\right].$$

(These equations are most easily established in this form by differentiating the two identities $P_0\dot{\ell}_{\theta_0,\eta} \equiv 0$ and $P_0 A_{\theta_0,\eta}h \equiv 0$ with respect to η under the expectation P_0, or by calculating the variance of the score function as in Murphy *et al.* 1997.) The first equation gives the information operator for the nuisance parameter η when θ is known. This is shown to be continuously invertible on the space of functions of bounded variation on $[0, \tau]$ in Lemma 4.3 of Murphy *et al.* (1997). Thus, we can define

$$h_0 = (A^*_{\theta_0,\eta_0}A_{\theta_0,\eta_0})^{-1}A^*_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0},$$

$$d\boldsymbol{\eta}_t(\theta, \eta) = (1 + (\theta - t)^T h_0)d\eta,$$

$$\dot{\ell}(t, \theta, \eta) = \log \text{lik}(t, \boldsymbol{\eta}_t(\theta, \eta)).$$

Then (2.3) holds, with the efficient score function for estimation of θ given by

$$\tilde{\dot{\ell}}_0(x) = \dot{\ell}_{\theta_0,\eta_0}(x) - A_{\theta_0,\eta_0}h_0(x).$$

See equation (4.12) of Murphy *et al.* (1997) for a verification of (2.1) with the above $\tilde{\dot{\ell}}_0$ and $\tilde{I}_0$ the variance of $\tilde{\dot{\ell}}_0$.

Let $\hat{\eta}_\theta$ be the maximizer of the log-likelihood for a fixed θ. We must verify that if $\tilde{\theta} \xrightarrow{P} \theta_0$, then $\|\eta_{\tilde{\theta}} - \eta_0\|_\infty \xrightarrow{P} 0$. To do this, restrict attention to a subsequence of n for which the convergence of $\tilde{\theta}$ is almost sure. Then a similar proof to the proof of Theorem 2.2 in Murphy *et al.* (1997) can be employed. Replace $\hat{\eta}$, θ_0 and $\hat{\theta}$ in their equations by $\hat{\eta}_{\tilde{\theta}}$, $\tilde{\theta}$ and

$\tilde{\theta}$, respectively. This proof implies that $\|\eta_{\tilde{\theta}} - \eta_0\|_\infty$ converges almost surely to zero along the subsequence. Since for any sequence of n such a subsequence can be found, we have convergence in probability.

Next, we employ Lemma 2.2 to verify (2.4) and (2.5). The function $\dot{\ell}$ is given by

$$\dot{\ell}(t, \theta, \eta)(x) = \ell_{t,\boldsymbol{\eta}_t(\theta,\eta)}(x) - A_{t,\boldsymbol{\eta}_t(\theta,\eta)}\left(\frac{h_0(x)}{1 + (\theta - t)^{\mathrm{T}} h_0(y)}\right)(x).$$

The set of all functions of the type $x \mapsto \dot{\ell}(t, \theta, \eta)(x)$ and $x \mapsto \ddot{\ell}(t, \theta, \eta)(x)$, with t and θ varying in a compact set in $\mathbb{R}^p$ and η varying in the set of non-decreasing functions with $\eta(\tau) \leqslant 2\eta_0(\tau)$, is Donsker and uniformly bounded. This can be seen by noting that the above functions can be written as a Lipschitz function of members of uniformly bounded Donsker classes and next employing Theorem 2.20.6 in van der Vaart and Wellner (1996). Note that $\int_0^y h \, \mathrm{d}(\boldsymbol{\eta}_t(\theta, \eta) - \eta_0)$ is uniformly bounded by a constant times the product of the variation of h and $\|\boldsymbol{\eta}_t(\theta, \eta) - \eta_0\|_\infty$. As a result, the maps $(t, \theta, \eta) \mapsto \dot{\ell}(t, \theta, \eta)(x)$ and $(t, \theta, \eta) \mapsto \ddot{\ell}(t, \theta, \eta)(x)$ are continuous at $(\theta_0, \theta_0, \eta_0)$ relative to the uniform topology on η. Thus, an application of Lemma 2.2 serves to verify (2.4) and (2.5).

To verify (2.6) in Theorem 2.1, we first derive a rate of convergence for the profile estimators $\hat{\eta}_\theta$ via Theorem 3.1. Define $\mathscr{H}$ to be the set of all functions h: $[0, \tau] \mapsto [0, 1]$ that are of variation bounded by 1. Define

$$W_{n1}(\theta, \eta) = \mathbb{P}_n \dot{\ell}_{\theta,\eta},$$

$$W_{n2}(\theta, \eta)h = \mathbb{P}_n A_{\theta,\eta} h, \qquad h \in \mathscr{H}.$$

Then $W_n(\theta, \eta) \in \mathbb{R}^p \times \ell^\infty(\mathscr{H})$. Since $\hat{\eta}_\theta$ maximizes the likelihood for fixed θ, we have that

$$W_{n2}(\theta, \hat{\eta}_\theta) = 0.$$

The expectation of W_n is given by

$$W_1(\theta, \eta) = P_0 \dot{\ell}_{\theta,\eta},$$

$$W_2(\theta, \eta)h = P_0 A_{\theta,\eta} h.$$

It is implicit in the proof Theorem 2.2 of Murphy *et al.* (1997) that the map W: $\mathbb{R}^p \times \operatorname{lin} H \mapsto \mathbb{R}^p \times \ell^\infty(\mathscr{H})$ is differentiable at (θ_0, η_0) with continuously invertible derivative $\dot{W}$ given by

$$(\theta - \theta_0, \eta - \eta_0) \mapsto \begin{pmatrix} \dot{W}_{11} & \dot{W}_{12} \\ \dot{W}_{21} & \dot{W}_{22} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ \eta - \eta_0 \end{pmatrix},$$

where

$$\dot{W}_{11}(\theta-\theta_0) = -P_0 \dot{\ell}_{\theta_0,\eta_0} \dot{\ell}^{\mathrm{T}}_{\theta_0,\eta_0}(\theta-\theta_0),$$

$$\dot{W}_{12}(\eta-\eta_0) = -\int A^*_{\theta_0,\eta_0} \dot{\ell}_{\theta_0,\eta_0}\, d(\eta-\eta_0),$$

$$\dot{W}_{21}(\theta-\theta_0)h = -P_0(A_{\theta_0,\eta_0}h)\dot{\ell}^{\mathrm{T}}_{\theta_0,\eta_0}(\theta-\theta_0),$$

$$\dot{W}_{22}(\eta-\eta_0)h = -\int A^*_{\theta_0,\eta_0}A_{\theta_0,\eta_0}\, hd(\eta-\eta_0).$$

Consequently, the $\dot{W}_2(0, \eta-\eta_0)$ in Theorem 3.1 is given by $\dot{W}_{22}(\eta-\eta_0)$, and $\|\hat{\eta}_{\tilde{\theta}}-\eta_0\|_\infty$ is of the order $\|\tilde{\theta}-\theta_0\| + n^{-1/2}$.

The left-hand side of (2.6) is equal to

$$\begin{aligned} P_0\dot{\ell}(\tilde{\theta}, \tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) &= W_1(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}) - W_2(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}})h_0 \\ &= \dot{W}_{11}(\tilde{\theta}-\theta_0) + \dot{W}_{12}(\hat{\eta}_{\tilde{\theta}}-\eta_0) - \dot{W}_{21}(\tilde{\theta}-\theta_0) - \dot{W}_{22}(\hat{\eta}_{\tilde{\theta}}-\eta_0)h_0 \\ &\quad + o_P(\|\tilde{\theta}-\theta_0\| + \|\hat{\eta}_{\tilde{\theta}}-\eta_0\|_\infty), \\ &= -\tilde{I}_0(\tilde{\theta}-\theta_0) + o_P(\|\tilde{\theta}-\theta_0\| + \|\hat{\eta}_{\tilde{\theta}}-\eta_0\|_\infty), \end{aligned}$$

by the definitions of $\dot{W}$ and h_0. This verifies (2.6).

