Semiparametrics via parametrics and contiguity

Adam Lee[1], Emil A. Stoltenberg[1], and Per A. Mykland[2]

[1]*Department of Data Science BI Norwegian Business School*
[2]*Department of Statistics, The University of Chicago*

January 16, 2025

Abstract

Inference on the parametric part of a semiparametric model is no trivial task. On the other hand, if one approximates the infinite dimensional part of the semiparametric model by a parametric function, one obtains a parametric model that is in some sense close to the semiparametric model; and inference may proceed by the method of maximum likelihood. Under regularity conditions, and assuming that the approximating parametric model in fact generated the data, the ensuing maximum likelihood estimator is asymptotically normal and efficient (in the approximating parametric model). Thus one obtains a sequence of asymptotically normal and efficient estimators in a sequence of growing parametric models that approximate the semiparametric model and, intuitively, the limiting 'semiparametric' estimator should be asymptotically normal and efficient as well. In this paper we make this intuition rigorous. Consequently, we are able to move much of the semiparametric analysis back into classical parametric terrain, and then translate our parametric results back to the semiparametric world by way of contiguity. Our approach departs from the sieve literature by being more specific about the approximating parametric models, by working under these when treating the parametric models, and by taking advantage of the mutual contiguity between the parametric and semiparametric models to lift conclusions about the former to conclusions about the latter. We illustrate our theory with two canonical examples of semiparametric models, namely the partially linear regression model and the Cox regression model. An upshot of our theory is a new, relatively simple, and rather parametric proof of the efficiency of the Cox partial likelihood estimator.

Keywords: Asymptotic equivalence; Cox model; efficiency; Fisher information; maximum likelihood; partially linear regression; profile likelihood; semiparametric models; sieves

1 Introduction

Suppose we observe an independent and identically distributed (i.i.d.) sample from a distribution $P_{\theta,\eta}$, with $\theta \in \Theta \subset \mathbb{R}^p$, and η ranging over a space of infinite dimension, and that our aim is to make inference on the parametric part θ. As is well known, the presence of the infinite dimensional nuisance parameter η makes this a nontrivial problem. In some models, an estimator for θ can be difficult to derive; in other models, an estimator might be simple to derive, but difficult to analyse; and in yet other models, both deriving and analysing an estimator for θ is a formidable task. To deal with these challenges, we borrow a simplifying strategy from the high-frequency econometrics literature, in particular Mykland and Zhang (2009). The basic idea is to pretend that the data stem from a parametric distribution, derive estimators and study their properties under this parametric distribution, then let the parametric distributions grow in such a way that they are mutually contiguous with respect to the full semiparametric distribution (i.e., asymptotically equivalent), and finally use the contiguity to switch the analysis back to the semiparametric model in the limit.

Working under parametric distributions has several advantages: For parametric (and dominated) models there is no ambiguity in defining a likelihood function; based on this function, estimators can be derived or defined without difficulty; and the required regularity conditions for these estimators to be consistent, asymptotically normal, and efficient are often simple to check. Given that the parametric distributions are constructed such that they are asymptotically equivalent to the semiparametric distribution, under what conditions are the parametric asymptotic statistical properties (asymptotic normality, efficiency) preserved when we switch back to the semiparametric model? This is the question to which we provide an answer in this article. Our answer is a set of conditions, and these conditions are, as we exemplify, not that difficult to check. In a certain sense, the question we pose in this article is a statistical analogue of the classical real analysis exam question relating to the condition under which we can conclude that a sequence of continuous functions has a continuous limit. Both our question and the real analysis exam question, pertains to conditions ensuring that certain properties of a sequence – asymptotic normality and efficiency in our case, continuity in the real analysis exam question – are preserved in the limit.

The ideas outlined above can be illustrated with the partially linear regression model (to which we return in greater detail in Section 5). In this model a real valued outcome Y of a subject with covariates $(W, Z) = (w, z)$ stems from the model

$$Y = \eta(z) + \theta w + \varepsilon, \tag{1.1}$$

for an unobservable noise term ε independent of (W, Z), where η is an infinite dimensional nuisance parameter (an unknown function), and θ is a Euclidean parameter on which we seek inference. If ε is assumed to be mean zero and normal with variance σ^2, and the covariates have a density, then the observation (W, Y, Z) has a density, but this density cannot be used to define a maximum likelihood estimator for (θ, η), because the maximiser for η will just interpolate the data.[1] A simple way to overcome these issues is to pretend that instead of the semiparametric model implied by (1.1), Y given $(W, Z) = (w, z)$ stems from the parametric model

$$Y = \beta_m(z)^{\mathrm{t}}\gamma + \theta w + \varepsilon, \tag{1.2}$$

where $\beta_m = (\beta_{m,1}, \dots, \beta_{m,k_m})^{\mathrm{t}}$ is a collection of orthonormal (or other basis) functions, $\gamma = (\gamma_1, \dots, \gamma_{k_m})^{\mathrm{t}}$ is a Euclidean parameter vector, and (W, Z, ε) has the same distribution as the similarly denoted random variable above. Now, maximum likelihood estimation is a least squares problem, and we readily obtain a maximum likelihood estimator for (θ, γ), say $(\widehat{\theta}_{m,n}, \widehat{\gamma}_{m,n})$, where m indicates the size of the parametric model, and n is the sample size. Assuming that the data in fact stem from the parametric model in (1.2), we get from standard parametric likelihood theory that $\sqrt{n}(\widehat{\theta}_{m,n} - \theta)$ converges in distribution to a mean zero normal with variance J_m^{-1}, where $J_m = \sigma^{-2}\big(\mathbb{E}\, W^2 - \sum_{j=1}^{k_m}(\mathbb{E}\,\{\beta_{m,j}(Z)W\})^2\big)$, and this is the efficient information under the model in (1.2). Using the orthonormality of the functions $\beta_{m,1}, \dots, \beta_{m,k_m}$, it follows from Parseval's identity that

$$J_m \to J := \sigma^{-2}\big(\mathbb{E}\, W^2 - (\mathbb{E}\,\{W \mid Z\})^2\big),$$

which is the efficient information under the model in (1.1) (cf. Bickel et al., 1998, p. 110). This raises two questions, to which we provide answers in this article: First, under what conditions does the above finding allow us to conclude that $\sqrt{n}(\widehat{\theta}_{m,n} - \theta)$ converges in distribution a mean zero normal with variance J^{-1} under the semiparametric model in (1.1) when m tends to infinity

[1] See Andersen et al. (1993, pp. 221–226) for a discussion of the difficulties in defining a likelihood function in the non- and semiparametric case.

with n; second, given that we have such convergence, under what conditions can we conclude that J is the efficient information?

The idea of approximating semiparametric models with growing parametric models is, of course, not new, as there is a well developed literature on sieve estimation. Compared to this broad literature however, the goal of the present paper is much more specific. We aim to find conditions under which we can transfer the classical efficiency results of maximum likelihood estimators of a Euclidean parameter in well-behaved parametric models to the semiparametric case. Whilst similar conclusions are reached in the sieve literature (see, e.g., Theorem 4 in Shen (1997)), these results typically follow as special cases of more general results permitting, for example, criterion functions which are not log-likelihoods. Our goal in this paper is to show that by focussing on the specific case of maximum likelihood estimation and being more specific about the behaviour of the approximating models permits a simpler and more user-friendly theory of semiparametric estimation. We show asymptotic normality and semiparametric efficiency of maximum likelihood estimators under conditions which are, we believe, simpler to both understand and verify than the often rather technical assumptions made in the sieve literature. In particular, we are able to show that this efficiency result holds in a broad class of models without imposing any empirical process type conditions. As we exemplify with the Cox model in Section 5.2, sieves might also just be used as a theoretical tool (and not for estimation), and allows for simpler efficiency proofs than those currently available.

2 General setup and parametric theory

Let $X_1, \ldots, X_n$ be i.i.d. replicates of X, where X stems from a distribution $P_0 := P_{\theta_0,\eta_0}$ belonging to a semiparametric family of distributions $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in \mathcal{H}\}$, where $\Theta \subset \mathbb{R}^p$ for some $p \geq 1$, and $\mathcal{H}$ is a space of infinite dimension. We suppose that $\mathcal{P}$ is dominated by some σ-finite measure μ, and write $p_{\theta,\eta}$ for the densities. The problem is to do inference on θ in the presence of the infinite dimensional nuisance parameter η.

The simplifying strategy discussed in the introduction involves the construction of certain parametric submodels: For each $m \geq 1$, let $\mathcal{H}_m$ be a family of parametric functions η_γ indexed by a parameter $\gamma \in \Gamma_m \subset \mathbb{R}^{k_m}$. Typically these shall be such that $\mathcal{H}_m \subset \mathcal{H}_{m+1}$ for all m with $\cup_{m\geq 1}\mathcal{H}_m$ dense in $\mathcal{H}$ in an appropriate topology. Let $T_m \colon \mathcal{H}_m \to \Gamma_m$ be an isomorphism between the function space $\mathcal{H}_m$ and Γ^m, thus $T_m\eta_\gamma = \gamma \in \Gamma_m$ and $T_m^{-1}\gamma = \eta_\gamma \in \mathcal{H}_m$ (see, e.g., Royden and Fitzpatrick, 2010, Corollary 13.5, p. 260 for the existence of such an isomorphism). Then, for each $m \geq 1$,

$$\mathcal{P}_m = \{P_{\theta,T_m^{-1}\gamma} \colon \theta \in \Theta, \gamma \in \Gamma_m\},$$

is a $p+k_m$ dimensional parametric model, and $P_{\theta,T_m^{-1}\gamma}$ has density $p_{\theta,T_m^{-1}\gamma}$ with respect to μ. As discussed above, the idea is to carry out the analysis *as if* $X_1, \ldots, X_n$ was an i.i.d. sample from a member of $\mathcal{P}_m$, let m increase with the sample size n, and use contiguity to switch back to P_{θ_0,η_0} in the limit. The natural choice is to suppose that the sample stems from $P_m := P_{\theta_0,T_m^{-1}\gamma_0}$, where θ_0 equals the true value in the big model P_0, and $(\gamma_0)_{m\geq 1} = (\gamma_{0,m})_{m\geq 1}$ is a sequence of growing vectors so that η_{γ_0} approaches the true η_0 as m tends to infinity. To declutter the notation, we avoid indexing γ and γ_0 by m, as the size of these parameter vectors will be clear from the context. The densities of P_0 and P_m with respect to μ are denoted p_0 and p_m, respectively, with similar subscripts for the expectations, $\mathbb{E}_0\, g(X) = \int g(x)\,\mathrm{d}P_0(x)$ and $\mathbb{E}_m\, g(X) = \int g(x)\,\mathrm{d}P_m(x)$. The product measures arising from an i.i.d. sample of size n is indicated by a superscript n, e.g., $P_{\theta,\eta}^n = P_{\theta,\eta} \times \cdots \times P_{\theta,\eta}$. The score functions with respect to θ and γ under the model $\mathcal{P}_m$ are

$$\dot{\ell}_{\theta,T_m^{-1}\gamma} = \frac{\partial}{\partial\theta}\log p_{\theta,T_m^{-1}\gamma}, \quad \text{and} \quad \dot{v}_{\theta,T_m^{-1}\gamma} = \frac{\partial}{\partial\gamma}\log p_{\theta,T_m^{-1}\gamma},$$

viewed as functions of $(\theta,\gamma)\in\Theta\times\Gamma_m$. When evaluated in (θ_0,γ_0) we write $\dot\ell_m=\dot\ell_{\theta_0,T_m^{-1}\gamma_0}$ and $\dot v_m=\dot v_{\theta_0,T_m^{-1}\gamma_0}$ for these functions. The log-likelihood function under $\mathcal{P}_m$ is $(\theta,\gamma)\mapsto\sum_{i=1}^n\log p_{\theta,T_m^{-1}\gamma}(X_i)$ for $(\theta,\gamma)\in\Theta\times\Gamma_m$, and the corresponding maximum likelihood estimator for θ is $\widehat\theta_{m,n}$. Let i_m be the Fisher information matrix under P_m and block partition it as follows

$$i_m=\begin{pmatrix} i_{m,00} & i_{m,01}\\ i_{m,10} & i_{m,11}\end{pmatrix}=\mathbb{E}_m\begin{pmatrix}\dot\ell_m\dot\ell_m^{\mathrm{t}} & \dot\ell_m\dot v_m^{\mathrm{t}}\\ \dot v_m\dot\ell_m^{\mathrm{t}} & \dot v_m\dot v_m^{\mathrm{t}}\end{pmatrix}.$$

The standard maximum likelihood theory for inference on θ in the presence of a finite dimensional nuisance parameter $\gamma\in\Gamma_m$ goes as follows: Under regularity conditions (see, e.g., van der Vaart (1998, Theorem 5.39, p. 65)), the maximum likelihood estimators are asymptotically linear in the parametric efficient influence function, that is, for fixed m,

$$\sqrt{n}(\widehat\theta_{n,m}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n J_m^{-1}\tilde\ell_m(X_i)+o_{P_m}(1),\tag{2.1}$$

as n tends to infinity, where $\tilde\ell_m$ is the efficient score function

$$\tilde\ell_m=\dot\ell_m-i_{m,01}i_{m,11}^{-1}\dot v_m=\dot\ell_m-\Pi_m\dot\ell_m,$$

with Π_m is the orthogonal projection onto the linear span of $\{\dot v_{m,j}:j=1,\ldots,k_m\}$ in $L^2(P_m)$, and

$$J_m=\mathbb{E}_m\,\tilde\ell_m(X)\tilde\ell_m(X)^{\mathrm{t}}=i_{m,00}-i_{m,01}i_{m,11}^{-1}i_{m,10}.$$

Provided the model $\mathcal{P}_m$ is differentiable in quadratic mean at (θ_0,γ_0) and the efficient information matrix J_m is nonsingular, then (2.1) is equivalent to $\widehat\theta_{m,n}$ being the best regular estimator (e.g., van der Vaart, 1998, Lemma 8.14). This means that as n tends to infinity, but m remains fixed, $\sqrt{n}(\widehat\theta_{n,m}-\theta_0)$ converges in distribution under P_m^n to a mean zero normal distribution with variance matrix J_m^{-1}, this being the smallest possible asymptotic variance matrix of any regular estimator.

So far in this section, the index m determining the size of the parametric model, has been kept fixed. When we let m grow, however, it is often the case that $J_m\to J$ for some nonsingular matrix J. It is then tempting to conclude that (i) $\sqrt{n}(\widehat\theta_{n,m_n}-\theta_0)\rightsquigarrow \mathrm{N}(0,J^{-1})$ under $P_{m_n}^n$, for some subsequence $(m_n)_{n\geq1}$, and hence also under P_0^n by a contiguity argument, i.e., joint convergence with the likelihood ratio; and (ii) J is the semiparametrically efficient information under P_0. Neither of these conclusions are true in general. Part of the impetus for this paper is to provide conditions under which (i) and (ii) *are* true, and thereby providing a recipe for semiparametric inference by way of parametrics and contiguity.