6. Logistic regression with a missing covariate

The following model is considered by Roeder *et al.* (1996), who use the profile likelihood to set a confidence interval in a study of the effect of cholesterol on heart disease. The model is expressed in terms of a basic random vector (D, W, Z), whose distribution is described in the following way (our parametrization is slightly different from that of Roeder *et al.*): D is a logistic regression on $\exp Z$ with intercept γ and slope β. W is a linear regression on Z with intercept α_0 and slope α_1, and an $N(0, \sigma^2)$ error. Given Z, the variables D and W are independent. Z has a completely unspecified distribution η. The unknown parameters are $\theta = (\beta, \alpha_0, \alpha_1, \gamma, \sigma)$ ranging over $\Theta \subset \mathbb{R}^4 \times (0, \infty)$ and the distribution η of the regression variable with support contained in a known, compact interval $\mathcal{Z} \subset \mathbb{R}$. The likelihood for the vector (D, W, Z) takes the form $p_\theta(d, w|z)\,\mathrm{d}\eta(z)$, with ϕ denoting the standard normal density,

$$p_\theta(d, w|z) = \left(\frac{1}{1+\exp(-\gamma-\beta \mathrm{e}^z)}\right)^d \left(\frac{\exp(-\gamma-\beta \mathrm{e}^z)}{1+\exp(-\gamma-\beta \mathrm{e}^z)}\right)^{1-d} \frac{1}{\sigma}\phi\left(\frac{w-\alpha_0-\alpha_1 z}{\sigma}\right)$$

and $\mathrm{d}\eta$ denoting the density of η with respect to a dominating measure.

Roeder *et al.* (1996) and Murphy and van der Vaart (1996) consider both a prospective and retrospective (or case–control) model. In the prospective model we observe two independent random samples of sizes n_{C} and n_{R} from the distributions of (D, W, Z) and (D, W), respectively. (The indexes C and R are for 'complete' and 'reduced', respectively.)

In the terminology of Roeder *et al*. (1996), the covariate Z in a full observation (D, W, Z) is a 'golden standard', but, in view of the costs of measurement, for a selection of observations only the 'surrogate covariate' W is available. In their example W is the natural logarithm of total cholesterol, Z is the natural logarithm of LDL cholesterol, and we are interested in heart disease $D = 1$.

We shall consider the situation that the number of complete and reduced observations are of comparable magnitude. More precisely, the proof applies to the situation that the fraction n_C/n_R is bounded away from 0 and ∞. For simplicity of notation, we shall henceforth assume that $n_C = n_R$. Then the observations can be paired and the observations in the prospective model can be summarized as n i.i.d. copies of $X = (Y_C, Z_C, Y_R)$ from the density

$$x = (y_C, z_C, y_R) \mapsto p_\theta(y_C|z_C)\mathrm{d}\eta(z_C)\int p_\theta(y_R|z)\mathrm{d}\eta(z) =: p_\theta(y_C|z_C)\mathrm{d}\eta(z_C)p_\theta(y_R|\eta).$$

Here we denote the complete sample components by $Y_C = (D_C, W_C)$ and Z_C and the reduced sample components by $Y_R = (D_R, W_R)$. In the complete sample part of the likelihood we use an empirical likelihood with $\eta\{z\}$, the measure of the point $\{z\}$,

$$\mathrm{lik}(\theta, \eta)(x) = p_\theta(y_C|z_C)\eta\{z_C\}\int p_\theta(y_R|z)\mathrm{d}\eta(z).$$

We shall concentrate on the regression coefficient, β, considering both the remaining coordinates of θ and η as nuisance parameters. (Thus, the parameter θ in the general results should be replaced by β throughout this section.) Note that the assumption of a known support means that in the maximum likelihood estimation, η is constrained to have support contained in $\mathscr{Z}$. Assuming that F_0 is non-degenerate, Murphy and van der Vaart (1996) show that the maximum likelihood estimator $(\hat{\theta}, \hat{\eta})$ is asymptotically normal. Consistency of $\hat{\eta}$ is relative to the weak topology. Here we shall verify that the conditions of Theorem 2.1 are satisfied, so that the asymptotic variance of the sequence $\sqrt{n}(\hat{\beta} - \beta)$ can be consistently estimated by minus the inverse of the curvature of the profile likelihood function. Since only the prospective model falls under the i.i.d. set-up of this paper, we shall concentrate on this model. However, since the profile likelihoods of the prospective and retrospective models are algebraically identical, the result can be extended to the retrospective model, as is shown for the maximum likelihood estimator in Murphy and van der Vaart (1996).

We start by introducing a least favourable submodel. The score function for θ, $\dot{\ell}_{\theta,\eta}$, is the sum of the score functions for θ for the conditional density $p_\theta(y_C|z_C)$ and the mixture density $p_\theta(y_R|\eta)$, given by

$$\dot{\ell}_\theta(y_C|z_C) = \frac{\partial}{\partial\theta}\log p_\theta(y_C|z_C), \qquad \dot{\ell}_{\theta_0,\eta_0}(y_R) = \frac{\int \dot{\ell}_\theta(y_R|z)\,p_\theta(y_R|z)\mathrm{d}\eta(z)}{p_\theta(y_R|\eta)}.$$

Furthermore, the score operator for η in the direction h (a bounded function satisfying $\int h\mathrm{d}\eta = 0$) is

$$A_{\theta,\eta}h(x) = h(z_C) + B_{\theta,\eta}h(y_R) = h(z_C) + \frac{\int h(z)\, p_\theta(y_R|z)\mathrm{d}\eta(z)}{p_\theta(y_R|\eta)}.$$

The operator $B_{\theta,\eta}$: $L_2(\eta) \mapsto L_2(p_\theta(\cdot|\eta))$ is the score operator for the mixture part of the model. A version of the Hilbert space adjoint $B^*_{\theta,\eta}$ of this operator is given by

$$B^*_{\theta,\eta}g(z) = \int g(y_R)\, p_\theta(y_R|z)\mathrm{d}\mu(y_R).$$

The efficient information matrix for θ when η is unknown is given by

$$\tilde{I}_0 = P_0 \dot{\ell}_{\theta_0,\eta_0} \dot{\ell}^{\mathrm{T}}_{\theta_0,\eta_0} + P_0 \dot{\ell}_{\theta_0} \dot{\ell}^{\mathrm{T}}_{\theta_0} - P_0(A_{\theta_0,\eta_0}(I + B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0})^{-1} B^*_{\theta_0,\eta_0} \dot{\ell}_{\theta_0,\eta_0}) \dot{\ell}^{\mathrm{T}}_{\theta_0,\eta_0}.$$

As in the proportional odds model, the least favourable direction, h_0, for the estimation of θ in the presence of the unknown η is given by $(A^*_{\theta_0,\eta_0}A_{\theta_0,\eta_0})^{-1}A^*_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0}$; however, it is easily shown that $A^*_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0} = B^*_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0}$ and $A^*_{\theta_0,\eta_0}A_{\theta_0,\eta_0} = I + B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0}$. The latter is the information operator for η when θ is known; in Section 8 of Murphy and van der Vaart (1996) it is shown that this operator is continuously invertible on the space of Lipschitz continuous functions. Additionally partition θ into $\theta = (\beta, \theta_2)$, where $\theta_2 = (\alpha_0, \alpha_1, \gamma, \sigma^2)$, and partition $\tilde{I}_0$ for θ into four submatrices accordingly. Then,

$$a_0^{\mathrm{T}} = (1, -\tilde{I}_{0,12}(\tilde{I}_{0,22})^{-1}),$$

$$h_0 = (I + B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0})^{-1}B^*_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0},$$

$$\mathrm{d}\boldsymbol{\eta}_t(\theta,\eta) = (1 + (\beta - t)a_0^{\mathrm{T}}(h_0 - \eta h_0))\mathrm{d}\eta,$$

$$\boldsymbol{\theta}_t(\theta,\eta) = \theta + (t - \beta)a_0,$$

where $\eta h = \int h\mathrm{d}\eta$ and $\eta_0 h_0 = 0$. In their Section 5, Murphy and van der Vaart (1996) show that the function h_0 is bounded. Thus $\boldsymbol{\eta}_t(\theta,\eta)$ has a positive density with respect to η for every sufficiently small $|\beta - t|$ and hence defines an element of the parameter set for η. Now we use the least favourable path

$$t \mapsto (\boldsymbol{\theta}_t(\theta,\eta)_2, \boldsymbol{\eta}_t(\theta,\eta))$$

in the parameter space for the nuisance parameter (θ_2, η). This leads to $\ell(t, \theta, \eta) = \log \mathrm{lik}(\boldsymbol{\theta}_t(\theta,\eta), \boldsymbol{\eta}_t(\theta,\eta))$. This submodel is least favourable at (θ_0, η_0) in that (2.3) is satisfied in the form

$$\frac{\partial}{\partial t}|_{t=\beta_0}\ell(t, \theta_0, \eta_0) = a_0^{\mathrm{T}}\tilde{\ell}_0,$$

where

$$\tilde{\ell}_0(x) = \dot{\ell}_{\theta_0}(y_C|z_C) + \dot{\ell}_{\theta_0,\eta_0}(y_R) - A_{\theta_0,\eta_0}h_0(y_R).$$

The function $\tilde{\ell}_0$ is the efficient influence function for the parameter θ in the presence of the nuisance parameter η, while the function $a_0^{\mathrm{T}}\tilde{\ell}_0$ is the efficient score function for β in the presence of the nuisance parameter (θ_2, η), both evaluated at (θ_0, η_0). See Section 7 of

Murphy and van der Vaart (1996). For the present purpose, the relevant information is that (2.1) is satisfied for the maximum likelihood estimator $\hat{\beta}$ substituted for $\hat{\theta}$, $\tilde{\ell}_0$ equal to $a_0^{\mathrm{T}}\dot{\ell}_0$ and $\tilde{I}_0$ equal to $a_0^{\mathrm{T}}\tilde{I}_0 a_0 = \tilde{I}_{0,11} - \tilde{I}_{0,12}\tilde{I}_{0,22}^{-1}\tilde{I}_{0,21}$.

Let $(\hat{\theta}_{2,\beta}, \hat{\eta}_\beta)$ be the profile likelihood estimator for (θ_2, η) when β is given so that $\hat{\theta}_\beta = (\beta, \hat{\theta}_{2,\beta})$. The profile likelihood estimator $(\hat{\theta}_{\tilde{\beta}}, \hat{\eta}_{\tilde{\beta}})$ can be shown to be consistent for (θ_0, η_0) as $\tilde{\beta} \xrightarrow{\mathrm{P}} \beta_0$, by the same proof as used for the full maximum likelihood estimator in Murphy and van der Vaart (1996). (Replace β_0 by $\tilde{\beta}$, $\hat{\beta}$ by $\tilde{\beta}$ and $(\hat{\theta}_2, \hat{\eta})$ by $(\hat{\theta}_{2,\tilde{\beta}}, \hat{\eta}_{\tilde{\beta}})$.) It now suffices to verify the conditions of Lemma 2.2. By direct calculation, and with the abbreviations $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t(\theta, \eta)$ and $\boldsymbol{\eta}_t = \boldsymbol{\eta}_t(\theta, \eta)$,

$$\dot{\ell}(t, \theta, \eta) = a_0^{\mathrm{T}}\dot{\ell}_{\boldsymbol{\theta}_t}(y_{\mathrm{C}}|z_{\mathrm{C}}) + a_0^{\mathrm{T}}\dot{\ell}_{\boldsymbol{\theta}_t,\boldsymbol{\eta}_t}(y_{\mathrm{R}}) - a_0^{\mathrm{T}}A_{\boldsymbol{\theta}_t,\boldsymbol{\eta}_t}\left(\frac{h_0 - \boldsymbol{\eta}_t h_0}{1 + (\beta - t)a_0^{\mathrm{T}}(h_0 - \boldsymbol{\eta}_t h_0)}\right)(y_{\mathrm{R}}).$$

The class of functions $\dot{\ell}(t, \theta, \eta)$, with t varying in a neighbourhood of β_0 and (θ, η) varying in a neighbourhood of (θ_0, η_0), is shown to be Donsker in Section 4 of Murphy and van der Vaart (1996). That the class of second derivatives, $x \mapsto \ddot{\ell}(t, \theta, \eta)(x)$, is Glivenko–Cantelli follows by similar, but simpler, arguments.

To verify condition (2.6), we apply Theorem 3.1 to study the profile estimators $\hat{\eta}_\theta$. Let $\mathscr{H}$ be the set of measurable functions $h: \mathscr{Z} \mapsto [0, 1]$ that are uniformly Lipschitz. Let $W_n = (W_{n1}, W_{n2})$ be the element of $\mathbb{R}^5 \times \ell^{*\infty}(\mathscr{H})$ given by

$$W_{n1}(\theta, \eta) = \mathbb{P}_n(\dot{\ell}_\theta(y_{\mathrm{C}}|z_{\mathrm{C}}) + \dot{\ell}_{\theta,\eta}(y_{\mathrm{R}})),$$

$$W_{n2}(\theta, \eta)h = \mathbb{P}_n A_{\theta,\eta}h(x, z) - P_{\theta,\eta}A_{\theta,\eta}h.$$

The maximum likelihood estimators $(\hat{\theta}, \hat{\eta})$ are zeros of the maps W_n,

$$W_n(\hat{\theta}, \hat{\eta}) \equiv 0.$$

Similarly the profile maximum likelihood estimator, $(\hat{\theta}_\beta, \hat{\eta}_\beta)$, satisfies

$$W_{n1,2}(\hat{\theta}_\beta, \hat{\eta}_\beta) = 0, \qquad W_{n2}(\hat{\theta}_\beta, \hat{\eta}_\beta) \equiv 0.$$

We shall identify each probability measure η on $\mathscr{Z}$ with an element of $\ell^\infty(\mathscr{H})$ through $\eta h = \int h \mathrm{d}\eta$. Then W_n can be viewed as a map from the space $\mathbb{R}^5 \times \ell^\infty(\mathscr{H})$ into itself with domain the product of Θ and the set of probability measures in $\ell^\infty(\mathscr{H})$ under the given identification. The expectation of W_n under the true distribution, $P_0 = P_{\theta_0,\eta_0}$ is the element $W = (W_1, W_2)$ of $\mathbb{R}^5 \times \ell^\infty(\mathscr{H})$ given by

$$\begin{aligned} W_1(\theta, \eta) &= P_0(\dot{\ell}_\theta(y_{\mathrm{C}}|z_{\mathrm{C}}) + \dot{\ell}_{\theta,\eta}(y_{\mathrm{R}})), \\ W_2(\theta, \eta)h &= P_0 A_{\theta,\eta}h - P_{\theta,\eta}A_{\theta,\eta}h. \end{aligned} \tag{6.1}$$

With this choice of centring function, we have $W(\theta_0, \eta_0) = 0$.

Conditions (3.1) and (3.2) are verified in Section 4 of Murphy and van der Vaart (1996). Furthermore, by Lemma 5.1 in the same paper, the map W is differentiable at (θ_0, η_0), with continuously invertible derivative

$$(\theta-\theta_0,\,\eta-\eta_0)\mapsto\begin{pmatrix}\dot W_{11} & \dot W_{12}\\ \dot W_{21} & \dot W_{22}\end{pmatrix}\begin{pmatrix}\theta-\theta_0\\ \eta-\eta_0\end{pmatrix},\tag{6.2}$$

where

$$\dot W_{11}(\theta-\theta_0)=-(P_0\dot\ell_{\theta_0,\eta_0}\dot\ell_{\theta_0,\eta_0}^{\mathrm T}+P_0\dot\ell_{\theta_0}\dot\ell_{\theta_0}^{\mathrm T})(\theta-\theta_0),$$

$$\dot W_{12}(\eta-\eta_0)=-\int B^*_{\theta_0,\eta_0}\dot\ell_{\theta_0,\eta_0}\,\mathrm d(\eta-\eta_0),$$

$$\dot W_{21}(\theta-\theta_0)h=-P_0A_{\theta_0,\eta_0}h\dot\ell_{\theta_0,\eta_0}^{\mathrm T}(\theta-\theta_0),$$

$$\dot W_{22}(\eta-\eta_0)h=-\int(I+B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0})h\,\mathrm d(\eta-\eta_0).$$

The above, combined with consistency of the profile maximum likelihood estimator, implies that $\|\hat\theta_{\tilde\beta}-\theta_0\|+\|\hat\eta_{\tilde\beta}-\eta_0\|_{\mathscr H}$ is of the order $\|\tilde\beta-\beta_0\|+n^{-1/2}$ by Theorem 3.1.