Remark 1. If $P_{m,n}^n\rightsquigarrow P_m$ for each m, and $P_m\rightsquigarrow P_0$, then there is a subsequence (n_m) such that $P_{m,n_m}^{n_m}\rightsquigarrow P_0$, see, e.g., Dudley (2002, p. 395). For our purposes, the subsequence in this result ends up in the wrong place, so to speak. A theorem getting the subsequence where we want it, is Theorem 4.2 in Billingsley (1968, p. 25), which says that if $Z_{m,n}\rightsquigarrow Z_m$ and $Z_m\rightsquigarrow Z$, and there is a subsequence (m_n) such that for any $\varepsilon>0$, $\limsup_n\Pr(d(Z_{m,n},Z_{m_n,n})\geq\varepsilon)\to0$ as $m\to\infty$ (for an appropriate metric d), then $Z_{m_n,n}\rightsquigarrow Z$. This theorem is not amenable to the product probability space setting of our paper.

3 Asymptotic efficiency

For the efficiency results of our parametric approximations to a semiparametric model to make sense, we need to impose some structure on the semiparametric model. In Section 3.1 we outline this structure and introduce some notation. In Section 3.2 the conditions we impose on the parametric approximations are presented and discussed.

3.1 The true semiparametric model

We concentrate on smooth models for i.i.d. data, as in the classical parametric theory. This is made precise by imposing a differentiability in quadratic mean (DQM) condition on the semiparametric model $\mathcal{P}$. Let B be a linear space. We will consider measures $P_{\theta_0+\tau/\sqrt{n},\eta_n(b)}$ for $h = (\tau, b) \in \mathbb{R}^p \times B$ where $\eta_n(b) \to \eta_0$ as n tends to infinity, and $\eta_n(0) = \eta_0$.

Assumption 1 (DQM)**.** *For each $h \in \mathbb{R}^p \times B$, $P_{\theta+\tau/\sqrt{n},\eta_n(b)} \in \mathcal{P}$ for all large enough n, and*

$$\lim_{n\to\infty} \int \{\sqrt{n}(p^{1/2}_{\theta_0+\tau/\sqrt{n},\eta_n(b)} - p_0^{1/2}) - \tfrac{1}{2} Ah\, p_0^{1/2}\}^2 \,\mathrm{d}\mu = 0,$$

where A is a bounded linear map between $\mathbb{R}^p \times B$ and $L^2(P_0)$.

It is automatic from Assumption 1 that $Ah \in L^2(P_0)$ and $\int Ah \,\mathrm{d}P_0 = 0$ (see van der Vaart (1998, Lemma 25.14, p. 363)). Since A is assumed to be linear we can split out the contributions of the parametric parameter of interest from the infinite dimensional nuisance as $Ah = \tau^{\mathrm{t}}\dot{\ell} + Db$, where $\dot{\ell}$ is the ordinary score function for θ in a model where the nuisance η is fixed; while $D\colon B \to \mathbb{R}$ is a linear operator, and Db has the interpretation of a score function for η with θ fixed. The efficiency bound for estimation of θ_0 is determined by the efficient score,

$$\tilde{\ell} = \dot{\ell} - \Pi\dot{\ell},$$

where Π denotes the orthogonal projection onto the closure of $\{Db : b \in B\}$ in $L^2(P_0)$. Define

$$J = \mathbb{E}_0\, \tilde{\ell}(X)\tilde{\ell}(X)^{\mathrm{t}}.$$

We will suppose that this matrix is nonsingular, which is necessary for the existence of regular estimators of θ (Chamberlain, 1986, Theorem 2, p. 194).

Assumption 2 (Nonsingularity)**.** *J is nonsingular.*

By the convolution theorem, any regular estimator of θ_0 has asymptotic covariance matrix bounded below by J^{-1}, the inverse of the efficient information (see, e.g., van der Vaart (1998, Theorem 25.20 p. 366 and Lemma 25.25, p. 369)).

3.2 The parametric approximations

We now introduce our assumption on the parametric approximations $P_m = P_{\theta_0, T_m^{-1}\gamma_0}$ to $P_0 = P_{\theta_0,\eta_0}$. In particular, we assume that the true semiparametric model P_0 can be approximated, in an appropriate sense, by a sequence of contiguous alternatives. Since P_0 is unknown, checking Assumption 3 means in practice checking it for any member of $\mathcal{P}$. For clarity, however, we state this and the upcoming assumptions for the specific member P_0 of $\mathcal{P}$ that generated the data.

Assumption 3 (Contiguity)**.** *There is a subsequence (m_n) and a function g such that $\mathbb{E}_0\, g(X) = 0$, $\mathbb{E}_0\, g(X)^2$ is finite, and $\mathbb{E}_0\, g(X)\tilde{\ell}(X) = 0$, and the log-likelihood ratio satisfies*

$$\log \frac{\mathrm{d}P^n_{m_n}}{\mathrm{d}P^n_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) - \tfrac{1}{2}\mathbb{E}_0\, g(X_1)^2 + o_{P_0^n}(1). \tag{3.1}$$

In our applications, we will seek that $g = 0$, but the orthogonality $\mathbb{E}_0\, g(X)\tilde{\ell}(X) = 0$ is all that is actually needed for our results. The log-likelihood ratio expansion in (3.1) is equivalent to the densities $p_{m_n} = \mathrm{d}P_{m_n}/\mathrm{d}\mu$ satisfying the DQM type condition

$$\lim_{n\to\infty} \int \{\sqrt{n}(p^{1/2}_{m_n} - p_0^{1/2}) - \tfrac{1}{2} g\, p_0^{1/2}\}^2 \,\mathrm{d}\mu = 0, \tag{3.2}$$

for g such that $\int g\tilde{\ell}\,\mathrm{d}P_0 = 0$, that is, for the function g appearing in (3.1). See for example Strasser (1985, Corollary 75.9) or Le Cam (1986, Prop. 17.3.2) for this equivalence. Thus, to check Assumption 3 it suffices to show either (3.1) or (3.2).

The log-likelihood ratio expansion in (3.1) is key to our results. In particular, since the data are assumed i.i.d., Assumption 3 and the central limit theorem yield

$$\log \frac{\mathrm{d}P^n_{m_n}}{\mathrm{d}P^n_0} \overset{P^n_0}{\rightsquigarrow} Z, \quad \text{where} \quad Z \sim \mathrm{N}\big(-\tfrac{1}{2}\mathbb{E}_0\, g(X)^2, \mathbb{E}_0\, g(X)^2\big).$$

Since $\exp(Z)$ is almost surely positive and $\mathbb{E}\exp(Z) = 1$, we get from Le Cam's first lemma that P^n_0 and $P^n_{m_n}$ are mutually contiguous (see van der Vaart (1988, Example 6.5, p. 89)). Thus, provided $\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0)$ converges jointly with $\log(\mathrm{d}P^n_0/\mathrm{d}P^n_{m_n})$ to a Gaussian limit under $P^n_{m_n}$, the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0)$ under P^n_0 can be recovered by Le Cam's third lemma. As such, Assumption 3 restricts the possible change in the limiting distribution of $\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0)$ resulting from a change of measure from the parametric P_{m_n} and back to the semiparametric P_0. In particular, provided the limiting distribution of $\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0)$ under $P^n_{m_n}$ is orthogonal to g in the sense that $\mathrm{Cov}(\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0), n^{-1/2}\sum_{i=1}^n g(X_i))$ tends to zero, changing the measure from P_{m_n} to P_0 does not affect the limiting distribution.

In addition to the contiguity condition in Assumption 3, we require that the efficient scores in the parametric submodels approximate the efficient score of the full semiparametric model in a statistically relevant sense. As we will take our limits along the subsequence $(m_n)_{n\in\mathbb{N}}$ of Assumption 3, it is sufficient that this approximation holds along this subsequence. This is convenient as Assumption 3 implies that $P_{m_n} \to P_0$ in total variation (via (3.2)), which can help to simplify the demonstration of (3.3) (see Lemma 3.2 below).

Assumption 4 (Efficient score approximation). *The efficient scores $\tilde{\ell}_{m_n}$ exist and*

$$\lim_{n\to\infty} \int \|\tilde{\ell}_{m_n} p^{1/2}_{m_n} - \tilde{\ell} p_0^{1/2}\|^2 \,\mathrm{d}\mu = 0, \tag{3.3}$$

where $\|x\| = (\sum_{j=1}^p x_j^2)^{1/2}$ is the Euclidean distance.

This assumption implies convergence of the efficient parametric information matrices to their semiparametric counterpart. To see this, notice that for any $a \in \mathbb{R}^p$ and assuming that all J_{m_n} and J are positive definite, the reverse triangle inequality and then Cauchy–Schwarz yield

$$\begin{aligned} |(a^{\mathrm{t}} J_{m_n} a)^{1/2} - (a^{\mathrm{t}} J a)^{1/2}|^2 &= |\|a^{\mathrm{t}}\tilde{\ell}_{m_n} p^{1/2}_{m_n}\|_\mu - \|a^{\mathrm{t}}\tilde{\ell} p_0^{1/2}\|_\mu|^2 \\ &\le \|a^{\mathrm{t}}(\tilde{\ell}_{m_n} p^{1/2}_{m_n} - \tilde{\ell} p_0^{1/2})\|^2_\mu \le \|a\|^2 \int \|\tilde{\ell}_{m_n} p^{1/2}_{m_n} - \tilde{\ell} p_0^{1/2}\|^2 \,\mathrm{d}\mu. \end{aligned}$$

By continuity of the square function, this entails that $a^{\mathrm{t}} J_{m_n} a \to a^{\mathrm{t}} J a$, and since the above is true for any $a \in \mathbb{R}^p$, $J_{m_n} \to J$ in the spectral norm.

A simplifying feature of our theory compared to the standard semiparametric theory, is that one may verify Assumption 4 without explicitly performing any projections, as seen by the following lemma. In this lemma the space B and the linear operator Db are as defined in connection with Assumption 1.

Lemma 3.1. *Suppose that the scores $\dot{\ell}_{m_n}$ and $\dot{v}_{m_n}$ exist in the DQM sense. If*

$$\lim_{n\to\infty} \int \|\dot{\ell}_{m_n}\, p^{1/2}_{m_n} - \dot{\ell} p_0^{1/2}\|^2 \,\mathrm{d}\mu = 0, \tag{3.4}$$

and for any $b \in B$ there are vectors $a_{m_n} \in \mathbb{R}^{k_{m_n}}$, such that

$$\lim_{n\to\infty} \int (a^{\mathrm{t}}_{m_n}\, \dot{v}_{m_n}\, p^{1/2}_{m_n} - Db p_0^{1/2})^2 \,\mathrm{d}\mu = 0, \tag{3.5}$$

then Assumption 4 holds.

Proof. Under the assumption of the lemma, $\tilde{\ell}_{m_n}$ exists as soon as $\dot{\ell}_{m_n}$ and $\dot{v}_{m_n}$ do. Given (3.4) and (3.5), an application of Theorem A.1 in the appendix yields (3.3). $\square$

We now present two lemmas which can help with the verification of (3.3), or of (3.4) and (3.5). Their straightforward proofs are deferred to Appendix A.

Lemma 3.2. *Suppose Assumption 3 holds. If $f_{m_n} = f_0 + o_{P_0}(1)$, $\int f_0^2 \, \mathrm{d}P_0 < \infty$ and $f_{m_n}^2$ is uniformly P_{m_n}-integrable then $\lim_{n\to\infty} \int \|f_{m_n} p_{m_n}^{1/2} - f_0 p_0^{1/2}\|^2 \, \mathrm{d}\mu = 0$.*

That $f_{m_n}^2$ is uniformly P_{m_n}-integrable means that $\lim_{K\to\infty} \sup_n \int_{|f_{m_n}|\geq K} f_{m_n}^2 \, \mathrm{d}P_{m_n} = 0$. An alternative approach which circumvents the requirement to establish uniform square integrability directly can be based on a lemma due originally to Riesz (1928) (cf. van der Vaart, 1998, Proposition 2.29, p. 22). This lemma also connects Assumption 4 with the intuitive argument based on observing that in certain models the sequence of parametric efficient information matrices J_{m_n} has a limit, as discussed in Section 2. Since it sheds considerable light on Assumption 4 and its connection to the discussion in Section 2, we include the following lemma.

Lemma 3.3. *Suppose that* (i) *$p_{m_n} \to p_0$ in μ-measure or* (ii) *for any measurable set A, $P(A) \leq \liminf_{n\to\infty} P_{m_n}(A)$. If $f_{m_n} \to f_0$ in μ-measure and $\limsup_{n\to\infty} \int f_{m_n}^2 \, \mathrm{d}P_{m_n} \leq \int f_0^2 \, \mathrm{d}P_0 < \infty$, then $\lim_{n\to\infty} \int \|f_{m_n} p_{m_n}^{1/2} - f_0 p_0^{1/2}\|^2 \, \mathrm{d}\mu = 0$.*

If we use these either of these lemmas to verify (3.3) directly, we are required to first find the efficient score under the semiparametric model $\mathcal{P}$, a task which involves performing sometimes complicated projections. Lemma 3.1, on the other hand, does not involve any projections. The following corollary permits us to find the efficient semiparametric score $\tilde{\ell}$ as a limit. It is used repeatedly in the applications of Section 5. Recall that the efficient score at $(\theta_0, T_m^{-1}\gamma_0)$ is $J_m = \mathbb{E}_m \, \tilde{\ell}_m \tilde{\ell}_m^{\mathrm{t}}$.

Corollary 3.4. *Suppose that (3.4) and (3.5) hold. If there is a vector f_0 of functions such that the conditions of either Lemma 3.2 or 3.3 hold for f_{m_n} with components P_{m_n}-a.s. equal to each of the components of $\tilde{\ell}_{m_n}$, then $f_0 = \tilde{\ell}$ P_0-a.s..*

Proof. Since Lemma 3.1 holds, we have that $\int \|\tilde{\ell}_m p_m^{1/2} - \tilde{\ell} p_0^{1/2}\|^2 \, \mathrm{d}\mu \to 0$, which is Assumption 4. By either Lemma 3.2 or 3.3 we have $\int \|\tilde{\ell}_m p_m^{1/2} - f_0 p_0^{1/2}\|^2 \, \mathrm{d}\mu \to 0$. Since L^2 limits are unique up to sets of measure zero, $f_0 p_0^{1/2} = \tilde{\ell} p_0^{1/2}$ μ-almost surely, hence $f_0 = \tilde{\ell}$ P_0-almost surely. $\square$

In order to approximate the semiparametric model, we let m increase with n. This means that, for our purposes, the property corresponding to the asymptotic linearity in the parametric efficient score function exhibited in (2.1), is

$$\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} J_{m_n}^{-1} \tilde{\ell}_{m_n}(X_i) + o_{P_{m_n}^n}(1). \tag{3.6}$$

Note that property (3.6) is *not* implied by (2.1), as (2.1) only requires that the remainder $\sqrt{n}(\widehat{\theta}_{m,n} - \theta_0) - n^{-1/2} \sum_{i=1}^{n} J_m^{-1} \tilde{\ell}_m(X_i)$ is $o_{P_m^n}(1)$ as $n \to \infty$, for fixed m. Verification of (3.6) of course depends on the definition of the estimator $\widehat{\theta}_{m_n,n}$. In Section 4 we use a profile likelihood technique to provide sufficient conditions for (3.6) to hold for a sequence of maximum likelihood estimators $\widehat{\theta}_{m_n,n}$ in growing parametric submodels.