The left-hand side of (2.6) is equal to

$$\begin{aligned}P_0\dot\ell(\tilde\beta,\hat\theta_{\tilde\beta},\hat\eta_{\tilde\beta})&=a_0^{\mathrm T}(W_1(\hat\theta_{\tilde\beta},\hat\eta_{\tilde\beta})-W_2(\hat\theta_{\tilde\beta},\hat\eta_{\tilde\beta})h_0)\\&=a_0^{\mathrm T}(\dot W_1(\hat\theta_{\tilde\beta}-\theta_0,\hat\eta_{\tilde\beta}-\eta_0)-\dot W_2(\hat\theta_{\tilde\beta}-\theta_0,\hat\eta_{\tilde\beta}-\eta_0)h_0)\\&\quad+o_P(\|\hat\theta_{\tilde\beta}-\theta_0\|+\|\hat\eta_{\tilde\beta}-\eta_0\|_{\mathscr H}),\\&=-a_0^{\mathrm T}\tilde I_0(\hat\theta_{\tilde\beta}-\theta_0)+o_P(\|\hat\theta_{\tilde\beta}-\theta_0\|+\|\hat\eta_{\tilde\beta}-\eta_0\|_{\mathscr H})\\&=-(\tilde I_{0,11}-\tilde I_{0,12}\tilde I_{0,22}^{-1}\tilde I_{0,21})(\tilde\beta-\beta_0)+o_P(\|\hat\theta_{\tilde\beta}-\theta_0\|+\|\hat\eta_{\tilde\beta}-\eta_0\|_{\mathscr H}),\end{aligned}$$

by the definitions of $\dot W$, h_0 and a_0. This verifies (2.6), because $\tilde I_{0,11}-\tilde I_{0,12}\tilde I_{0,22}^{-1}\tilde I_{0,21}$ is the efficient information for estimating β in the presence of the nuisance parameter (θ_2,η).

7. Semi-parametric penalized logistic regression

In this model the observations are n i.i.d. copies of $X=(Y,W,Z)$ for a 0–1 variable Y such that

$$P(Y=1|W,Z)=F(\theta W+\eta(Z)),$$

where $F(u)=\mathrm e^u/(1+\mathrm e^u)$ is the logistic distribution. Both W and Z are assumed to have bounded support, which we take to be a subset of $[0,1]^2$. The unknown parameters are the scalar θ, and η, a function in the Sobolev class of functions on $[0,1]$ whose $(k-1)$th derivative exists and is absolutely continuous with $J(\eta)<\infty$, where

$$J^2(\eta)=\int_0^1(\eta^{(\mathrm k)}(z))^2\,\mathrm dz.$$

Here, $k \geqslant 1$ is a fixed integer and $\eta^{(j)}$ is the jth derivative of η with respect to z. Mammen and van de Geer (1997) study the estimators for θ and η obtained by maximizing the penalized log-likelihood, given by

$$\mathbb{P}_n \log p_{\theta,\eta} - \tilde{\lambda}^2 J^2(\eta),$$

where $\tilde{\lambda}$ is a 'smoothing parameter' and

$$p_{\theta,\eta}(x) = F(\theta w + \eta(z))^y (1 - F(\theta w + \eta(z))^{1-y} f^{W,Z}(w, z).$$

The smoothing parameter may depend on the data and hence can, for instance, be chosen by cross-validation. The estimator $\hat{\eta}$ of η is a weighted sum of a finite number of basis functions determined by $\{Z_1, \ldots, Z_n\}$ (O'Sullivan *et al.* 1986).

For the purpose of (first-order efficient) inference concerning θ, there is considerable freedom in the choice of the smoothing parameter. Following Mammen and van de Geer (1997), we assume that

$$\tilde{\lambda}^2 = o_P(n^{-1/2}) \quad \text{and} \quad \tilde{\lambda}^{-1} = O_P(n^{k/(2k+1)}). \tag{7.1}$$

To ensure the identifiability of the parameters we assume that $\mathrm{E}_0 \operatorname{var}(W|Z)$ is positive, and that the support of Z (the smallest closed set with mass 1) contains at least k distinct points in $[0, 1]$. Finally, we assume that the function h_0 given by (7.2) has a version with $J(h_0) < \infty$.

Under the above assumptions, the arguments of Mammen and van de Geer (1997) can be refined to prove that $\|\hat{\eta} - \eta_0\|_2 = O_P(\tilde{\lambda})$, where $\|a\|_2 = E_0 a^2(Z)$, and that $\hat{\theta}$ is asymptotically efficient in the sense of (2.1).

Our purpose is to show that the second derivative of the profile penalized log-likelihood yields a consistent estimator of minus the inverse of the asymptotic variance of $\hat{\theta}$. To do this, we follow the general scheme of the paper, with the log-likelihood equal to the penalized log-likelihood

$$\log \operatorname{lik}(\theta, \eta)(x) = \log p_{\theta,\eta}(x) - \tilde{\lambda}^2 J^2(\eta).$$

Assumption (7.1) ensures that even though this function depends on n and possibly on the observations through $\tilde{\lambda}$, the arguments are unaffected, in the sense that Theorem 2.1 and its proof go through with minor notational adaptations.

The score function for θ takes the form

$$\dot{\ell}_{\theta,\eta}(x) = (y - F(\theta w + \eta(z)))w.$$

As in the previous examples, for h a function with $J(h) < \infty$, we may differentiate the log-likelihood (the true one, with $\tilde{\lambda} = 0$) along the submodel $\eta_t = \eta + th$ at $t = 0$ to obtain a score function for η, given by

$$A_{\theta,\eta} h(x) = (y - F(\theta w + \eta(z)))h(z).$$

The efficient score function is given by

$$\tilde{\dot{\ell}}_0 = \dot{\ell}_{\theta_0,\eta_0} - A_{\theta_0,\eta_0} h_0 = (y - F(\theta_0 w + \eta_0(z)))(w - h_0(z)).$$

Here h_0 minimizes the distance $P_0(\dot{\ell}_{\theta_0,\eta_0} - A_{\theta_0,\eta_0} h_0)^2$, and is given by

$$h_0(z) = \frac{E_0[Wf(\theta_0 W + \eta_0(Z))|Z = z]}{E_0[f(\theta_0)W + \eta_0(Z))|Z = z]}. \tag{7.2}$$

(Note that $F(1 - F) = f$, the derivative of F.) Thus, we define as least favourable submodel

$$\boldsymbol{\eta}_t(\theta, \eta) = \eta + (\theta - t)h_0,$$

$$\ell(t, \theta, \eta) = \log \text{lik}(t, \boldsymbol{\eta}_t(\theta, \eta)).$$

Differentiation of $\ell(t, \theta_0, \eta_0)$ with respect to t and evaluation at $t = \theta_0$ and $\tilde{\lambda} = 0$ yields the efficient score function $\tilde{\ell}_0$.

Let $\hat{\eta}_\theta$ be the maximizer of the penalized log-likelihood for a fixed θ and the same stochastic smoothing parameter $\tilde{\lambda}$ as the one used to arrive at the estimator $(\hat{\theta}, \hat{\eta})$. Recall that $\hat{\psi}_\theta = (\theta, \hat{\eta}_\theta)$. In Lemmas 7.1–7.4 we prove that

$$\tilde{\lambda} J(\hat{\eta}_{\tilde{\theta}}) + \| |\hat{\eta}_{\tilde{\theta}} - \eta_0| \wedge 1 \|_2 = O_P(\tilde{\lambda} + \|\tilde{\theta} - \theta_0\|). \tag{7.3}$$

We shall verify (2.4′)–(2.5′) and (2.6), where we take $\overline{\psi} \xrightarrow{P} \psi_0$ to mean $\overline{\theta} \xrightarrow{P} \theta_0$, and $\| |\overline{\eta} - \eta_0| \wedge 1 \|_2 \xrightarrow{P} 0$. We have

$$\dot{\ell}(t, \psi)(x) = (y - F(tw + \eta(z) + (\theta - t)h_0(z)))(w - h_0(z))$$

$$+ 2\tilde{\lambda}^2 \int_0^1 (\eta + (\theta - t)h_0)^{(k)}(z)(h_0)^{(k)}(z)\, dz,$$

$$\ddot{\ell}(t, \psi) = -f(tw + \eta(z) + (\theta - t)h_0(z))(w - h_0(z))^2 - 2\tilde{\lambda}^2 J^2(h_0).$$

The penalty terms do not play a role in the verification of (2.4′)–(2.5′) and (2.6), since $\tilde{\lambda}^2 = o_P(n^{-1/2})$ by assumption and

$$\tilde{\lambda}^2 J(\hat{\eta}_{\tilde{\theta}}) = O_P(\tilde{\lambda}^2 + \tilde{\lambda}|\tilde{\theta} - \theta_0|).$$

Therefore, without loss of generality we may set $\tilde{\lambda} = 0$ for this part of the argument. If $(\overline{\theta}, \overline{\psi}) \to (\theta_0, \psi_0)$, then, in view of the continuity of F and f, $\dot{\ell}(\overline{\theta}, \overline{\psi})$ converges a.e. to $\dot{\ell}_0$ and $\ddot{\ell}(\overline{\theta}, \overline{\psi})(x)$ converges a.e. to $-f(\theta_0 w + \eta_0(z))(w - h_0(z))^2$, at least along subsequences. By the dominated convergence theorem, $-P_0\ddot{\ell}(\overline{\theta}, \overline{\psi})$ converges to the efficient information

$$\tilde{I}_0 = P_0 f(\theta_0 w + \eta_0(z))(w - h_0(z))^2.$$

Thus, for (2.4)–(2.5) it certainly suffices to show that the classes of functions $\dot{\ell}(t, \psi)$ and $\ddot{\ell}(t, \psi)$, respectively, with (t, ψ) ranging over a neighbourhood of (θ_0, ψ_0), are P_0-Donsker and P_0-Glivenko–Cantelli with square-integrable and integrable envelope functions, respectively. If h_n in (1.3) is chosen such that $h_n = O_P(\tilde{\lambda})$, we have that $J(\hat{\eta}_{\tilde{\theta}}) = O_P(1)$ by (7.3). Since it suffices to prove (2.4)–(2.5) for $\overline{\psi}$ of the form $\overline{\psi} = (\overline{\theta}, \hat{\eta}_{\overline{\theta}})$ with $|\overline{\theta} - \theta_0| \leqslant |\hat{\theta} + h_n - \theta_0|$, we may then assume a priori that $J(\overline{\eta}) = O_P(1)$. Under the condition that $J(\eta)$ is uniformly bounded, the classes of functions $\dot{\ell}(t, \psi)$ and $\ddot{\ell}(t, \psi)$ can be seen to be Donsker and Glivenko–Cantelli by entropy calculations as in Lemma 7.2, and the uniform entropy central limit theorem and uniform entropy Glivenko–Cantelli theorem, respectively – see, for example, Theorems 2.5.2 and 2.4.3 in van der Vaart and Wellner

(1996). Without the condition that $h_n = O_P(\tilde{\lambda})$, we must refine the argument and can verify (2.4′)–(2.5′) rather than (2.4)–(2.5). This is done in Lemma 7.5 below.

In order to verify (2.6) we follow the second intuitive justification given in Section 2. We may still assume that $\tilde{\lambda} = 0$. By the formula for $\dot{\ell}$, with $\hat{g}_{\tilde{\theta}}(w, z) = \tilde{\theta}w + \hat{\eta}_{\tilde{\theta}}(z)$ and $g_0(w, z) = \theta_0 w + \eta_0(z)$,

$$\begin{aligned} P_0\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) &= P_0(y - F(\hat{g}_{\tilde{\theta}}))(w - h_0(z)) \\ &= P_0(F(g_0) - F(\hat{g}_{\tilde{\theta}}))(w - h_0(z)). \end{aligned}$$

By Taylor's formula,

$$|F(g) - F(g_0) - f(g_0)(g - g_0)| \leqslant \tfrac{1}{2}\|f'\|_\infty |g - g_0|^2.$$

The function $f(g_0)^{-1}(F(g_0) - F(g) - f(g_0)(\theta - \theta_0)w)$ is uniformly bounded. Consequently, for a sufficiently large constant M,

$$|F(g) - F(g_0) - f(g_0)(\theta - \theta_0)w - f(g_0)[\eta - \eta_0]_M| \leqslant \|f'\|_\infty(|\theta - \theta_0|^2 + |\eta - \eta_0|^2), \quad (7.4)$$

where $[\eta]_M$ is η truncated to the interval $[-M, M]$. Since the left-hand side is bounded, the right-hand side can be truncated at a sufficiently large constant and inequality (7.4) will still hold. Since $P_0 f(g_0)a(z)(w - h_0(z))$ is zero for every a,

$$\begin{aligned} P_0\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) = &- P_0(F(\hat{g}_{\tilde{\theta}}) - F(g_0) - f(g_0)[(\tilde{\theta} - \theta_0)w + [\hat{\eta}_{\tilde{\theta}} - \eta_0]_M])(w - h_0(z)) \\ &- (\tilde{\theta} - \theta_0)P_0 f(g_0)w(w - h_0(z)). \end{aligned}$$

The first term on the right is bounded by a multiple of $|\tilde{\theta} - \theta_0|^2 + P_0[(\hat{\eta}_{\tilde{\theta}} - \eta_0)^2 \wedge 1]$. This is negligible to the desired order by (7.3). The second term is equal to $-(\tilde{\theta} - \theta_0)\tilde{I}_0$.

We finish this section with a careful proof of the rate of convergence (7.3). For a function g of (y, w, z) let $\|g\|_2$ denote the square of $P_0 g^2(Y, W, Z)$. This norm does not depend on the parameters (θ, η) and can be taken as fixed in the following.

Lemma 7.1. *Let (7.1) hold and assume that $P_0 \operatorname{var}(W|Z)$ is positive. Furthermore, suppose that the support of Z contains at least k distinct points. If $|\tilde{\theta} - \theta_0| \xrightarrow{P} 0$, then*

$$\|p_{\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}} - p_{\theta_0, \eta_0}\|_2 + \tilde{\lambda}J(\hat{\eta}_{\tilde{\theta}}) = O_P(\tilde{\lambda} + |\tilde{\theta} - \theta_0|).$$

This implies (7.3). If $\tilde{\theta} - \theta_0 = O_P(\tilde{\lambda})$, then this also implies that $J(\hat{\eta}_{\tilde{\theta}}) = O_P(1)$, and next that $\|\hat{\eta}_{\tilde{\theta}} - \eta_0\|_2 = O_P(\tilde{\lambda} + |\tilde{\theta} - \theta_0|)$.

Proof. We apply Theorem 3.2, where we let the θ of this theorem include the smoothing parameter λ, and where

$$m_{\theta,\lambda,\eta} = \log \frac{p_{\theta,\eta} + p_{\theta,\eta_0}}{2p_{\theta,\eta_0}} - \tfrac{1}{2}\lambda^2(J^2(\eta) - J^2(\eta_0)).$$

Within this context we write $\hat{\eta}_{\theta,\lambda}$ rather than $\hat{\eta}_\theta$. By the concavity of the logarithmic function and the definition of $\hat{\eta}_{\theta,\lambda}$,

$$\mathbb{P}_n m_{\theta,\lambda,\hat{\eta}_{\theta,\lambda}} \geqslant \tfrac{1}{2}\mathbb{P}_n \log \frac{p_{\theta,\hat{\eta}_{\theta,\lambda}}}{p_{\theta,\eta_0}} - \tfrac{1}{2}\lambda^2(J^2(\hat{\eta}_{\theta,\lambda}) - J^2(\eta_0)) \geqslant 0 = \mathbb{P}_n m_{\theta,\lambda,\eta_0}.$$

In view of (7.1), we may restrict (θ, λ) a priori to the set $\Theta_n = \{|\theta - \theta_0| < \varepsilon, \lambda \geqslant \lambda\}$ for a small $\varepsilon > 0$ and for λ a sufficiently large multiple of $n^{-k/(2k+1)}$. Suppose that it can be shown that $\|\hat{\eta}_{\tilde{\theta},\tilde{\lambda}}\|_\infty = O_P(J(\hat{\eta}_{\tilde{\theta},\tilde{\lambda}}) + 1)$. Then we may also restrict η to the set $H_n = \{\eta: \|\eta\|_\infty \leqslant CJ(\eta) + C\}$ for a large constant C. (Strictly speaking, we must let $\lambda^{k/(2k+1)}$ and C tend to infinity, but there is no loss of generality in giving the proof for a fixed but arbitrary large constant only.)