Combined with Assumptions 1–4, the linearity in the influence function displayed in (3.6) implies the asymptotic efficiency of the estimator $\widehat{\theta}_{m_n,n}$. The essential idea for proving this is outlined in the following argument: If Assumption 4 holds, then $J_{m_n}^{-1} \tilde{\ell}_{m_n}$ in (3.6) may be swapped with $J^{-1}\tilde{\ell}$, thus (3.6) becomes $\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0) = n^{-1/2} \sum_{i=1}^{n} J^{-1} \tilde{\ell}(X_i) + o_{P_{m_n}^n}(1)$. Combined with Assumption 3, this asymptotic linearity ensures that $\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0)$ converges jointly with $\log(\mathrm{d}P_0^n/\mathrm{d}P_{m_n}^n)$ under $P_{m_n}^n$, and hence, by Le Cam's third lemma, one can change the measure from P_{m_n} back to P_0 at the cost of adding a bias term of $J^{-1}\mathbb{E}_0\,\tilde{\ell}(X)g(X)$ to the limiting distribution of $\sqrt{n}(\widehat{\theta}_{m_n,n}-\theta_0)$ under P_0. But by the condition on g in Assumption 3, this bias term is zero.

Theorem 3.5. *If Assumptions 1 & 2 hold, and (m_n) is a subsequence such that Assumptions 3 & 4 hold, and (3.6) is satisfied, then $\widehat{\theta}_{m_n,n}$ is best regular in $\mathcal{P}$.*

Proof. Assumptions 3 and 4 along with the i.i.d. assumption on the data, verify the conditions of Proposition A.10 in van der Vaart (1988, p. 185). Applied to our setting, this proposition gives that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{\ell}_{m_n}(X_i)-\tilde{\ell}(X_i))=\sqrt{n}(\mathbb{E}_{m_n}\tilde{\ell}_{m_n}(X)-\mathbb{E}_0\,\tilde{\ell}(X)\,)-\mathbb{E}_0\,\tilde{\ell}(X)g(X)+o_{P_0^n}(1)=o_{P_0^n}(1).$$

The first equality follows from the cited proposition. The second equality ensues because $\mathbb{E}_{m_n}\tilde{\ell}_{m_n}(X)=0$, $\mathbb{E}_0\,\tilde{\ell}(X)=0$, and $\mathbb{E}_0\,\tilde{\ell}(X)g(X)=0$ by Assumption 3. Since Assumption 3 implies that $P_{m_n}^n$ and P_0^n are mutually contiguous, we can swap the $o_{P_{m_n}^n}(1)$ in (3.6) with $o_{P_0^n}(1)$, so that (3.6) reads $\sqrt{n}(\widehat{\theta}_{m_n,n}-\theta_0)=n^{-1/2}\sum_{i=1}^n J_{m_n}^{-1}\tilde{\ell}_{m_n}(X_i)+o_{P_0^n}(1)$. Moreover, we have that $J_{m_n}\to J$ by Assumption 4, and therefore $J_{m_n}^{-1}\to J^{-1}$ since the inverse is a continuous operation. Combining this with $n^{-1/2}\sum_{i=1}^n(\tilde{\ell}_{m_n}(X_i)-\tilde{\ell}(X_i))=o_{P_0^n}(1)$ and (3.6), we conclude that

$$\sqrt{n}(\hat{\theta}_{m_n,n}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}J^{-1}\tilde{\ell}(X_i)+o_{P_0^n}(1). \tag{3.7}$$

Given Assumptions 1 and 2, the result then follows from Lemma 25.23 and Lemma 25.25 in van der Vaart (1998, pp. 367–369). □

4 Sieved profile likelihood

In addition to the four regularity conditions 1–4, Theorem 3.5 requires that the linear expansion (3.6) holds. In this section we provide two sets of conditions under which this linear expansion is satisfied for maximum likelihood estimators in growing parametric models. Both sets of conditions are sieved, i.e. growing parametric, versions of a profile likelihood theorem due to Murphy and van der Vaart (2000). These authors provide conditions under which the semiparametric profile likelihood admits a quadratic expansion which, in turn, implies a condition like (3.7). In this section we argue similarly, but replace the *semiparametric* profile likelihood with a *sieved* profile likelihood which permits us to conclude that (3.6) holds. Whilst the expansion in (3.6) requires further work, namely Theorem 3.5, in order to conclude efficiency of the sieved maximum likelihood estimator, there is a key technical advantage of working with the sieved profile likelihood in place of the semiparametric profile likelihood. The latter requires the careful construction of 'approximately least favourable submodels'. As can be seen from the examples in that article, such submodels can be quite complicated to construct. In contrast, as the likelihoods we work are parametric, exact least favourable submodels can be constructed straightforwardly, as they always take the same form.

To introduce the profile likelihood, let $(\theta,\gamma)\mapsto L_m(\theta,\gamma)(x)=p_{\theta,T_m^{-1}\gamma}(x)$ be the likelihood function under $\mathcal{P}_m$, and $L_{m,n}(\theta,\gamma)=\prod_{i=1}^n L_m(\theta,\gamma)(X_i)$ be the likelihood based on an i.i.d. sample $X_1,\dots,X_n$. Denote by $\mathrm{pl}_{m,n}(\theta)$ the profile likelihood based on the m-th parametric model,

$$\mathrm{pl}_{m,n}(\theta)=\sup_{\gamma\in\Gamma_m}L_{m,n}(\theta,\gamma),$$

For each θ, let $\widehat{\gamma}(\theta)$ be the value that achieves this supremum, that is $\mathrm{pl}_{m,n}(\theta)=L_{m,n}(\theta,\widehat{\gamma}(\theta))$.

4.1 Quadratic expansion & log concavity

A relatively straightforward set of sufficient conditions for (3.6) (or (3.7)) can be obtained using the results of Hjort and Pollard (1993). Specifically, if the function $h \mapsto A_{m,n}(h)$ is concave, where we define $A_{m,n}(h) = \log \mathrm{pl}_{m,n}(\theta_0 + h/\sqrt{n}) - \log \mathrm{pl}_{m,n}(\theta_0)$, and one manages to find a subsequence (m_n) such that

$$A_{m_n,n}(h) = \frac{h^{\mathrm{t}}}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}_{m_n}(X_i) - \tfrac{1}{2} h^{\mathrm{t}} J_{m_n} h + o_{P^n_{m_n}}(1), \tag{4.1}$$

for each h, then the 'Basic Corollary' in Hjort and Pollard (1993) immediately delivers (3.6). This setting covers a large class of semiparametric models of practical interest, as our examples demonstrate. For cases, such as our examples, in which (4.1) can be shown to hold, this provides a complete proof of asymptotic normality and efficiency of the sieved maximum likelihood estimator without requiring any empirical process type arguments. We summarise this in a proposition.

Proposition 4.1. *Suppose that Assumptions 1 & 2 hold and that (m_n) is a subsequence such that Assumptions 3 & 4 and the quadratic expansion in (4.1) hold. If $h \mapsto A_{m,n}(h)$ is concave, then (3.6) holds, and $\widehat{\theta}_{m_n,n}$ is best regular in $\mathcal{P}$.*

Proof. This follows from the Basic Corollary in Hjort and Pollard (1993). In particular, since $J_{m_n} \to J$ under Assumption 4 we have $A_{m_n,n}(h) = h^{\mathrm{t}} n^{-1/2} \sum_{i=1}^{n} \tilde{\ell}_{m_n}(X_i) - \frac{1}{2} h^{\mathrm{t}} J h + o_{P^n_{m_n}}(1)$. Since $h \mapsto A_{m,n}(h)$ is concave, $\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta_0) = n^{-1/2} \sum_{i=1}^{n} J^{-1} \tilde{\ell}_{m_n}(X_i) + o_{P^n_{m_n}}(1)$ by the Basic Corollary, and $\widehat{\theta}_{m_n,n}$ being best regular in follows from Theorem 3.5. □

Compared to Theorem 3.5, the above proposition allows us to replace (3.6) with (4.1), provided $h \mapsto A_{m,n}(h)$ is concave. One key advantage of this is that (4.1) and concavity give us *both* consistency and asymptotic normality of $\widehat{\theta}_{m_n,n}$, while to establish (3.6) without log concavity, one typically first needs to establish that this estimator is consistent.

We now present a theorem giving conditions under which (4.1) holds. This theorem is a stripped down and sieved version of Theorem 1 in Murphy and van der Vaart (2000), and requires that we introduce some quantities inspired by that paper. In the construction of these quantities we gain a lot in simplicity by working with growing parametric models, compared to attacking the semiparametric model directly. This is because under $P_m = P_{\theta_0, T_m^{-1}\gamma_0}$, the least favourable submodel for estimating θ_0 always takes the form $\theta \mapsto P_{\theta, T_m^{-1}\gamma(\theta)}$, where $\gamma(\theta) = \gamma_0 + i_{m,11}^{-1} i_{m,10}(\theta_0 - \theta)$. In view of this, define

$$\gamma_t^{\mathrm{sub}}(\theta, \gamma) = \gamma + i_{m,11}^{-1} i_{m,10}(\theta - t),$$

where we suppress the dependence of $\gamma_t^{\mathrm{sub}}(\theta, \gamma)$ upon m from the notation. For each m and each $(\theta, \gamma) \in \Theta \times \Gamma_m$ define the mappings

$$t \mapsto l_m(t, \theta, \gamma) \coloneqq \log L_m(t, \eta_{\gamma_t^{\mathrm{sub}}(\theta,\gamma)}).$$

These functions bridge the gap between the log profile likelihood and the efficient score. To see this, notice that $\eta_{\gamma_\theta^{\mathrm{sub}}}(\theta, \gamma) = \gamma$, and that the derivative of $l_m(t, \theta, \gamma)$ with respect to t is

$$\dot{l}_m(t, \theta, \gamma) = \dot{\ell}_{t, T_m^{-1}\gamma_t^{\mathrm{sub}}(\theta,\gamma)} - i_{m,01} i_{m,11}^{-1} \dot{v}_{t, T_m^{-1}\gamma_t^{\mathrm{sub}}(\theta,\gamma)} = \tilde{\ell}_{t, T_m^{-1}\gamma_t^{\mathrm{sub}}(\theta,\gamma)},$$

in particular, $\dot{l}_m(\theta_0, \theta_0, \gamma_0) = \tilde{\ell}_m$. At the same time, mimicking the sandwiching technique of Murphy and van der Vaart (2000), we have, writing $\mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f(X_i)$ for the integral with respect to the empirical measure, that for arbitrary $\tilde{\theta}_n$

$$\mathbb{P}_n l_{m_n}(\tilde{\theta}_n, \theta_0, \widehat{\gamma}(\theta_0)) - \mathbb{P}_n l_{m_n}(\theta_0, \theta_0, \widehat{\gamma}(\theta_0)) \le \frac{1}{n}(\log \mathrm{pl}_{m_n,n}(\tilde{\theta}_n) - \log \mathrm{pl}_{m_n,n}(\theta_0)), \tag{4.2}$$

and

$$\frac{1}{n}(\log \mathrm{pl}_{m_n,n}(\tilde\theta_n) - \log \mathrm{pl}_{m_n,n}(\theta_0)) \le \mathbb{P}_n l_{m_n}(\tilde\theta_n, \tilde\theta_n, \widehat\gamma(\tilde\theta_n)) - \mathbb{P}_n l_{m_n}(\theta_0, \tilde\theta_n, \widehat\gamma(\tilde\theta_n)). \tag{4.3}$$

In particular, this means that the process $h \mapsto A_{m,n}(h) = \log \mathrm{pl}_{m,n}(\theta_0 + h/\sqrt{n}) - \log \mathrm{pl}_{m,n}(\theta_0)$ can be squeezed between two quantities approximating the 'efficient LAN expansion' of (4.1).

Theorem 4.2. *Set $\tilde\theta_n = \theta_0 + h/\sqrt{n}$ for some fixed h. Assume that $t \mapsto l_m(t, \theta, \gamma)(x)$ is twice continuously differentiable for each m, (θ, γ) and x, with derivatives $\dot l_m$ and $\ddot l_m$ and that there is a subsequence (m_n) such that for $\tilde\psi$ equal to either $(\tilde\theta_n, \widehat\gamma(\tilde\theta_n))$ or $(\theta_0, \widehat\gamma(\theta_0))$,*

$$\sqrt{n}\mathbb{P}_n \dot l(\theta_0, \tilde\psi) = \sqrt{n}\mathbb{P}_n \tilde\ell_{m_n} + o_{P^n_{m_n}}(1), \quad \text{and} \quad \mathbb{P}_n \ddot l(s_n, \tilde\psi) = -J_{m_n} + o_{P^n_{m_n}}(1).$$

for any deterministic sequence $s_n \to \theta_0$. Then (4.1) holds.

Proof. A Taylor expansion keeping $\tilde\psi$ fixed, yields

$$n\mathbb{P}_n l_{m_n}(\theta_0 + h/\sqrt{n}, \tilde\psi) - n\mathbb{P}_n l_{m_n}(\theta_0, \tilde\psi) = h^{\mathrm{t}}\sqrt{n}\,\mathbb{P}_n \dot l_{m_n}(\theta_0, \tilde\psi) + \tfrac{1}{2}h^{\mathrm{t}}\,\mathbb{P}_n \ddot l_{m_n}(s_n, \tilde\psi)h,$$

for an s_n between $\tilde\theta_n$ and θ_0. Replacing $\mathbb{P}_n \ddot l_{m_n}(s_n, \tilde\psi)$ by $-J_{m_n} + o_{P^n_{m_n}}(1)$ and $\sqrt{n}\,\mathbb{P}_n \dot l_{m_n}(\theta_0, \tilde\psi)$ by $\sqrt{n}\mathbb{P}_n \tilde\ell_{m_n} + o_{P^n_{m_n}}(1)$ in both of the sandwiching bounds in (4.2) and (4.3), proves the claim. ☐

4.2 Random quadratic expansions

The log concavity assumption that we worked under in the previous section can be substituted with a consistency assumption. Suppose that for any random sequence $\tilde\theta_n$ such that $\tilde\theta_n = \theta_0 + o_{P^n_{m_n}}(1)$ for some subsequence (m_n), with $\tilde h_n = \tilde\theta_n - \theta_0$, we have

$$\log \mathrm{pl}_{m_n,n}(\tilde\theta_n) - \log \mathrm{pl}_{m_n,n}(\theta_0) = \tilde h_n^{\mathrm{t}} \sum_{i=1}^n \tilde\ell_{m_n}(X_i) - \tfrac{1}{2}n\tilde h_n^{\mathrm{t}} J_{m_n}\tilde h_n + r_n(\tilde\theta_n), \tag{4.4}$$

where $r_n(\tilde\theta_n) = o_{P^n_{m_n}}(\sqrt{n}\|\tilde h_n\| + n\|\tilde h_n\|^2 + 1)$. This gives a proposition that is similar to Proposition 4.1, replacing log concavity with consistency.

Proposition 4.3. *Suppose Assumptions 1–4 hold, and let $\widehat\theta_{m,n}$ be the maximiser of $\mathrm{pl}_{m,n}(\theta)$. If $\widehat\theta_{m_n,n} = \theta_0 + o_{P^n_{m_n}}(1)$ and (4.4) holds, then (3.6) holds, and consequently $\widehat\theta_{m_n,n}$ is best regular in $\mathcal{P}$.*

The following theorem, which is analogous to Theorem 4.2, provides conditions under which the expansion given in (4.4) holds.