The function $p_{\theta_0,\eta_0}(x)/f^{W,Z}(w, z)$ is bounded away from zero and infinity uniformly in x. Therefore, by continuity $p_{\theta,\eta_0}(x)/f^{W}, Z(w, z)$ is bounded away from zero and infinity, uniformly in x and θ varying over a neighbourhood of θ_0. This implies that $m_{\theta,0,\eta}(x)$ is uniformly bounded in θ, η and x. Since $\mathbb{G}_n m_{\theta,\lambda,\eta} = \mathbb{G}_n m_{\theta,0,\eta}$, this shows that Lemma 3.3 can be applied to verify (3.6).

By the well-known inequality relating Kullback–Leibler divergence and Hellinger distance (see, for example, the proof of Lemma 5.35 in van der Vaart 1998),

$$\begin{aligned} P_0(m_{\theta,\lambda,\eta} - m_{\theta,\lambda,\eta_0}) &= P_0(m_{\theta,\lambda,\eta} - m_{\theta_0,\lambda,\eta_0}) - P_0(m_{\theta,\lambda,\eta_0} - m_{\theta_0,\lambda,\eta_0}) \\ &\leqslant -h^2(p_{\theta,\eta}, p_{\theta_0,\eta_0}) - \lambda^2 J^2(\eta) + \|\theta - \theta_0\|^2 + \lambda^2 \\ &\leqslant -\|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2^2 - \lambda^2 J^2(\eta) + |\theta - \theta_0|^2 + \lambda^2, \end{aligned}$$

since $p_{\theta_0,\eta_0}/f^{W,Z}$ is bounded away from zero. This suggests the choice of

$$d^2_{\theta,\lambda}(\eta, \eta_0) = \|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2^2 + \lambda^2 J^2(\eta)$$

and the Euclidean norm for (θ, λ). Since the derivative of the function $p \mapsto \log(p + p_0)$ is bounded, uniformly in p_0 that are bounded away from zero,

$$P_0(m_{\theta,0,\eta} - m_{\theta_0,0,\eta_0})^2 \leqslant \|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2^2 + |\theta - \theta_0|^2.$$

If $(\theta, \lambda) \in \Theta_n$ and $d_{\theta,\lambda}(\eta, \eta_0) < \delta$, then $\|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2 < \delta$ and $J(\eta) < \delta/\lambda$, and hence $\|\eta\|_\infty \leqslant \delta/\lambda$ by our working assumption that $\eta \in H_n$. By a result of Birman and Solomjak (1967),

$$\log N(\varepsilon, \{\eta: J(\eta) \leqslant M, \|\eta\|_\infty \leqslant M\}, \|\cdot\|_\infty) \leqslant \left(\frac{M}{\varepsilon}\right)^{1/k}.$$

The class of functions $w \mapsto w\theta$ for θ varying over a compact has polynomial bracketing numbers. Since the transformation $(\theta, \eta) \mapsto m_{\theta,0,\eta}$ is Lipschitz and essentially monotone, it follows that

$$\log N_{[]}(\varepsilon, \{m_{\theta,0,\eta}: (\theta, \lambda) \in \Theta_n, \eta \in H_n, d_{\theta,\lambda}(\eta, \eta_0) \leqslant \delta\}, L_2(P_0)) \leqslant \left(\frac{1 + \delta/\lambda}{\varepsilon}\right)^{1/k}.$$

Thus, by Lemma 3.3 condition (3.6) is satisfied with ϕ_n and $J = J_n$ related as in Lemma 3.3 and

$$J_n(\delta) \leqslant \left(1 + \frac{\delta}{\lambda_n}\right)^{1/2k} \delta^{1-1/2k}.$$

By Theorem 3.2 we obtain that

$$d_{\tilde{\theta},\tilde{\lambda}}(\hat{\eta}_{\tilde{\theta},\tilde{\lambda}}, \eta_0) = O_P(\tilde{\lambda} + |\tilde{\theta} - \theta_0| + (n\lambda_n^{1/k})^{-1/2} + n^{-k/(2k+1)}) = O_P(\tilde{\lambda} + |\tilde{\theta} - \theta_0|).$$

This is the first assertion of the lemma. The other assertions follow by Lemma 7.4.

To show that $\|\hat{\eta}_{\tilde{\theta},\tilde{\lambda}}\|_\infty = O_P(J(\hat{\eta}_{\tilde{\theta},\tilde{\lambda}}) + 1)$ we apply Theorem 3.2 in a crude manner, with a different maximal inequality. We still assume that $(\theta, \lambda) \in \Theta_n$, but drop the assumption that $\eta \in H_n$. By Lemma 7.2, and a maximal inequality due to Kim and Pollard (1990) (see Theorem 2.14.1 of van der Vaart and Wellner 1996), condition (3.6) is satisfied for

$$J_n(\delta) \leqslant \left(1 + \frac{\delta}{\lambda_n}\right)^{1/2k}.$$

In view of Theorem 3.2, this means that $d_{\tilde{\theta},\tilde{\lambda}}(\hat{\eta}_{\tilde{\theta},\tilde{\lambda}}, \eta_0) = O_P(\delta_n)$ for any $\delta_n \downarrow 0$ such that $\delta_n \geqslant (n^k\lambda_n)^{-1/(4k-1)}$. In particular, $d_{\tilde{\theta},\tilde{\lambda}}(\hat{\eta}_{\tilde{\theta},\tilde{\lambda}}, \eta_0) \xrightarrow{P} 0$. The result then follows from Lemma 7.3(i). □

Lemma 7.2.

$$\sup_Q \log N(\varepsilon, \{p_{\theta,\eta}: \theta \in \mathbb{R}, J(\eta) \leqslant M\}, L_2(Q)) \leqslant \left(\frac{1+M}{\varepsilon}\right)^{1/k}.$$

Proof. The functions $p_{\theta,\eta}$ are transformations of the functions $F(\theta w + \eta(z))$ (and the 0–1 variable y). It suffices to give the same bound for the entropy of the latter collection of functions.

For every η with $J(\eta) < \infty$, there exists a polynomial $\tilde{\eta}$ of degree at most $k-1$ such that $\|\eta - \tilde{\eta}\|_\infty \leqslant J(\eta)$. (By the Cauchy–Schwarz inequality $|\eta^{(k-1)}(z) - \eta^{(k-1)}(0)| \leqslant J(\eta)$ for every z. Next integrate this $k-1$ times.) For a fixed function η, let $\mathscr{F}_\eta$ be the set of all functions $F(\theta w + p(z) + \eta(z))$ with θ ranging over $\mathbb{R}$ and p ranging over the set of all polynomials of degree at most $k-1$. Then our set of functions is the union of all $\mathscr{F}_\eta$ with η ranging over the set H of all functions with $J(\eta) \leqslant M$ and $\|\eta\|_\infty \leqslant M$.

By Birman and Solomjak (1967) the $\|\cdot\|_\infty$-entropy of the class H is of the order $(1/\varepsilon)^{1/k}$.

Each class $\mathscr{F}_\eta$ is Vapnik–Chervonenkis of index at most $k+3$ and uniformly bounded. (See, for example, Lemmas 2.6.15 and 2.6.18(viii) of van der Vaart and Wellner 1996.) Thus its covering numbers are polynomial.

We can construct a net over $\cup_{\eta \in H} \mathscr{F}_\eta$ by first choosing an ε-net over the set H, and next, for every η in the net, choosing an ε-net over $\mathscr{F}_\eta$. The total number of functions will be bounded as in the lemma, and will constitute an ε'-net over the functions of interest, for ε' a fixed multiple of ε. □

Lemma 7.3. *(i) For every sufficiently small $\delta > 0$ there exists a constant C depending only on P_0 such that $\|\eta\|_\infty \leqslant C(J(\eta) + 1)$ whenever $|\theta - \theta_0| < \delta$ and $\|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2 < \delta$.*

(ii) For any η we have $\|\eta\|_\infty \leqslant J(\eta) + \|\eta\|_2$.