Theorem 4.4. *Assume that $t \mapsto l_m(t, \theta, \gamma)(x)$ is twice continuously differentiable for each m, (θ, γ) and x, with derivatives $\dot l_m$ and $\ddot l_m$; and that there is a subsequence (m_n) such that for any random sequence $\tilde\theta_n$ with $\tilde\theta_n = \theta_0 + o_{P^n_{m_n}}(1)$, we have*

$$\sqrt{n}\mathbb{P}_n \dot l_{m_n}(\theta_0, \tilde\theta_n, \widehat\gamma(\tilde\theta_n)) = \sqrt{n}\mathbb{P}_n \tilde\ell_{m_n} + o_{P^n_{m_n}}(\sqrt{n}\|\tilde\theta_n - \theta_0\| + 1), \tag{4.5}$$

and

$$\mathbb{P}_n \ddot l_{m_n}(s_n, \tilde\theta_n, \widehat\gamma(\tilde\theta_n)) = -J_{m_n} + o_{P^n_{m_n}}(1), \tag{4.6}$$

for any random sequence $s_n = \theta_0 + o_{P^n_{m_n}}(1)$. Then (4.4) holds.

Proposition 4.3 and Theorem 4.4 are sieved versions of Corollary 1 and Theorem 1 (respectively) in Murphy and van der Vaart (2000). As they are proven by making what are essentially notational adjustments to the proofs of the cited results, we defer their proofs to Appendix B. In Appendix B we also provide sufficient conditions for the assumptions made in Theorem 4.4.

5 Applications

In this section we continue the example of the partially linear model, and also apply our theory to the Cox regression model. Both models satisfy the background assumptions on the semiparametric model (Assumption 1 and 2), consequently we concentrate on the assumptions made on the parametric approximations, that is Assumptions 3 and 4, along with (3.6). For both models we take, for simplicity, $\theta \in \Theta \subset \mathbb{R}$, and η a real valued s-times continuously differentiable function on the unit interval. The parametric approximations employed take the form

$$\beta_m(z)^{\mathrm{t}}\gamma = (T_m^{-1}\gamma)(z) \tag{5.1}$$

where for each m, $\beta_m = (\beta_{m,1}, \dots, \beta_{m,k_m})^{\mathrm{t}}$ is a collection of orthonormal functions, and $\gamma = (\gamma_1, \dots, \gamma_{k_m})^{\mathrm{t}} \in \Gamma_m \subset \mathbb{R}^{k_m}$ are coefficients such that $\beta_m^{\mathrm{t}}\gamma \to \eta$ in $L^2([0,1], \nu)$, where ν is an appropriate finite measure.[2] To not overburden the notation, the vectors γ are not indexed by m. Some of the details of the two following examples are left to appendices C.1 and C.2.

5.1 The partially linear model

We have n i.i.d. replicates of $X = (W, Y, Z)$, for an outcome Y and covariates (W, Z) with values in $\mathbb{R} \times [0,1]$. The observation X stems from the model $P_0 = P_{\theta_0,\eta_0}$, where each member $P_{\theta,\eta} \in \mathcal{P}$ has density

$$p_{\theta,\eta}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\tfrac{1}{2}(y - \eta(z) - \theta w)^2/\sigma^2) f_{W,Z}(w,z),$$

with respect to Lebesgue measure, and $f_{W,Z}(w,z)$ is the joint density of (W,Z). In other words, given $(W,Z) = (w,z)$, $Y = \eta(z) + \theta w + \varepsilon$, with ε and (W,Z) independent. Let P_Z be the law of Z. We assume that that $\mathbb{E}\,|W|^{2+\delta} < \infty$ for some $\delta > 0$. The densities in $\mathcal{P}_m$ are of the same form, with η replaced by $\beta_m^{\mathrm{t}}\gamma$. The score functions under $\mathcal{P}_m$ are

$$\dot{\ell}_{\theta,T_m^{-1}\gamma}(x) = \frac{w}{\sigma^2}(y - \beta_m(z)^{\mathrm{t}}\gamma - \theta w), \quad \text{and} \quad \dot{v}_{\theta,T_m^{-1}\gamma}(x) = \frac{\beta_m(z)}{\sigma^2}(y - \beta_m(z)^{\mathrm{t}}\gamma - \theta w). \tag{5.2}$$

Due to the orthonormality of $\beta_m = (\beta_{m,1}, \dots, \beta_{m,k_m})^{\mathrm{t}}$, the Fisher information matrix under $\mathcal{P}_m$ takes an appealing form, in particular $i_{m,11} = \sigma^{-2} I_{k_m}$ where I_{k_m} is the k_m-dimensional identity matrix, and $i_{m,01} = \sigma^{-2}\,\mathbb{E}\,W\beta_m(Z)^{\mathrm{t}}$. The efficient score under $\mathcal{P}_m$ is therefore

$$\tilde{\ell}_{\theta,T_m^{-1}\gamma}(x) = \frac{1}{\sigma^2}\big(w - \beta_m(z)^{\mathrm{t}}\mathbb{E}\,[W\beta_m(Z)]\big)(y - \beta_m(z)^{\mathrm{t}}\gamma - \theta w).$$

Define $b_0(z) = \mathbb{E}\,(W \mid Z = z)$ and $b_m(z) = \beta_m(z)^{\mathrm{t}}\mathbb{E}\,[W\beta_m(Z)] = \sum_{j=1}^{k_m}\beta_{m,j}(z)\langle\beta_{m,j}, b_0\rangle$. As the β_m form an orthonormal basis of $L_2([0,1], P_Z)$, we have $b_m \to b_0$ in $L^2([0,1], P_Z)$. We first turn to Assumption 3. The log-likelihood ratio of P_m^n with respect to P_0^n can be written

$$\log\frac{\mathrm{d}P_{m_n}^n}{\mathrm{d}P_0^n} = \frac{1}{\sqrt{n}}\sum_{i=1}^n h_{m,n}(Z_i)(\varepsilon_i/\sigma) - \frac{1}{2n}\sum_{i=1}^n h_{m,n}(Z_i)^2,$$

in terms of $h_{m,n}(z) = (\sqrt{n}/\sigma)(\beta_m(z)^{\mathrm{t}}\gamma_0 - \eta_0(z))$ and $\varepsilon_i = Y_i - \eta_0(Z_i) - \theta_0 W_i$. Assume that there is a subsequence $(m_n)_{n\geq 1}$ and a function h such that $h_{m_n,n} \to h$ in $L^2(P_0)$.[3] Under this assumption it follows from the fact that the data are i.i.d. with ε independent of Z and the law of large numbers that, with $g(x) = h(z)(y - \eta_0(z) - \theta_0 w)/\sigma$, we have

$$\log\frac{\mathrm{d}P_{m_n}^n}{\mathrm{d}P_0^n} = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i) - \tfrac{1}{2}\mathbb{E}_0\,g(X_1)^2 + o_{P_0^n}(1), \tag{5.3}$$

[2] This is always possible, see e.g. Theorem 14.3.1 in Szegő (1975).

[3] Conditions under which this holds are given as (pl3) & (pl4) in Appendix section C.1.

as $\mathbb{E}_0\, h(Z)^2 = \mathbb{E}_0\, h(Z)^2(\varepsilon/\sigma)^2$. To conclude that Assumption 3 holds, we need also show that $\mathbb{E}_0\, g(X)\tilde{\ell}(X) = 0$, which requires that we determine the form of the efficient score. Working formally, we expect that $\tilde{\ell}_{\theta_0, T_m^{-1}\gamma_0}$ will converge to

$$u(x) := \frac{1}{\sigma^2}\big(w - b_0(z)\big)\big(y - \eta_0(z) - \theta_0 w\big).$$

We verify that this is true in $L^1(P_0)$ in Appendix C.1. As $\tilde{\ell}_{\theta_0, T_m^{-1}\gamma_0}(X) \sim \sigma^{-2}(W - b_m(Z))\varepsilon$ under P_m and

$$\mathbb{E}\,(|W|^{2+\delta})\,\mathbb{E}\,(|\varepsilon|^{2+\delta}) < \infty \quad \text{and} \quad \mathbb{E}\,[\varepsilon^2(b_m(Z) - b_0(Z))^2] = \sigma^2\mathbb{E}\,[(b_m(Z) - b_0(Z))^2] \to 0, \quad (5.4)$$

it follows that $\tilde{\ell}_{\theta_0, T_{m_n}^{-1}\gamma_0}$ is square uniformly P_{m_n}-integrable. Since also $P_{m_n} \to P_0$ in total variation by (5.3) (via (3.2)), Lemma A.2 in the appendix yields

$$\lim_{n\to\infty} \int \|\tilde{\ell}_{\theta_0, T_{m_n}^{-1}\gamma_0} p_{m_n}^{1/2} - u p_0^{1/2}\|^2 \, \mathrm{d}\mu = 0.$$

Since the (3.4) and (3.5) also hold (as shown in Appendix C.1), it follows from Corollary 3.4 that $u = \tilde{\ell}$ P_0-almost surely. We now have an expression for the efficient score, and can verify that, indeed

$$\mathbb{E}_0\, g(X)\tilde{\ell}(X) = \frac{1}{\sigma^2}\mathbb{E}\, h(Z)(W - b_0(Z)) = \frac{1}{\sigma^2}\mathbb{E}\, h(Z)\mathbb{E}\,\{(W - b_0(Z)) \,|\, Z\} = 0,$$

and we conclude that Assumption 3 holds. Equations (3.4) and (3.5) are shown to hold in Appendix C.1. Thus Assumption 4 also holds by Lemma 3.1. From the developments so far, we also see that the semiparametric efficient information is $J = \sigma^{-2}(\mathbb{E}\, W^2 - \|b_0\|^2) = \sigma^{-2}\mathbb{E}\,\big(W^2 - (\mathbb{E}\,[W \,|\, Z])^2\big)$, and we note that

$$J_{m_n} = \mathbb{E}_{m_n}\, \tilde{\ell}_{\theta_0, T_{m_n}^{-1}\gamma_0}(X)^2 = \frac{1}{\sigma^2}\big(\mathbb{E}\, W^2 - \sum_{j=1}^{k_{m_n}} |\langle \beta_{m_n,j}, b_0\rangle|^2\big) \to \frac{1}{\sigma^2}\big(\mathbb{E}\, W^2 - \|b_0\|^2\big) = J,$$

which may also be verified directly using Parseval's identity.

Let $(\widehat{\theta}_{m,n}, \widehat{\gamma}_{m,n})$ be the maximum likelihood estimator under $\mathcal{P}_m$. To establish that $\widehat{\theta}_{m_n,n}$ is best regular, we show that $h \mapsto \log \mathrm{pl}_{m,n}(\theta_0 + h/\sqrt{n}) - \log \mathrm{pl}_{m,n}(\theta_0)$ is concave and admits an expansion as in (4.1). The conclusion then follows from Proposition 4.1. The value achieving the supremum in $\sup_{\gamma \in \Gamma_m} L_{m,n}(\theta, \gamma)$ is the least squares solution $\widehat{\gamma}(\theta) = B_{m,n}^{-1}\sum_{i=1}^n \beta_m(Z_i)(Y_i - \theta W_i)$, where $B_{m,n} = \sum_{i=1}^n \beta_m(Z_i)\beta_m(Z_i)^{\mathrm{t}}$. The log profile likelihood can then be expressed as

$$\log \mathrm{pl}_{m,n}(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^n \big(Y_i - \breve{Y}_{m,n,i} - \theta(W_i - \breve{W}_{m,n,i})\big)^2,$$

where $\breve{W}_{m,n,i} = \beta_m(Z_i)^{\mathrm{t}} B_{m,n}^{-1}\sum_{j=1}^n \beta_m(Z_j)W_j$ and $\breve{Y}_{m,n,i} = \beta_m(Z_i)^{\mathrm{t}} B_{m,n}^{-1}\sum_{j=1}^n \beta_m(Z_j)Y_j$, and we see that $\log \mathrm{pl}_{m,n}(\theta)$ is concave in θ. The expansion (4.1) is verified Appendix C.1 under conditions (pl1) and (pl2). In summary, under the basic conditions given here along with (pl1)–(pl4), the conditions of Proposition 4.1 are satisfied and consequently the estimator $\widehat{\theta}_{m_n,n}$ is best regular in $\mathcal{P}$.

5.2 Cox regression

We have n i.i.d. replicates of $X = (T, \delta, W)$ observed over $[0, 1]$, with $T = \min(T', C)$ and $\delta = I(T' \leq C)$, where the lifetime T' and the censoring time C are independent given W, and T' given $W = w$ follows a Cox model, i.e., its hazard rate is of the form $\eta(t)\exp(\theta w)$ where $\theta \in \Theta \subset \mathbb{R}$

for simplicity, and η belongs to the space $\mathcal{H}$ of one time continuously differentiable functions $\eta\colon [0,1] \to (0,\infty)$. Notation wise, we will not distinguish between random and observed δ. As above, $P_0 = P_{\theta_0,\eta_0}$ is the true data generating process, while $P_m = P_{\theta_0,T_m^{-1}\gamma_0}$ are the parametric approximations. Here we take the basis function

$$\beta_{m,j}(t) = I_{V_{m,j}}(t)/\Delta_m, \quad \text{for } j = 1,\dots,k_m, \tag{5.5}$$

where $V_{m,1},\dots,V_{m,k_m}$ are disjoint intervals whose union make up $[0,1]$, and each interval is of length Δ_m (or proportional to it). With this basis, a natural choice is to work under the coefficients $\gamma_0 = (\gamma_{0,1},\dots,\gamma_{0,k_m})^{\mathrm{t}}$ where $\gamma_{0,j} = \eta_0((j-1)\Delta_m)\Delta_m$ for $j = 1,\dots,k_m$. The observation X stems from $P_0 \in \mathcal{P}$, where each member $P_{\theta,\eta} \in \mathcal{P}$ has density

$$p_{\theta,\eta}(x) = \big(e^{\theta w}\eta(t)\big)^{\delta} \exp\{-e^{\theta w}H_\eta(t)\}\bar{G}_{C\,|\,W}(t\,|\,w)^{\delta} g_{C\,|\,W}(t\mid w)^{1-\delta} f_W(w),$$

with respect to Lebesgue measure μ on $[0,1]^2 \times \mathbb{R}$; $\bar{G}_{C\,|\,W} = 1 - G_{C\,|\,W}$ where $G_{C\,|\,W}$ is the conditional c.d.f. of C given W, with density $g_{C\,|\,W}$; f_W is the marginal density of W; and $H_\eta(t) = \int_0^t \eta(s)\,\mathrm{d}s$ is the cumulative baseline hazard. The densities of the distributions in $\mathcal{P}$ are of the same form with η replaced by $\beta_m^{\mathrm{t}}\gamma$. We note that the mapping $\eta \mapsto p_{\theta_0,\eta}$ is continuous for each θ and x. The score functions under $\mathcal{P}_m$ are

$$\dot{\ell}_{\theta,T_m^{-1}\gamma}(x) = w(\delta - e^{\theta w}H_{T_m^{-1}\gamma}(t)), \quad \text{and} \quad \dot{v}_{\theta,T_m^{-1}\gamma}(x) = \delta\frac{\beta_m(t)}{\beta_m(t)^{\mathrm{t}}\gamma} - e^{\theta w}\int_0^1 \beta_m(s)\,\mathrm{d}s. \tag{5.6}$$

To compute the second moments of these functions it is convenient to appeal to the counting process approach to survival analysis. Introduce the counting process $N(t) = \delta I(T \le t)$, the at-risk process $Y(t) = I(T \ge t)$, and let $(\mathcal{F}_t^n)_{t\in[0,1]}$ be the filtration generated by the i.i.d. replicates $\{(N_i(s), Y_i(s))_{s\le t}, W_i\}_{1\le i\le n}$ of $\{(N(s), Y(s))_{s\le t}, W\}$, with $(W_1,\dots,W_n)$ measurable with respect to $\mathcal{F}_0^n$. We assume that the processes

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\eta_0(s)\exp(\theta_0 X_i)\,\mathrm{d}s,$$

are square integrable martingales with respect to $((\mathcal{F}_t^n)_{t\in[0,1]}, P_0^n)$, and that

$$M_i^{(m)}(t) = N_i(t) - \int_0^t Y_i(s)\beta_m(s)^{\mathrm{t}}\gamma_0\exp(\theta_0 X_i)\,\mathrm{d}s,$$

are square integrable martingales with respect to $((\mathcal{F}_t^n)_{t\in[0,1]}, P_m^n)$. Using the relations $\delta f(T) = \int_0^1 f(s)\,\mathrm{d}N(s)$ for any f and $f(T) = \int_0^1 Y(s)\,\mathrm{d}f(s)$ for any f such that $f(0) = 0$, the score functions above are $\dot{\ell}_m(X) = WM^{(m)}(1)$ and $\dot{v}_m(X) = \int_0^1 \beta_m(s)(\beta_m(s)^{\mathrm{t}}\gamma_0)^{-1}\,\mathrm{d}M^{(m)}(s)$ when evaluated in $(\theta_0, T_m^{-1}\gamma_0)$. Using the Itô isometry, we then find

$$i_{m,01} = \mathbb{E}_m\,\dot{\ell}_m(X)\dot{v}_m(X)^{\mathrm{t}} = \int_0^1 \beta_m(t)s_m^{(1)}(t)\,\mathrm{d}t,$$

and

$$i_{m,11} = \mathbb{E}_m\,\dot{v}_m(X)\dot{v}_m(X)^{\mathrm{t}} = \int_0^1 \frac{\beta_m(t)\beta_m(t)^{\mathrm{t}}}{\beta_m(t)^{\mathrm{t}}\gamma_0}s_m^{(0)}(t)\,\mathrm{d}t.$$

Inserting the basis function $\beta_m = (\beta_{m,1},\dots,\beta_{m,k_m})^{\mathrm{t}}$ given in (5.5), gives that the efficient score under $\mathcal{P}_m$ is

$$\tilde{\ell}_m(x) = \sum_{j=1}^{k_m} I_{V_{m,j}}(t)\bigg(w - \frac{\int_{V_{m,j}} s_m^{(1)}(u)\,\mathrm{d}u}{\int_{V_{m,j}} s_m^{(0)}(u)\,\mathrm{d}u}\bigg)\{\delta - e^{\theta w}H_m(t)\},$$

where $s_m^{(k)}(t) = \mathbb{E}_m\, Y(t)W^k \exp(W\theta_0)$ for $k = 0, 1, 2$, and $H_m(t) \int_0^t \beta_m(s)^{\mathrm{t}}\gamma_0 \,\mathrm{d}s$. In Appendix C.2 we show that under certain continuity assumptions on the $s_m^{(k)}$ functions,

$$\tilde{\ell}_m(x) \to \int_0^1 \big(w - \frac{s^{(1)}(t)}{s^{(0)}(t)}\big)\{\delta - e^{\theta w} H_0(t)\} := u(x),$$

for μ-almost all x, where $H_0(t) = \int_0^t \eta_0(s)\,\mathrm{d}s$ and $s^{(k)} = \mathbb{E}_0\, Y(t)W^k \exp(W\theta_0)$ are the limits of $s_m^{(k)}$ as $m \to \infty$. Both score functions above can be expressed using the counting process martingales, as

$$\tilde{\ell}_m(X) = \sum_{j=1}^{k_m} \int_{V_{m,j}} \big(W - \frac{\int_{V_{m,j}} s_m^{(1)}(u)\,\mathrm{d}u}{\int_{V_{m,j}} s_m^{(0)}(u)\,\mathrm{d}u}\big)\,\mathrm{d}M^{(m)}(t), \quad \text{and} \quad \tilde{\ell}(X) = \int_0^1 \big(w - \frac{s^{(1)}(t)}{s^{(0)}(t)}\big)\,\mathrm{d}M(t),$$

and these are the expressions we use in showing the convergence

$$J_m = \mathbb{E}_m\, \tilde{\ell}_m(X)^2 \to \int_0^1 \Big(\frac{s^{(2)}(t)}{s^{(0)}(t)} - \big(\frac{s^{(1)}(t)}{s^{(0)}(t)}\big)^2\Big) s^{(0)}(t)\eta_0(t)\,\mathrm{d}t = \mathbb{E}_0\, u(X)^2.$$

Since also (3.4) and (3.5) hold (as we show in Appendix C.2), it follows from the above and Corollary 3.4 that Assumption 4 is satisfied, that $u = \tilde{\ell}\ P_0$-almost surely, and that $J = \mathbb{E}_0\, u(X)^2$, where $\tilde{\ell}$ and J are the efficient score and information under P_0, respectively. Next, we turn to Assumption 3. Writing $\bar{N}_n(t) = n^{-1}\sum_{i=1}^n N_i(t)$, $\bar{Y}_n(t) = n^{-1}\sum_{i=1}^n Y_i(t)$, and $\bar{M}_n(t) = n^{-1}\sum_{i=1}^n M_i(t)$, the log-likelihood ratio can be written

$$\log \frac{\mathrm{d}P_m^n}{\mathrm{d}P_0^n} = \sqrt{n}\int_0^1 g_{m,n}\,\mathrm{d}\bar{M}_n - \tfrac{1}{2}\int_0^1 g_{m,n}^2\,\mathrm{d}\bar{N}_n + r_{m,n},$$

where $g_{m,n}(t) = h_{m,n}(t)/\eta_0(t)$ with $h_{m,n}(t) = \sqrt{n}(\beta_m(t)^{\mathrm{t}}\gamma_0 - \eta_0(t))$, while the remainder is $r_{m,n} = \int_0^1 g_{m,n}^2 R(g_{m,n}/\sqrt{n})\,\mathrm{d}\bar{N}_n$, and R is the function defined via the Taylor expansion $\log(1 + a) = a - \frac{1}{2}a^2 + a^2 R(a)$. For any subsequence (m_n) such that $h_{m_n,n} \to h$ in $L^2(\mathrm{d}t)$ we have that $\log(\mathrm{d}P_{m_n}^n/P_0^n) = \sqrt{n}\int_0^1 g\,\mathrm{d}\bar{M}_n(t) - \frac{1}{2}\int_0^1 g^2\,\mathrm{d}\bar{N}_n(t) + o_{P_0}(1)$, where $g = h/\eta_0$. We see that $\mathbb{E}_0 g(X)\tilde{\ell}(X) = 0$, and conclude that Assumption 3 is satisfied. Note that with the basis functions in (5.5) we only have $h_{m_n,n} \to h$ in $L^2(\mathrm{d}t)$ when $\sqrt{n}\Delta_{m_n} \to 0$, and consequently $h = 0$ (see the introduction in Appendix C for the details).

It remains to verify that there is a subsequence (m_n) satisfying Assumption 3 such that the process $A_{m_n,n}(h) = \mathrm{pl}_{m_n,n}(\theta_0 + h/\sqrt{n}) - \mathrm{pl}_{m,n}(\theta_0)$ admits a quadratic expansion as in (4.1). For θ fixed, the maximum likelihood estimator for γ is $\widehat{\gamma}(\theta) = (\widehat{\gamma}_1(\theta), \ldots, \widehat{\gamma}_{k_m}(\theta))^{\mathrm{t}}$, where

$$\widehat{\gamma}_j(\theta) = \frac{\int_{V_{m,j}} \mathrm{d}\bar{N}_n(t)}{\int_{V_{m,j}} S_n^{(0)}(t,\theta)\,\mathrm{d}t},$$

with $S_n^{(k)}(t,\theta) = n^{-1}\sum_{i=1}^n Y_i(t)W_i^k \exp(\theta W_i)$ for $k = 0, 1, 2$. This gives the log profile likelihood function

$$\log \mathrm{pl}_{m,n}(\theta) = \sum_{j=1}^{k_m}\sum_{i=1}^{n} \int_{V_{m,j}} \big\{\theta W_i - \log \int_{V_{m,j}} S_n^{(0)}(s,\theta)\,\mathrm{d}s\big\}\,\mathrm{d}N_i(t) + \text{const.},$$

which is concave in θ (see Andersen and Gill (1982, p. 1103)). The log profile likelihood ratio process is

$$A_{m,n}(h) = \frac{h}{\sqrt{n}}\sum_{i=1}^{k_m}\sum_{i=1}^{n} \int_{V_{m,j}} \big(W_i - \frac{\int_{V_{m,j}} S_n^{(1)}(s,\theta_0)\,\mathrm{d}s}{\int_{V_{m,j}} S_n^{(0)}(s,\theta_0)\,\mathrm{d}s}\big)\,\mathrm{d}M_i^{(m)}(t)$$
$$- \frac{h^2}{2}\sum_{i=1}^{k_m}\Big\{\frac{\int_{V_{m,j}} S_n^{(2)}(s,\xi_n)\,\mathrm{d}s}{\int_{V_{m,j}} S_n^{(0)}(s,\xi_n)\,\mathrm{d}s} - \Big(\frac{\int_{V_{m,j}} S_n^{(1)}(s,\xi_n)\,\mathrm{d}s}{\int_{V_{m,j}} S_n^{(0)}(s,\xi_n)\,\mathrm{d}s}\Big)^2\Big\}\{\bar{N}_n(t_j) - \bar{N}_n(t_{j-1})\}.$$

Note that in this expression $\mathrm{d}M_i^{(m)}$ can be replaced by $\mathrm{d}M_i$. In Appendix C.2 we show that under a standard uniform convergence assumption on $S_n^{(k)}$ (essentially the same as imposed in Andersen and Gill (1982)), we have that the process $A_{m_n,n}(h)$ satisfies the quadratic expansion in (4.1) for any subsequence (m_n). From Proposition 4.1 we can then conclude that provided (m_n) is such that $\sqrt{n}\Delta_{m_n} \to 0$, then the maximiser of $\mathrm{pl}_{m_n,n}(\theta)$ is best regular in $\mathcal{P}$.

5.3 Efficiency of the Cox partial likelihood estimator

The conclusion of the previous section leads to a new, relatively simple, and rather parametric proof of the efficiency of the Cox partial likelihood estimator (see for example (Bickel et al., 1998, pp. 77–81) or Kosorok (2008) for proofs that differ from ours). Let $L_n^{\mathrm{cox}}(\theta)$ be Cox's partial likelihood function (Gill (1984) is an excellent introduction). We may then define the process $A_n^{\mathrm{cox}}(h) = \log L_n^{\mathrm{cox}}(\theta_0 + h/\sqrt{n}) - \log L_n^{\mathrm{cox}}(\theta_0)$, which is concave in h, and is given by

$$A_n^{\mathrm{cox}}(h) = \sum_{i=1}^n \int_0^1 \big(X_i \frac{h}{\sqrt{n}} - \log \frac{S_n(t, \theta_0 + h/\sqrt{n})}{S_n(t, \theta_0)}\big)\,\mathrm{d}N_i(t).$$

Under standard regularity conditions in the Cox regression setting, see (cx1)–(cx4) in Appendix C.2, the process $A_n^{\mathrm{cox}}(h)$ admits the expansion $A_n^{\mathrm{cox}}(h) = hn^{-1/2}\sum_{i=1}^n \tilde{\ell}(X_i) - \frac{1}{2}h^2 J + o_{P_0}(1)$, where $\tilde{\ell}$ and J are as defined in Section 5.2. Since $h \mapsto A_n^{\mathrm{cox}}(h)$ is concave, the Basic Corollary in Hjort and Pollard (1993) entails that the maximiser of $L_n^{\mathrm{cox}}(\theta)$, say $\widehat{\theta}_n^{\mathrm{cox}}$, satisfies (3.7), that is

$$\sqrt{n}(\widehat{\theta}_n^{\mathrm{cox}} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n J^{-1}\tilde{\ell}(X_i) + o_{P_0^n}(1).$$

But from the application of our theory in Section 5.2 we know that J and $\tilde{\ell}$ are the *efficient* semiparametric information and score, and consequently, we can conclude that the Cox estimator is best regular in $\mathcal{P}$.

A Technical results

Theorem A.1. *Let H be a Hilbert space, $h_n, h \in H$, and L_n, L closed (proper) linear subspaces of H. Let $g_n := \Pi(h_n|L_n)$ and $g := \Pi(h|L)$. If* (i) *$h_n \to h$ and* (ii) *for each $f \in L$, there is a sequence $(f_n)_{n\in\mathbb{N}}$ and a $N \in \mathbb{N}$ such that $f_n \to f$ and $f_n \in L_n$ for $n \geq N$, then $g_n \to g$.*

Proof. Let Π_n be the orthogonal projection onto L_n and Π that onto L. First suppose $h_n = h$ ($n \in \mathbb{N}$). As $(g_n)_{n\in\mathbb{N}}$ is bounded, any subsequence contains a weakly convergent subsequence, say $g_{n_k} \rightharpoonup g^\star$. By self-adjointness and idempotency

$$\langle g_{n_k}\,,\, g_{n_k}\rangle = \langle \Pi_{n_k} h\,,\, \Pi_{n_k} h\rangle = \langle h\,,\, \Pi_{n_k} h\rangle \to \langle h\,,\, g^\star\rangle\,. \tag{A.1}$$

Let $f \in L$. By hypothesis there are $(f_n)_{n\in\mathbb{N}}$ with $f_n \to f$ and $f_n \in L_n$ for $n \geq N_1$. So $f_{n_k} \to f$ and $f_{n_k} \in L_{n_k}$ for $k \geq K_1$. Since $h - \Pi_{n_k} h \rightharpoonup h - g^\star$, by Proposition 16.7 in Royden and Fitzpatrick (2010) and the fact that $h - g_{n_k} \in L_{n_k}^\perp$ for each k, $\langle h - g^\star\,,\, f\rangle = \lim_{k\to\infty} \langle h - g_{n_k}\,,\, f_{n_k}\rangle = 0$. Hence $g^\star = \Pi h = g$. By self-adjointness and idempotency of Π and (A.1), $\lim_{k\to\infty}\langle g_{n_k}\,,\, g_{n_k}\rangle = \langle h\,,\, \Pi h\rangle = \langle \Pi h\,,\, \Pi h\rangle = \langle g\,,\, g\rangle$, and hence $g_{n_k} \to g$ by the Radon–Riesz Theorem. As the initial subsequence was arbitrary, $g_n \to g$. To complete the proof, for $h_n \to h$ an arbitrary convergent sequence, $\|g_n - g\| \leq \|h_n - h\| + \|\Pi_n h - \Pi h\|$. The first right hand side term is $o(1)$ by assumption; the second by the case with $h_n = h$. □

Lemma A.2. *Let $P_n, P_0 \ll \mu$ with densities p_n, p_0. Suppose that* (i) *$P_n \to P_0$ in total variation;* (ii) *f_n converges to f_0 in P_0-probability; and* (iii) *f_n is uniformly square P_n-integrable and f_0 is square P_0-integrable. Then $\lim_{n\to\infty}\int \big(f_n p_n^{1/2} - f_0 p_0^{1/2}\big)^2\,\mathrm{d}\mu = 0$.*

Proof. Expansion of the square yields

$$\int (f_n p_n^{1/2} - f_0 p_0^{1/2})^2 \, \mathrm{d}\mu = \int f_n^2 \, \mathrm{d}P_n + \int f_0^2 \, \mathrm{d}P_0 - 2 \int f_n f_0 p_n^{1/2} p_0^{1/2} \, \mathrm{d}\mu.$$

Combining (i), (ii), (iii) and a version of Vitali's convergence theorem (Corollary 2.9 in Feinberg et al., 2016) gives $\lim_{n\to\infty} \int f_n^2 \, \mathrm{d}P_n = \int f_0^2 \, \mathrm{d}P_0$. Hence the proof will be complete if we show that $\lim_{n\to\infty} \int f_n f_0 p_n^{1/2} p_0^{1/2} \, \mathrm{d}\mu = \int f_0^2 \, \mathrm{d}P_0$.