Proof. (i) By assumption there exist disjoint intervals $[a_i, b_i]$ such that $F_Z(b_i) - F_Z(a_i) > 0$ for each $i = 1, \ldots, k$. If $\|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2 < \delta$, then, for every $0 < a < b < 1$,

$$\int_a^b \int (F(\theta w + \eta(z)) - F(\theta_0 w + \eta_0(z)))^2 F_{W|Z}(\mathrm{d}w|z) F_Z(\mathrm{d}z) < \delta^2.$$

Therefore, there exist $z_i \in (a_i, b_i]$ for which

$$\int (F(\theta w + \eta(z_i)) - F(\theta_0 w + \eta_0(z_i)))^2 F_{W|Z}(\mathrm{d}w|z_i)(F_Z(b_i) - F_Z(a_i)) < \delta^2$$

for each $i = 1, \ldots, k$. Next for each z_i there exists a w_i which satisfies

$$(F(\theta w_i + \eta(z_i)) - F(\theta_0 w_i + \eta_0(z_i)))^2 \leqslant \frac{\delta^2}{F_Z(b_i) - F_Z(a_i)}.$$

Since $F(\theta_0 w_i + \eta_0(z_i))$ is bounded away from zero and one, this implies that, for sufficiently small $\delta > 0$, the numbers $F(\theta w_i + \eta(z_i))$ are bounded away from zero and one as well, whence the numbers $\theta w_i + \eta(z_i)$ are uniformly bounded by a constant that depends on δ and (θ_0, η_0) only. Since $\|\theta - \theta_0\| < \delta$, this in turn implies that $|\eta(z_i)| \leqslant K_\delta$ for some constant K_δ.

For every η there exists a polynomial $\tilde{\eta}$ of degree smaller than $k-1$ such that $\|\eta - \tilde{\eta}\|_\infty \leqslant J(\eta)$. See the proof of Lemma 7.2. It follows that the numbers $|\tilde{\eta}(z_i)|$ are bounded by $K_\delta + J(\eta)$. If $\tilde{\eta}(z) = \sum a_j z^j = (1, z, \ldots, z^{k-1}) \cdot a$, then

$$\|a\| \leqslant \left\| \begin{pmatrix} 1 & z_1 & \cdots & z_1^{k-1} \\ \vdots & \vdots & & \vdots \\ 1 & z_k & \cdots & z_k^{k-1} \end{pmatrix}^{-1} \right\| \left\| \begin{pmatrix} \tilde{\eta}(z_1) \\ \vdots \\ \tilde{\eta}(z_k) \end{pmatrix} \right\| \leqslant L\sqrt{k}(K_\delta + J(\eta)),$$

where L can be chosen to correspond to the worst possible choice of the points $z_i \in (a_i, b_i]$. Consequently, $\|\tilde{\eta}\|_\infty \leqslant \|a\| \leqslant K_\delta + J(\eta)$, and $\|\eta\|_\infty$ is bounded similarly.

(ii) Since $\|\eta - \tilde{\eta}\|_\infty \leqslant J(\eta)$, we have $\|\tilde{\eta}\|_2 \leqslant J(\eta) + \|\eta\|_2$. By the non-singularity of the matrix $P_0 \phi \phi^{\mathrm{T}}$, for $\phi = (1, z, \ldots, z^{k-1})$, this implies that $\|a\| \leqslant J(\eta) + \|\eta\|_2$, whence $\|\tilde{\eta}\|_\infty$ is bounded similarly. □

Lemma 7.4. *(i)* $\|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2 \geqslant (|\theta - \theta_0| \wedge 1 + \||\eta - \eta_0| \wedge 1\|_2) \wedge 1$.

(ii) There exists a constant C depending on M only such that, whenever $J(\eta) < M$, $\|p_{\theta,\eta} - p_{\theta_0,\eta_0}\|_2 \geqslant C(|\theta - \theta_0| + \|\eta - \eta_0\|_2) \wedge 1$.

Proof. (i) If $p_{\theta,\eta} \to p_{\theta_0,\eta_0}$ in L_2, then $\theta \to \theta_0$ and $\eta \to \eta_0$ in measure, whence $\||\eta - \eta_0| \wedge 1\|_2 \to 0$. Thus it suffices to prove the inequality for small values of $|\theta - \theta_0|$ and $\||\eta - \eta_0| \wedge 1\|_2$.

By a Taylor expansion (cf. equation (7.4)), uniformly in (w, z),

$$|F(\theta w + \eta(z)) - F(\theta_0 w + \eta_0(z)) - f(g_0)[(\theta - \theta_0)w + [\eta - \eta_0]_M]|$$

$$\leqslant (|\theta - \theta_0|^2 + |\eta - \eta_0|^2) \wedge 1.$$

Conclude that

$$P_0(F(\theta w + \eta(z)) - F(\theta_0 w + \eta_0(z)))^2$$

$$\geqslant (P_0((\theta - \theta_0)w + [\eta - \eta_0]_M)^2 - O(|\theta - \theta_0| \wedge 1)^4 - O(P_0|\eta - \eta_0|^4 \wedge 1)$$

$$\geqslant |\theta - \theta_0|^2 + P_0[\eta - \eta_0]_M^2 - o(|\theta - \theta_0| \wedge 1)^2 - o(P_0|\eta - \eta_0|^2 \wedge 1),$$

by the assumption $P_0 \operatorname{var}(W|Z) > 0$. Inequality (i) follows.

(ii) If $p_{\theta,\eta} \to p_{\theta_0,\eta_0}$ in L_2 and $J(\eta) = O(1)$, then, by Lemma 7.3 $\|\eta\|_\infty = O(1)$. Hence the conclusion in the first paragraph of the proof of (i) can be strengthened to $\theta \to \theta_0$ and $\|\eta - \eta_0\|_2 \to 0$. The proof proceeds along the same lines, substituting $\|\eta - \eta_0\|_2$ for $\||\eta - \eta_0| \wedge 1\|_2$. □

Lemma 7.5. *Under (7.1) we have for every random sequence* $\tilde{\theta} \xrightarrow{P} \theta_0$ and $\overline{\theta} \xrightarrow{P} \theta_0$,

$$\mathbb{G}_n(\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) - \dot{\ell}_0) = o_P(1 + \sqrt{n}|\tilde{\theta} - \theta_0|),$$

$$(\mathbb{P}_n - P_0)\ddot{\ell}(\tilde{\theta}, \hat{\psi}_{\overline{\theta}}) \xrightarrow{P} 0.$$

Proof. In view of (7.1) there is no loss of generality in assuming that $\tilde{\lambda}$ is bounded below by a multiple of $\lambda_n = n^{-k/(2k+1)}$ and bounded above by $\varepsilon n^{-1/4}$ for an arbitrary $\varepsilon > 0$. By combining Lemmas 7.3(i) and 7.1, $\|\hat{\eta}_{\tilde{\theta}}\|_\infty$ is bounded in probability by a multiple of $J(\hat{\eta}_{\tilde{\theta}}) + 1$, which by equation (7.3) is bounded by a multiple of $1 + |\tilde{\theta} - \theta_0|/\lambda_n$. Furthermore, by Taylor series arguments as used for the proof of (7.4),

$$P_0\left(\frac{\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) - \dot{\ell}_0}{1 + \sqrt{n}|\tilde{\theta} - \theta_0|}\right)^2 \leqslant \frac{|\tilde{\theta} - \theta_0|^2 + P_0|\hat{\eta}_{\tilde{\theta}} - \eta_0|^2 \wedge 1}{(1 + \sqrt{n}|\tilde{\theta} - \theta_0|)^2} \leqslant O_P\left(\frac{1}{n} + \tilde{\lambda}^2\right) = O_P(\varepsilon^2 n^{-1/2}).$$

Define $\mathscr{F}_n$ as the class of functions

$$\left\{\frac{\dot{\ell}(\theta, \theta, \eta) - \dot{\ell}_0}{1 + \sqrt{n}|\theta - \theta_0|} : J(\eta) \leqslant 1 + \frac{|\theta - \theta_0|}{\lambda_n}, \|\eta\|_\infty \leqslant 1 + J(\eta), |\theta - \theta_0| < \delta\right\}$$

$$\cap \{f \in L_2(P_0) : P_0 f^2 \leqslant \varepsilon^2 n^{-1/2}\}.$$

Then it follows that on a set of probability arbitrarily close to 1 we can bound $\mathbb{G}_n(\dot{\ell}(\tilde{\theta}, \hat{\psi}_{\tilde{\theta}}) - \dot{\ell}_0)/(1 + \sqrt{n}|\tilde{\theta} - \theta_0|)$ by $\|\mathbb{G}_n\|_{\mathscr{F}_n}$.