Let Q_n be the probability measure with μ-density $q_n \coloneqq c_n p_n^{1/2} p_0^{1/2}$ where c_n is the normalising constant. We have $c_n^{-1} \coloneqq \int p_n^{1/2} p_0^{1/2} \, \mathrm{d}\mu = 1 - \frac{1}{2} \int (p_n^{1/2} - p_0^{1/2})^2 \, \mathrm{d}\mu \to 1$ as $\int (p_n^{1/2} - p_0^{1/2})^2 \, \mathrm{d}\mu \le 2 d_{\mathrm{TV}}(P_n, P_0)$, with d_{TV} the total variation distance (e.g., Strasser, 1985, Lemma 2.15). Similarly,

$$\begin{aligned} \int |q_n - p_0| \, \mathrm{d}\mu &\le \int p_0^{1/2} |c_n| |p_n^{1/2} - p_0^{1/2}| \, \mathrm{d}\mu + \int p_0 |c_n - 1| \, \mathrm{d}\mu \\ &\le |c_n| \Big(\int p_0 \, \mathrm{d}\mu \Big)^{1/2} \Big(\int (p_n^{1/2} - p_0^{1/2})^2 \, \mathrm{d}\mu \Big)^{1/2} + |c_n - 1| \to 0, \end{aligned}$$

implying that $d_{\mathrm{TV}}(Q_n, P_0) \to 0$. Now, let $g_n \coloneqq f_n f_0$. Note that $h_n \coloneqq g_n + |g_n| \ge 0$. Since $|g_n| \to f_0^2$ in P_0-probability, we have $\liminf_{n\to\infty} \int |g_n| \, \mathrm{d}Q_n \ge \int f_0^2 \, \mathrm{d}P_0$ by a version of Fatou's lemma (Corollary 2.3 in Feinberg et al., 2016). Additionally, by the Cauchy–Schwarz inequality

$$\limsup_{n\to\infty} \int |g_n| \, \mathrm{d}Q_n \le \limsup_{n\to\infty} c_n \Big(\int f_n^2 \, \mathrm{d}P_n \Big)^{1/2} \Big(\int f_0^2 \, \mathrm{d}P_0 \Big)^{1/2} = \int f_0^2 \, \mathrm{d}P_0.$$

In consequence,

$$\lim_{n\to\infty} \int |g_n| \, \mathrm{d}Q_n = \int f_0^2 \, \mathrm{d}P_0. \tag{A.2}$$

An entirely analogous argument applied to $h_n \ge 0$ shows that

$$\lim_{n\to\infty} \int h_n \, \mathrm{d}Q_n = 2 \int f_0^2 \, \mathrm{d}P_0. \tag{A.3}$$

Combining equations (A.2) and (A.3) yields

$$\lim_{n\to\infty} \int f_n f_0 \, \mathrm{d}Q_n = \lim_{n\to\infty} \int h_n - |g_n| \, \mathrm{d}Q_n = \int f_0^2 \, \mathrm{d}P_0.$$

Since $\int f_n f_n p_n^{1/2} p_0^{1/2} \, \mathrm{d}\mu = c_n^{-1} \int f_n f_0 \, \mathrm{d}Q_n$ and $c_n \to 1$, this completes the proof. □

We next provide proofs of Lemmas 3.2 and 3.3.

Proof of Lemma 3.2. Since Assumption 3 holds, so does (3.2) and hence $P_{m_n} \to P_0$ in total variation. Apply Lemma A.2. □

Proof of Lemma 3.3. In case (i) we have $f_{m_n} p_{m_n}^{1/2} \to f_0 p_0^{1/2}$ in μ-measure. The conclusion follows from Proposition 2.29 in van der Vaart (1998) as $\limsup_n \int (f_{m_n} p_{m_n}^{1/2})^2 \, \mathrm{d}\mu \le \int (f_0 p_0^{1/2})^2 \, \mathrm{d}\mu < \infty$. In case (ii) the result follows directly from Proposition S3.1 in the supplementary material to Hoesch et al. (2024). □

B Additional details for Section 4.2

B.1 Proofs of Proposition 4.3 and Theorem 4.4

Proof of Proposition 4.3. Let $\Delta_n \coloneqq n^{-1/2} \sum_{i=1}^n \tilde{\ell}_{m_n}(X_i)$ and $\widehat{h}_n \coloneqq \sqrt{n}(\hat{\theta}_{n,m_n} - \theta)$. Applying (4.4) with $\tilde{\theta}_n = \widehat{\theta}_{m_n,n}$ gives

$$\log \mathrm{pl}_{n,m_n}(\widehat{\theta}_{m_n,n}) = \log \mathrm{pl}_{n,m_n}(\theta_0) + \widehat{h}_n^{\mathrm{t}} \Delta_n - \frac{1}{2} \widehat{h}_n^{\mathrm{t}} J_{m_n} \widehat{h}_n + r_n(\hat{\theta}_{n,m_n}),$$

where $r_n(\widehat{\theta}_{m_n,n}) = o_{P^n_{m_n}}(\|\widehat{h}_n\| + 1)^2$. Since $J_{m_n} = \mathbb{E}_{m_n}\tilde{\ell}_{m_n}(X)\tilde{\ell}_{m_n}(X)^{\mathrm{t}}$, Assumptions 2 & 4 ensure that $J^{-1}_{m_n}\Delta_n = O_{P^n_{m_n}}(1)$. By (4.4) with $\tilde{\theta}_n = \theta + n^{-1/2}J^{-1}_{m_n}\Delta_n$ and $r_n(\tilde{\theta}_n) = o_{P^n_{m_n}}(1)$,

$$\log \mathrm{pl}_{n,m_n}(\tilde{\theta}_n) = \log \mathrm{pl}_{n,m_n}(\theta_0) + \Delta_n^{\mathrm{t}} J^{-1}_{m_n}\Delta_n - \frac{1}{2}\Delta_n^{\mathrm{t}} J^{-1}_{m_n}\Delta_n + r_n(\tilde{\theta}_n).$$

By definition, $\log \mathrm{pl}_{n,m_n}(\widehat{\theta}_{m_n,n})$ is larger than $\log \mathrm{pl}_{n,m_n}(\tilde{\theta}_n)$. Hence

$$\widehat{h}_n^{\mathrm{t}}\Delta_n - \frac{1}{2}\widehat{h}_n^{\mathrm{t}} J_{m_n}\widehat{h}_n - \frac{1}{2}\Delta_n^{\mathrm{t}} J^{-1}_{m_n}\Delta_n \geq -o_{P^n_{m_n}}(\|\widehat{h}_n\| + 1)^2.$$

The left hand side of the preceding display is equal to the left hand side of

$$-\frac{1}{2}\left(\widehat{h}_n - J^{-1}_{m_n}\Delta_n\right)^{\mathrm{t}} J_{m_n}\left(\widehat{h}_n - J^{-1}_{m_n}\Delta_n\right) \leq -\frac{c}{2}\|\widehat{h}_n - J^{-1}_{m_n}\Delta_n\|^2,$$

where $0 < c \leq \lambda_{\min}(J_{m_n})$ for all sufficiently large n, where $\lambda_{\min}$ is the smallest eigenvalue of J_{m_n}, and we use that these are bounded below for n sufficiently large. Combination of the preceding two displays yields

$$\|\widehat{h}_n - J^{-1}_{m_n}\Delta_n\| = o_{P^n_{m_n}}(\|\widehat{h}_n\| + 1).$$

Since $\|J^{-1}_{m_n}\Delta_n\| = O_{P^n_{m_n}}(1)$, the triangle inequality implies that

$$\|\widehat{h}_n\| \leq \|\widehat{h}_n - J^{-1}_{m_n}\Delta_n\| + \|J^{-1}_{m_n}\Delta_n\| = o_{P^n_{m_n}}(\|\widehat{h}\|_n + 1) + O_{P^n_{m_n}}(1) = O_{P^n_{m_n}}(1).$$

Using this in the penultimate display yields $\|\widehat{h}_n - J^{-1}_{m_n}\Delta_n\| = o_{P^n_{m_n}}(1)$, which implies (3.6). □

Remark 2. In the proof of Proposition 4.3 the expansion (4.4) is used for two different $\tilde{\theta}_n$, namely $\tilde{\theta}_n = \widehat{\theta}_{m_n,n}$ and $\tilde{\theta}_n \coloneqq \breve{\theta}_{m_n,n} \coloneqq n^{-1/2}J^{-1}_{m_n}\Delta_n$, where $\Delta_n = n^{-1/2}\sum_{i=1}^n \tilde{\ell}_{m_n}(X_i)$. Under Assumptions 1-4,

$$\sqrt{n}(\breve{\theta}_{m_n,n} - \theta_0) = J^{-1}_{m_n}\Delta_n = O_{P^n_{m_n}}(1),$$

as noted in the proof. Therefore, if one establishes that also $\sqrt{n}(\widehat{\theta}_{m_n,n} - \theta) = O_{P_{m_n}}(1)$, then it suffices to show that $r_n(\tilde{\theta}_n) = o_{P_{m_n}}(1)$ for all $\tilde{\theta}_n$ such that $\sqrt{n}(\tilde{\theta}_n - \theta) = O_{P_{m_n}}(1)$.

Proof of Theorem 4.4. Let $\tilde{h}_n = \tilde{\theta}_n - \theta_0$. For fixed $\tilde{\psi}$, a Taylor expansion yields

$$\mathbb{P}_n l_{m_n}(\tilde{\theta}_n, \tilde{\psi}) - \mathbb{P}_n l_{m_n}(\theta_0, \tilde{\psi}) = \tilde{h}_n^{\mathrm{t}}\mathbb{P}_n \dot{l}_{m_n}(\theta_0, \tilde{\psi}) + \tfrac{1}{2}\tilde{h}_n^{\mathrm{t}}\mathbb{P}_n \ddot{l}_{m_n}(s_n, \tilde{\psi})\tilde{h}_n,$$

for a s_n between $\tilde{\theta}_n$ and θ_0. Multiplying through by n and replacing $\mathbb{P}_n\ddot{l}_{m_n}(s_n, \tilde{\psi})$ by $-J_{m_n} + o_{P^n_{m_n}}(1)$ and $\sqrt{n}\mathbb{P}_n\dot{l}_{m_n}(\theta_0, \tilde{\psi})$ by $\sqrt{n}\mathbb{P}_n\tilde{\ell}_{m_n} + o_{P^n_{m_n}}(\sqrt{n}\|\tilde{h}_n\| + 1)$ on the right hand side gives

$$n\mathbb{P}_n l_{m_n}(\tilde{\theta}_n, \tilde{\psi}) - n\mathbb{P}_n l_{m_n}(\theta_0, \tilde{\psi}) = \tilde{h}_n^{\mathrm{t}}\sum_{i=1}^n \tilde{\ell}_{m_n}(X_i) - \tfrac{1}{2}n\tilde{h}_n^{\mathrm{t}} J_{m_n}\tilde{h}_n + o_{P^n_{m_n}}((\sqrt{n}\|\tilde{h}_n\| + 1)^2),$$

for $\tilde{\psi}$ equal to either $(\tilde{\theta}_n, \widehat{\gamma}(\tilde{\theta}_n)$ or $(\theta_0, \widehat{\gamma}(\theta_0))$. Applying the sandwiching bounds in (4.2) and (4.3) gives (4.4). □

B.2 Sufficient conditions for Theorem 4.4

As demonstrated in the examples in the main text, in some models, our approach of working under the parametric models P_m allows (3.6) to be established directly. For cases where this is not possible, Proposition 4.3 and Theorem 4.4 provide a general result for such estimators. This result is based on Murphy and van der Vaart (2000) with the key difference being that we consider a sieved profile likelihood in which, at each step m, true least-favourable submodels necessarily exist. This avoids the requirement to construct 'approximate least favourable submodels' as in Murphy and van der Vaart (2000). Nevertheless, as the theoretical analysis of this sieve profile likelihood estimator proceeds very similarly to the analysis of the semiparametric profile likelihood estimator considered in Murphy and van der Vaart (2000), Theorem 4.4 states the result under high level conditions. Here we give lower-level structural conditions which imply the conditions (4.5) and (4.6) required by Theorem 4.4. The conditions are similar to those given by Murphy and van der Vaart (2000), but require some adjustment as we deal with a sequence of objective functions. This is the price we pay for the guaranteed existence of true least-favourable submodels.

The following lemma splits condition (4.5) into a no-bias condition and a condition relating to the empirical process $\mathbb{G}_n$, here defined by $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - P_{m_n} f)$.

Lemma B.1. *Suppose that for any $\tilde{\theta}_n = \theta_0 + o_{P^n_{m_n}}(1)$,*

$$\mathbb{E}_{m_n} \dot{l}_{m_n}(\theta_0, \tilde{\theta}_n, \widehat{\gamma}(\tilde{\theta}_n)) = o_{P^n_{m_n}}(\|\tilde{\theta}_n - \theta_0\| + n^{-1/2}) \ , \tag{B.1}$$

and

$$\mathbb{G}_n \dot{l}_{m_n}(\theta_0, \tilde{\theta}_n, \widehat{\gamma}(\tilde{\theta}_n)) = \mathbb{G}_n \tilde{\ell}_{m_n} + o_{P^n_{m_n}}(1). \tag{B.2}$$

Then (4.5) holds.

Proof. By (B.1) and (B.2) and since $\mathbb{E}_{m_n} \tilde{\ell}_{m_n} = 0$,

$$\begin{aligned}
\sqrt{n}\mathbb{P}_n \dot{l}_{m_n}(\theta_0, \tilde{\theta}_n, \widehat{\gamma}(\tilde{\theta}_n)) &= \mathbb{G}_n \dot{l}_{m_n}(\theta_0, \tilde{\theta}_n, \widehat{\gamma}(\tilde{\theta}_n)) + \sqrt{n}\mathbb{E}_{m_n} \dot{l}_{m_n}(\theta_0, \tilde{\theta}_n, \widehat{\gamma}(\tilde{\theta}_n)) \\
&= \mathbb{G}_n \tilde{\ell}_{m_n} + o_{P^n_{m_n}}(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1) \\
&= \sqrt{n}\mathbb{P}_n \tilde{\ell}_{m_n} + o_{P^n_{m_n}}(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1).
\end{aligned}$$

□

Condition (B.1) is a 'no-bias' condition, cf. the discussion in Murphy and van der Vaart (2000). Condition (B.2) can be shown to hold if the nuisance parameter estimator $\widehat{\gamma}$ satisfies a consistency condition and a stochastic equicontinuity type condition is satisfied.

For any m, $\gamma = \gamma_m \in \mathbb{R}^{k_m}$ may be viewed as an eventually zero sequence in $\mathbb{R}^{\mathbb{N}}$. Let $\Gamma^{\mathbb{N}}_m$ denote the subset of $\mathbb{R}^{\mathbb{N}}$ corresponding to vectors in Γ_m. We equip $\mathbb{R}^{\mathbb{N}}$ with a topology, τ, and shall need the following consistency condition to hold: For any $\tilde{\theta}_n = \theta_0 + o_{P^n_{m_n}}(1)$,

$$\widehat{\gamma}(\tilde{\theta}_n) - \gamma_0 = o_{P^n_{m_n}}(1). \tag{B.3}$$

That is, for any neighbourhood $\mathcal{U}$ of zero in $(\mathbb{R}^{\mathbb{N}}, \tau)$

$$\lim_{n\to\infty} P_{m_n}(\widehat{\gamma}(\tilde{\theta}_n) - \gamma_0 \in \mathcal{U}) = 1,$$

for $\widehat{\gamma}(\tilde{\theta}_n), \gamma_0$ considered as elements in $\mathbb{R}^{\mathbb{N}}$. The topology τ is arbitrary. It may, for instance, be that induced by $\mathcal{H}$ (if $\mathcal{H}$ is a topological vector space). However, for this condition to be useful, the topology needs to be strong enough to imply certain continuity conditions.

Lemma B.2. *Suppose that $\tilde{\theta}_n = \theta_0 + o_{P^n_{m_n}}(1)$ and that (B.3) holds. Let $\mathcal{V}$ be a neighbourhood of $(0,0) \in \mathbb{R}^p \times \mathbb{R}^{\mathbb{N}}$. Suppose that that for each $\varepsilon, \upsilon > 0$ there are random $\Delta_n(\varepsilon, \upsilon) \geq 0$ and $N(\varepsilon, \upsilon)$ such that if $n \geq N(\varepsilon, \upsilon)$ then* (i) *$P_{m_n}(\Delta_n(\epsilon, \upsilon) > \upsilon) < \varepsilon$ and* (ii)

$$\sup_{v \in \mathcal{V}_n} \|\mathbb{G}_n \dot{l}_{m_n}(\theta_0, \psi_0 + v) - \mathbb{G}_n \dot{l}_{m_n}(\theta_0, \psi_0)\| \leq \Delta_n(\varepsilon, \upsilon), \qquad \psi_0 := (\theta_0, \gamma_0),$$

where $\mathcal{V}_n := \{v \in \mathcal{V} : \psi_0 + v \in \Theta \times \Gamma^{\mathbb{N}}_{m_n}\}$. Then (B.2) holds.

Proof. Let $X_n(v) \coloneqq \mathbb{G}_n \dot{l}_{m_n}(\theta_0, \psi_0 + v)$ and note that $X_n(0) = \mathbb{G}_n \dot{l}_{m_n}(\theta_0, \psi_0) = \mathbb{G}_n \tilde{\ell}_{m_n}$. Fix $\varepsilon, \upsilon > 0$. By (B.3) there is a N_1 such that $n \geq N_1$ implies

$$P_{m_n}(\hat{v}_n \in \mathcal{V}) \geq 1 - \varepsilon/2.$$

where $\hat{v}_n \coloneqq (\tilde{\theta}_n - \theta_0, \widehat{\gamma}(\tilde{\theta}_n) - \gamma_0)$. Note that, by definition, if $\hat{v}_n \in \mathcal{V}$ then $\hat{v}_n \in \mathcal{V}_n$. Using this, (i) and (ii) we have for all $n \geq \max\{N_1, N(\varepsilon/2, \upsilon)\}$,

$$\begin{aligned} P_{m_n}(\|X_n(\hat{v}) - X_n(0)\| > \upsilon) &\leq P_{m_n}(\hat{v}_n \notin \mathcal{V}) + P_{m_n}(\{\|X_n(\hat{v}) - X_n(0)\| > \upsilon\} \cap \{\hat{v}_n \in \mathcal{V}_n\}) \\ &< \varepsilon/2 + P_{m_n}\left(\sup_{v \in \mathcal{V}_n} \|X_n(v) - X_n(0)\| > \upsilon\right) \\ &\leq \varepsilon/2 + P_{m_n}(\Delta_n(\varepsilon/2, \upsilon) > \upsilon) < \varepsilon. \end{aligned}$$

□

Finally, condition (4.6) is an approximate information equality. Provided the information equality approximately holds in the least-favourable parametric submodels, this will hold under continuity and moment conditions.

Lemma B.3. *Suppose that $\tilde{\vartheta}_n = \theta_0 + o_{P^n_{m_n}}(1)$, $\tilde{\theta}_n = \theta_0 + o_{P^n_{m_n}}(1)$, that (B.3) holds and*

(i) $P_m[\dot{l}_m(\theta_0, \theta_0, \gamma_0)\dot{l}_m(\theta_0, \theta_0, \gamma_0)^{\mathrm{t}}] = -P_m[\ddot{l}_m(\theta_0, \theta_0, \gamma_0)] + o(1)$ *as $m \to \infty$;*

(ii) $(\tilde{\vartheta}_n, \tilde{\theta}_n, \widehat{\gamma}(\tilde{\theta}_n)) - (\theta_0, \theta_0, \gamma_0) = o_{P^n_{m_n}}(1)$ *implies that*

$$\lim_{n \to \infty} \mathbb{E}_{m_n} \|\ddot{l}_{m_n}(\tilde{\vartheta}_n, \tilde{\theta}_n, \widehat{\gamma}(\tilde{\theta}_n)) - \ddot{l}_{m_n}(\theta_0, \theta_0, \gamma_0)\| = 0;$$

(iii) $\|\ddot{l}_{m_n}(\theta_0, \theta_0, \gamma_0)\|$ *is uniformly P_{m_n}-integrable.*

Then (4.6) holds.

Proof. By the construction of the least favourable submodels and (i),

$$J_m = \mathbb{E}_m \tilde{\ell}_m \tilde{\ell}_m^{\mathrm{t}} = \mathbb{E}_m \dot{l}_m(\theta_0, \theta_0, \gamma_0)\dot{l}_m(\theta_0, \theta_0, \gamma_0)^{\mathrm{t}} = -\mathbb{E}_m \ddot{l}_m(\theta_0, \theta_0, \gamma_0) + o(1).$$

Write $Y_{n,i} \coloneqq \ddot{l}_{m_n}(\theta_0, \theta_0, \gamma_0)(X_i)$ and $\tilde{Y}_{n,i} \coloneqq \ddot{l}_{m_n}(\tilde{\vartheta}_n, \tilde{\theta}_n, \widehat{\gamma}(\tilde{\theta}_n))(X_i)$. Then (4.6) holds if both (a) $n^{-1}\sum_{i=1}^n (\tilde{Y}_{n,i} - Y_{n,i})$ and (b) $n^{-1}\sum_{i=1}^n (Y_{n,i} - \mathbb{E}_{m_n} Y_{n,i})$ are $o_{P^n_{m_n}}(1)$. Requirement (a) follows as condition (ii) implies $\mathbb{E}_{m_n}\|\tilde{Y}_{n,i} - Y_{n,i}\| \to 0$ and hence $n^{-1}\sum_{i=1}^n (\tilde{Y}_{n,i} - Y_{n,i})$ converges to zero in probability by Markov's inequality. For (b), under condition (iii) $n^{-1}\sum_{i=1}^n (Y_{n,i} - \mathbb{E}_{m_n} Y_{n,i})$ converges to zero by the weak law of large numbers (see, e.g., Gut, 1992). □

C The applications

In our application to the Cox model, the approximations to η_0 take the form $\beta_m(x)^{\mathrm{t}}\gamma_0 = \sum_{j=1}^{k_m} I_{V_{m,j}}(x)\eta_0(x_j)$, writing $x_j = j\Delta_m$, so that x_j is the lower limit of the interval $V_{m,j} = [x_{j-1}, x_j) = [(j-1)\Delta_m, j\Delta_m]$, and $V_{m,1} \cup \cdots \cup V_{m,k_m} = [0,1]$ for each m. In the verification of Assumption 3 the sequence of functions $(h_{m,n})$ defined by $h_{m,n}(x) = \sqrt{n}\big(\beta_m(x)^{\mathrm{t}}\gamma_0 - \eta_0(x)\big)$, plays an essential role. With the piecewise constant basis functions inserted, this function is

$$h_{m,n}(x) = \sqrt{n}\sum_{j=1}^{k_m}\big(\eta_0(x_{j-1}) - \eta_0(x)\big).$$

If we assume that η_0 is two times differentiable, then a Taylor expansion of each summand yields $\eta_0(x_{j-1}) - \eta_0(x) = \eta_0'(x)(x_{j-1} - x) + \frac{1}{2}\eta_0''(\tilde{x}_j)(x_{j-1} - x)^2$ for some $\tilde{x}_j \in V_{m,j}$. Here x_{j-1} is the biggest number on the grid $\{j\Delta_m \colon j = 1, \ldots, k_m\}$ smaller than x. Thus for $x \in V_{m,j}$

$$x_{j-1} = \lfloor x/\Delta_m \rfloor \Delta_m,$$

where $\lfloor y \rfloor = \max\{z \in \mathbb{N} \colon z \leq y\}$ is the floor function. The sawtooth function $\{y\}$ is the defined as the fractional part of y, that is $\{y\} = y - \lfloor y \rfloor$, so $0 \leq \{y\} < 1$. But then, for $x \in V_{m,j}$

$$x_{j-1} - x = \lfloor x/\Delta_m \rfloor \Delta_m - x = (x/\Delta_m - \{x/\Delta_m\})\Delta_m - x = -\{x/\Delta_m\}\Delta_m.$$

Inserting this back into the Taylor expansion above, and then inserting the Taylor expansion back into $h_{m,n}$, we see that

$$h_{m,n}(x) = -\sqrt{n}\{x/\Delta_m\}\Delta_m \eta_0'(x) + \tfrac{1}{2}\sqrt{n}\{x/\Delta_m\}^2\Delta_m^2 \sum_{j=1}^{k_m} I_{V_{m,j}}(x)\eta_0''(\tilde{x}_j). \tag{C.1}$$

As its name suggests, the sawtooth function $\{x/\Delta_m\}$ does not converge, it jumps up and down between zero and one. Since the sawtooth function is bounded by 1 however,

$$|\sqrt{n}\{x/\Delta_m\}\Delta_m \eta_0'|(x) \leq \sqrt{n}\Delta_m |\eta_0'(x)|,$$

and we see that the first term on the right in (C.1) converges to zero if and only if (m_n) is such that $\sqrt{n}\Delta_{m_n} \to 0$. Provided $\sum_{j=1}^{k_m} I_{V_{m,j}}(x)\eta_0''(\tilde{x}_j) = O(1)$ as $m \to \infty$, the second term will tend to zero under this condition as well.

If we assume that η_0 is continuously differentiable, as we do in the paper, a similar argument gives that $h_{m_n,n}(x) \to 0$ if and only if $\sqrt{n}\Delta_{m_n} \to 0$, and that $h \equiv 0$ is the only possible limit of $h_{m,n}(x)$ as $m, n \to \infty$. We note that in this case,

$$\int_0^1 h_{m_n,n}(x)^2 \,\mathrm{d}x \leq n\Delta_{m_n}^2 \sum_{j=1}^{k_m} \int_{V_{m,j}} \eta_0'(\tilde{x}_j)^2 \,\mathrm{d}x \leq n\Delta_{m_n}^2 \sup_{x \in [0,1]} \eta_0'(x)^2 \to 0,$$

using that $\eta_0'(x)$ is continuous on $[0,1]$, and hence uniformly bounded on this interval. Finally, we note that if $f(x)$ is a continuous density on $[0,1]$ and is bounded below, then a sequence converges in $L^2([0,1], \mathrm{d}x)$ if and only if it converges in $L^2([0,1], f(x)\mathrm{d}x)$.

C.1 The partially linear model

In this section we first verify that (3.4) and (3.5) hold in the setup of Section 5.1. This entails, by way of Lemma 3.1, that the partially linear model satisfies Assumption 4. We then establish that $\tilde{\ell}_{\theta_0, T_m^{-1}\gamma_0}$ converges to the semiparametric efficient score. Finally, we consider the special case of piecewise constant basis functions for which we establish (3.6).

Consider submodels of the form $\tau \mapsto p_{\theta_0 + a\tau, \eta_0 + b\tau}$ where $a \in \mathbb{R}$ and $b \in B \subset \mathcal{H}$. Differentiating with respect to τ, and evaluating in $\tau = 0$

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \log p_{\theta + a\tau, \eta + b\tau}\big|_{\tau=0} = a\dot{\ell}_{\theta,\eta} + \frac{b(z)}{\sigma^2}(y - \eta(z) - \theta w),$$

where $\dot{\ell}_{\theta,\eta}(x) = \sigma^{-2} w(y - \eta(z) - \theta w)$. We now show that these are the limits, in the sense of (3.4) and (3.5), of their parametric counterparts $\dot{\ell}_{\theta, T_m^{-1}\gamma}$ and $\dot{v}_{\theta, T_m^{-1}\gamma}$ in (5.2).

We will use Lemma A.2, and note for future reference that (a) $d_{\mathrm{TV}}(P_{m_n}, P_0) \to 0$ by Assumption 3 and (3.2), and (b) that for any $b \in B$, the sequence defined by $\tilde{b}_m(z) = \beta_m^{\mathrm{t}}(z)\mathbb{E}\{b(Z)\beta_m(Z)\}$ is such that $\tilde{b}_m \to b$ in $L^2(P_Z)$. Note also that under P_m, $\dot{\ell}_m(X) \sim \sigma^{-2} W\varepsilon$ and with $\tilde{\gamma} = \mathbb{E}\{b(Z)\beta_m(Z)\}$, $\tilde{\gamma}^{\mathrm{t}} \dot{v}_m(X) \sim \sigma^{-2}\tilde{b}_m(Z)\varepsilon$. The uniform square P_{m_n}-integrability required by Lemma A.2 then follows from (5.4). Additionally, it is straightforward to check that $\dot{\ell}_{\theta,\eta}(X)$ and $b(Z)(Y - \eta(Z) - \theta W)$ are square integrable under P_0. Finally we show $L^1(P_0)$ convergence of the parametric scores to the semiparametric scores. In particular,

$$\int \|\dot{\ell}_m - \dot{\ell}\| \,\mathrm{d}P_0 = \sigma^{-2}\,\mathbb{E}\,\|W(\eta_0(Z) - \eta_m(Z))\| \to 0,$$

by the Cauchy–Schwarz inequality as $\eta_m(z) := \beta_m(z)^{\mathrm{t}}\gamma_0 \to \eta_0(z)$ in $L^2(P_Z)$. Similarly with $\tilde{\gamma} = \mathbb{E}\{b(Z)\beta_m(Z)\}$ as above,

$$\int \|\tilde{\gamma}^{\mathrm{t}}\dot{v}_m - Db\| \,\mathrm{d}P_0 = \sigma^{-2}\mathbb{E}\,\|\epsilon(\tilde{b}_m(Z) - b(Z)) + \tilde{b}_m(Z)(\eta_0(Z) - \eta_m(Z))\| \to 0,$$

as $\tilde{b}_m(z) \to b(z)$ in $L^2(P_Z)$. Applying Lemma A.2 then verifies (3.4) and (3.5) of Lemma 3.1, meaning that Assumption 4 holds.