We now apply the maximal inequality Lemma 3.4.2 in van der Vaart and Wellner (1996) to $\|\mathbb{G}_n\|_{\mathscr{F}_n}$. Since $\dot{\ell}(\theta, \theta, \eta)$ depends on (θ, η) in a Lipschitz and essentially monotone manner,

$$\begin{aligned}
&\log N_{[]}(\varepsilon, \mathscr{F}_n, L_2(P_0)) \\
&\quad \leqslant \log N_{[]}\left(\varepsilon, \left\{\frac{\eta}{1+\sqrt{n}|\theta-\theta_0|}: J(\eta) \leqslant 1 + \frac{|\theta-\theta_0|}{\lambda_n}, \|\eta\|_\infty \leqslant 1 + J(\eta)\right\}, L_2(P_0)\right) \\
&\qquad + \log \frac{1}{\varepsilon} \\
&\quad \leqslant \left(\frac{1+(\sqrt{n}\lambda_n)^{-1}}{\varepsilon}\right)^{1/k},
\end{aligned}$$

by Birman and Solomjak (1967), since

$$J\left(\frac{\eta}{1+\sqrt{n}|\theta-\theta_0|}\right) = \frac{J(\eta)}{1+\sqrt{n}|\theta-\theta_0|} \leqslant 1 + \frac{1}{\sqrt{n}\lambda_n}.$$

Therefore, the relevant entropy integral is equal to

$$\int_0^{\varepsilon n^{-1/4}} \left(\frac{1+(\sqrt{n}\lambda_n)^{-1}}{\varepsilon}\right)^{1/2k} \mathrm{d}\varepsilon \leqslant (\varepsilon n^{-1/4})^{(1-1/2k)}\left(1+(\sqrt{n}\lambda_n)^{-1}\right)^{1/2k}.$$

By Lemma 3.4.2 in van der Vaart and Wellner (1996), we conclude that $E^*\|\mathbb{G}_n\|_{\mathscr{F}_n} \to 0$. This concludes the proof of the first assertion, which is the verification of (2.4′).

To prove the second assertion, we need a Glivenko–Cantelli theorem for classes of functions that change with n. A suitable extension of the uniform entropy Glivenko–Cantelli theorem is as follows. If $\mathscr{F}_n$ are suitably measurable classes of functions with uniformly integrable envelope functions and $\log N(\varepsilon, \mathscr{F}_n, L_1(\mathbb{P}_n)) = o_P^*(n)$, then $\|\mathbb{P}_n - P_0\|_{\mathscr{F}_n} \xrightarrow{P} 0$ for every $\varepsilon > 0$. The proof of Theorem 2.4.3 in van der Vaart and Wellner (1996) applies with minor notational changes. We apply this theorem to the set $\ddot{\mathscr{F}}_n$ of functions $\ddot{\ell}(t, \theta, \eta)$ with t and θ_0 ranging over a neighbourhood of θ_0 and $\lambda_n J(\eta)$ bounded by a constant. By arguments as in Lemma 7.2,

$$\sup_Q \log N(\varepsilon, \mathscr{F}_n, L_1(Q)) \leqslant \left(\frac{1+\lambda_n^{-1}}{\varepsilon}\right)^{1/k}.$$

Thus the present classes $\ddot{\mathscr{F}}_n$ certainly satisfy the entropy condition. Moreover, they are uniformly bounded. Since the functions $\dot{\ell}(\tilde{\theta}, \overline{\theta}, \hat{\eta}_{\overline{\theta}})$ are contained in $\ddot{\mathscr{F}}_n$ with probability tending to 1, the second assertion of the lemma follows. □

Acknowledgements

The research of the first author was partially supported by NSF grant DMS-9307255 and NIDA grant A P50 DA 10075-01.

References

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

Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J. (1993) *Efficient and Adaptive Estimation for Semi-parametric Models*. Baltimore, MD: Johns Hopkins University Press.

Birman, M.S. and Solomjak, M.Z. (1967) Piecewise-polynomial approximation of functions of the classes W_p. *Math. USSR-Sb.*, **73**, 295–317.

Chang, M.N. (1990) Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.*, **18**, 391–404.

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

Gaydos, B.L. and Lindsay, B.G. (1996) Use of the efficient score to calculate standard errors in a semi-parametric errors-in-variables model. Preprint.

Gill, R.D. (1989) Non- and semi-parametric maximum likelihood estimators and the von-Mises method (part I). *Scand. J. Statist.*, **16**, 97–128.

Gill, R.D., van der Laan, M.J. and Wijers, B.J. (1995) The line segment problem. Preprint.

Gu, M.G. and Zhang, C.H. (1993) Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann. Statist.*, **21**, 611–624.

Huang, J. (1996) Efficient estimation for the Cox model with interval censoring. *Ann. Statist*., **24**, 540–568.

Huang, J. and Wellner J.A. (1995) Efficient estimation for the Cox model with Case 2 interval censoring. Preprint.

Kim, J. and Pollard, P. (1990) Cube root asymptotics. *Ann. Statist.*, **18**, 191–219.

Mammen, E. and van de Geer, S. (1997) Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.*, **25**, 1014–1035.

Murphy, S.A. (1995) Asymptotic theory for the frailty model. *Ann. Statist.*, **23**, 182–198.

Murphy, S.A. and van der Vaart, A.W. (1996) Semiparametric mixtures in case–control studies. Preprint.

Murphy, S.A., Rossini, A.J. and van der Vaart, A.W. (1997) MLE in the proportional odds model *J. Amer. Statist. Assoc*., **92**, 968–976.

Murphy, S.A. and van der Vaart, A.W. (1997) Semiparametric likelihood ratio inference. *Ann. Statist.*, **25**, 1471–1509.

Nielsen, G.G., Gill, R.D., Andersen, P.K. and Sørensen, T.I. (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statisti.*, **19**, 25–44.

O'Sullivan, F., Yandell, B.S. and Raynor, W.J. (1986) Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc*., **81**, 96–103.

Parner, E. (1998) Asymptotic normality in the correlated gamma-frailty model. *Ann. Statist*., **26**, 183–214.

Qin, J. (1993) Empirical likelihood in biased sample problems. *Ann. Statist.*, **21**, 1182–1196.

Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.

Qin, J. and Wong, A. (1996) Empirical likelihood in a semi-parametric model. *Scand. J. Statist.*, **23**, 209–220.

Roeder, K., Carroll, R.J. and Lindsay, B.G. (1996) A semiparametric mixture approach to case–control studies with errors in covariables. *J. Amer. Statist. Assoc*., **91**, 722–732.

Severini, T.A. and Staniswalis, J.G. (1994) Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc*., **89**, 501–511.

Severini, T.A. and Wong, W.H. (1992) Profile likelihood and conditionally parametric models. *Ann. Statist.*, **20**, 1768–1802.

van der Laan, M.J. (1993) Efficient and inefficient estimation in semiparametric models. Doctoral dissertation, University of Utrecht, The Netherlands.

van der Vaart, A.W. (1994a) Infinite dimensional *M*-estimation. In B. Grigelionis, J. Kubilius, H. Pragarauskas and V. Statulevicius (eds), *Probability Theory and Mathematic Statistics*, Proceedings of the 6th Vilnius Conference, pp. 715–734. Zeist: VSP International Science Publishers, and Vilnius: TEV Ltd Publishers Service Group.

van der Vaart, A.W. (1994b) On a model of Hasminskii and Ibragimov. In A. Zaitsev (ed.), *Proceedings Kolmogorov Semester at the Euler International Mathematical Institute, St. Petersburg*. Amsterdam: North-Holland.

van der Vaart, A.W. (1994c) Maximum likelihood estimation with partially censored observations. *Ann. Statist.*, **22**, 1896–1916.

van der Vaart, A.W. (1996) Efficient estimation in semiparametric models. *Ann. Statist.*, **24**, 862–878.

van der Vaart, A.W. (1998) *Asymptotic Statistics.* Cambridge: Cambridge University Press.

van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

Received December 1996 and revised August 1998.