To verify that $\tilde{\ell}_m$ converges to $\tilde{\ell}$ in $L^1(P_0)$ as claimed in Section 5.1 we note that similarly,

$$\int \|\tilde{\ell}_m - u\| \,\mathrm{d}P_0 = \sigma^{-2}\mathbb{E}\{\|\varepsilon(b_0(Z) - b_m(Z)) + (W - b_m(Z))(\eta_0(Z) - \eta_m(Z))\} \to 0.$$

To establish (4.1), we use Theorem 4.2. We have

$$l_m(t, \theta, \gamma)(x) = -\frac{1}{2\sigma^2}\left(y - wt - \beta_m(z)^{\mathrm{t}}(\gamma + \mathbb{E}\{\beta_m(Z)W\}(\theta - t))\right)^2 + C(w, z),$$

for $C(w, z)$ a term which does not depend on (t, θ, γ). Thus

$$\dot{l}_m(t, \theta, \gamma)(x) = \sigma^{-2}\left(y - wt - \beta_m(z)^{\mathrm{t}}(\gamma + \mathbb{E}\{\beta_m(Z)W\}(\theta - t))\right)\left(w - \beta_m(z)^{\mathrm{t}}\mathbb{E}\{\beta_m(Z)W\}\right),$$

and

$$\ddot{l}_m(t, \theta, \gamma)(x) = -\sigma^{-2}\left(w - \beta_m(z)^{\mathrm{t}}\mathbb{E}\{\beta_m(Z)^{\mathrm{t}}W\}\right)^2 = -\sigma^{-2}\left(w - b_m(z)\right)^2.$$

We first note that $\mathbb{E}_m\ddot{l}_m(t, \theta_0, \gamma_0) = -J_m$. Thus for the second part of the condition in Theorem 4.2 it suffices to show that $n^{-1}\sum_{i=1}^n \sigma^{-2}(W_i - b_{m_n}(Z_i))^2 - J_{m_n} = o_{P^n_{m_n}}(1)$. This follows from the weak law of large numbers as the $\sigma^{-2}(W_i - b_{m_n}(Z_i))^2$ are uniformly P_{m_n}-integrable (see Gut (1992)). To establish the first part of Theorem 4.2, let $h \in \mathbb{R}$ and set $\tilde{\theta}_n = \theta_0 + h/\sqrt{n}$. We then have

$$\dot{l}_{m_n}(\theta_0, \tilde{\theta}_n, \hat{\gamma}(\tilde{\theta}_n))(X_i) = \frac{1}{\sigma^2}\left(Y_i - W_i\theta_0 - \beta_{m_n}(Z_i)^{\mathrm{t}}\left[\hat{\gamma}(\tilde{\theta}_n) + \mathbb{E}\{\beta_m(Z)W\}\frac{h}{\sqrt{n}}\right]\right)(W_i - b_m(Z_i)),$$

with $\hat{\gamma}(\tilde{\theta}_n) = B^{-1}_{m_n,n}n^{-1}\sum_{i=1}^n \beta_m(Z_i)(Y_i - (\theta_0 + \frac{h}{\sqrt{n}})W_i)$, and $B_{m_n,n} = n^{-1}\sum_{i=1}^n \beta_{m_n}(Z_i)\beta_{m_n}(Z_i)^{\mathrm{t}}$ as previously defined. Therefore, the difference

$$\begin{aligned}&\dot{l}_{m_n}(\theta_0, \tilde{\theta}_n, \hat{\gamma}(\tilde{\theta}_n))(X_i) - \tilde{\ell}_{m_n}(X_i)\\ &= \sigma^{-2}[W_i - b_{m_n}(Z_i)]\beta_{m_n}(Z_i)^{\mathrm{t}}\left[(\hat{\gamma}(\theta_0) - \gamma) + \frac{h}{\sqrt{n}}\left(\mathbb{E}\{\beta_m(Z)W\} - B^{-1}_{m_n,n}\frac{1}{n}\sum_{i=1}^n \beta_m(Z_i)W_i\right)\right].\end{aligned}$$

In consequence, to verify the remaining condition of Theorem 4.2, it suffices to show that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n (W_i - b_0(Z_i))\beta_{m_n}(Z_i)^{\mathrm{t}}(\hat{\gamma}(\theta_0) - \gamma) &= o_{P_{m_n}}(1)\\ \frac{1}{\sqrt{n}}\sum_{i=1}^n (b_0(Z_i) - b_{m_n}(Z_i))\beta_{m_n}(Z_i)^{\mathrm{t}}(\hat{\gamma}(\theta_0) - \gamma) &= o_{P_{m_n}}(1) \qquad \text{(C.2)}\\ \frac{1}{n}\sum_{i=1}^n (W_i - b_{m_n}(Z_i))\beta_{m_n}(Z_i)^{\mathrm{t}}\left[\mathbb{E}\{\beta_{m_n}(Z)W\} - \tilde{B}^{-1}_{m_n,n}\frac{1}{n}\sum_{i=1}^n \beta_{m_n}(Z_i)W_i\right] &= o_{P_{m_n}}(1)\end{aligned}$$

In order to verify these conditions, we impose the following assumptions, which may be relaxed if more conditions are imposed upon W, see below. Let $\xi_m := \sup_{z\in[0,1]}\|\beta_{m_n}(z)\|$ and let a_n be such that $\|b_0(Z) - b_{m_n}(Z)\|_{L_2(\nu)} \le a_n$. (Upper bounds on a_n are available from approximation theory under, for example, smoothness conditions on b_0.)

(pl1) $\mathbb{E}(W^2 \mid Z) \le C$, P_Z-almost surely;

(pl2) $k_{m_n}^2/\sqrt{n} \to 0$, $\xi_{m_n} k_{m_n}/\sqrt{n} \to 0$ and $a_n \sqrt{k_{m_n}} \xi_{m_n} \to 0$.

For the first, note that as for $k \neq i$

$$\mathbb{E}\{(W_i - b_0(Z_i))\beta_{m_n,j}(Z_i)(W_k - b_0(Z_k))\beta_{m_n,j}(Z_k)\} = 0$$

then for some positive constant C_0

$$\begin{aligned} P_{m_n}\big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (W_i - b_0(Z_i))\beta_{m_n,j}(Z_i) \ge M\big) &\le M^{-1}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\{(W_i - b_0(Z_i))^2 \beta_{m_n,j}(Z_i)^2\} \\ &\le 2M^{-1}\mathbb{E}\{(\mathbb{E}(W^2 \mid Z) + b_0(Z)^2)\beta_{m_n,j}(Z)^2\} \\ &\le C_0 M^{-1} \end{aligned}$$

which can be made arbitrary small by taking M large enough. We also have

$$\sum_{j=1}^{k_{m_n}} |\widehat{\gamma}_j(\theta_0) - \gamma_j| \le \sqrt{k_{m_n}}\big(\sum_{j=1}^{k_{m_n}} |\widehat{\gamma}_j(\theta_0) - \gamma_j|^2\big)^{1/2} = O_{P_{m_n}}(k_{m_n}/\sqrt{n}) \tag{C.3}$$

by the Cauchy–Schwarz inequality and Theorem 4.1 in Belloni et al. (2015). It follows that by taking $M = M_n = k_{m_n}$ above and using the union bound that

$$\sum_{j=1}^{k_{m_n}} (\hat{\gamma}_j(\theta_0) - \gamma_j)\frac{1}{\sqrt{n}}\sum_{i=1}^n \beta_{m_n,j}(Z_i)(W_i - b_0(Z_i)) = O_{P_{m_n}}(k_{m_n}^2 n^{-1/2}) = o_{P_{m_n}}(1),$$

which verifies the first term. The second holds as by Theorem 4.1 in Belloni et al. (2015) again,

$$\begin{aligned} \big|\frac{1}{\sqrt{n}}\sum_{i=1}^n (b_0(Z_i) - b_{m_n}(Z_i))\beta_{m_n}(Z_i)^{\mathrm{t}}(\hat{\gamma}(\theta_0) - \gamma)\big| &\le \sqrt{n}\xi_{m_n}\|\hat{\gamma}(\theta_0) - \gamma\|\frac{1}{n}\sum_{i=1}^n |b_0(Z_i) - b_{m_n}(Z_i)| \\ = O_{P_{m_n}}\left(\sqrt{n}\xi_{m_n}\sqrt{k_{m_n}} n^{-1/2} a_n\right) &= O_{P_{m_n}}\left(\xi_{m_n}\sqrt{k_{m_n}} a_n\right) = o_{P_{m_n}}(1) \end{aligned}$$

as

$$P_{m_n}\big(\frac{1}{n}\sum_{i=1}^n |b_0(Z_i) - b_{m_n}(Z_i)| \ge M a_n\big) \le M^{-1} a_n^{-1}\|b_0 - b_{m_n}\|_{L^1(P_Z)} \le M^{-1}.$$

Finally, we can rewrite the left hand side of the third condition in (C.2) as $n^{-1}\sum_{i=1}^n (W_i - b_{m_n}(Z_i))\beta_{m_n}(Z_i)^{\mathrm{t}}(\pi - \widehat{\pi})$, where $\pi = \mathbb{E}\{\beta_{m_n}(Z)W\}$ and $\widehat{\pi}$ is its empirical counterpart. Using, once more, Theorem 4.1 in Belloni et al. (2015) we have

$$\begin{aligned} \big|\frac{1}{n}\sum_{i=1}^n (W_i - b_{m_n}(Z_i))\beta_{m_n}(Z_i)^{\mathrm{t}}(\pi - \widehat{\pi})\big| &\le \xi_{m_n}\|\pi - \widehat{\pi}\|_2 \frac{1}{n}\sum_{i=1}^n |W_i - b_{m_n}(Z_i)| \\ &= O_{P_{m_n}}(\xi_{m_n}\sqrt{k_{m_n}/n}), \end{aligned}$$

as

$$P_{m_n}\big(n^{-1}\sum_{i=1}^n |W_i - b_{m_n}(Z_i)| \ge M\big) \le M^{-1}(\mathbb{E}|W| + \mathbb{E}|b_0(Z)| + a_n).$$

Here we remark that the rate conditions in (pl2) can be relaxed if we impose additional conditions on W. In particular, if we replace (pl2) with

(pl2*) W is sub-exponential with parameters (ν, α), $\xi_{m_n} k_{m_n}/\sqrt{n} \to 0$ and $a_n \sqrt{k_{m_n}} \xi_{m_n} \to 0$

then the second and third terms can be shown to be $o_{P_{m_n}}(1)$ exactly as before but we can refine the argument relating to the first term. In particular, (C.3) holds as above, but the probabilistic bound on $n^{-1/2}\sum_{i=1}^n(W_i - b_0(Z_i))\beta_{m_n,j}(Z_i)$ can be improved. Since $b_0(Z_i)$ is bounded by, say C_0 and $\beta_{m_n,j}(Z_i)$ is bounded by ξ_{m_n}, $(W - b_0(Z))\beta_{m_n,j}(Z)$ is also subexponential with parameters $(\xi_{m_n}(\nu + C_0), \alpha)$. Hence for all $t \geq 0$ the Bernstein-type bound below holds (e.g. Wainwright, 2019, Eq. (2.18), p. 29)

$$P_{m_n}\left(\left|\frac{1}{n}\sum_{i=1}^n(W_i - b_0(Z_i))\beta_{m_n,j}(Z_i)\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{nt}{2\alpha}, \frac{nt^2}{2\xi_{m_n}^2(\nu+C_0)^2}\right\}\right).$$

Hence taking $t = t_n = 2(\nu + C_0)\xi_{m_n}\sqrt{\log k_{m_n}}/\sqrt{n}$ we have that

$$\begin{aligned}k_{m_n}P_{m_n}\left(\left|\frac{1}{n}\sum_{i=1}^n(W_i - b_0(Z_i))\beta_{m_n,j}(Z_i)\right| \geq t\right) &\leq k_{m_n}2\exp\left(-\frac{nn^{-1}4\xi_{m_n}^2(\nu+C_0)^2\log k_{m_n}}{2\xi_{m_n}^2(\nu+C_0)^2}\right)\\ &= 2k_{m_n}\exp(-2\log k_{m_n}) = 2k_{m_n}^{-1} \to 0.\end{aligned}$$

Thus by (C.3) and the union bound

$$\sum_{j=1}^{k_{m_n}}(\widehat{\gamma}_j(\theta_0) - \gamma_j)\frac{1}{\sqrt{n}}\sum_{i=1}^n\beta_{m_n,j}(Z_i)(W_i - b_0(Z_i)) = O_{P_{m_n}}(\xi_{m_n}\sqrt{\log k_{m_n}}k_{m_n}n^{-1}) = o_{P_{m_n}}(1).$$

Finally we provide an example of β_{m_n} for which $h_{m_n,n} \to h$ in $L_2(P_0)$. In particular, let β_{m_n} be (orthonormalised) B-splines of degree l with equally spaced knots. Then if

(pl3) η is s-times continuously differentiable and $l + 1 \geq s$

(pl4) $k_{m_n}^{-s}\sqrt{n} \to 0$

it follows from Theorem 20.3 in Powell (1981) that if γ is chosen optimally

$$\|\beta_{m_n}(z)^{\mathrm{t}}\gamma - \eta_0(z)\|_\infty = O(k_{m_n}^{-s}),$$

hence

$$\sqrt{n}\|\beta_{m_n}^{\mathrm{t}}\gamma - \eta_0\|_{L_2(P_Z)} = O(\sqrt{n}k_{m_n}^{-s}) = o(1)$$

and thus $h_{m_n,n} \to 0$ in $L^2(P_0)$.

Remark 3. The locally constant basis function discussed in the introduction to this appendix are easy to work with, and provide very intelligible results. Consider therefore the sequence of orthonormal basis function $\beta_{m,j}$ that take the form

$$\beta_{m,j}(z) = \frac{I_{V_{m,j}}(z)}{(\mathbb{E}\, I_{V_{m,j}}(Z))^{1/2}}, \quad j = 1, \ldots, k_m, k_m \geq 1$$

where $V_{m,j} = \{z\colon (j-1)\Delta_m \leq z < j\Delta_m\}$ for $j = 1, \ldots, k_m$, and $V_{m,1} \cup \cdots \cup V_{m,k_m} = [0, 1]$ for all m, with Δ_m a step size. Impose the following conditions:

(lc1) Z has density f_Z that is continuously differentiable and bounded below on $[0, 1]$;
(lc2) The function $z \mapsto \mathbb{E}\,(W \mid Z = z)$ is continuously differentiable;
(lc3) The function $z \mapsto \mathbb{E}\,(W^2 \mid Z = z)$ is continuously differentiable.
(lc4) $\eta_0 \in \mathcal{H}$, where $\mathcal{H}$ consists of all continuously differentiable functions $\eta\colon [0, 1] \to \mathbb{R}$.

With these basis function, one may use Theorem 4.2 to show that the quadratic expansion in (4.1) holds provided the subsequence (m_n) is chosen so that $n\Delta_{m_n} \to \infty$; and that Assumption 3, i.e., contiguity, is satisfied provided $\sqrt{n}\Delta_{m_n} \to 0$.

C.2 The Cox model

The quantities

$$S_n^{(k)}(t,\theta) = \frac{1}{n}\sum_{i=1}^n Y_i(t) W_i^k \exp(\theta W_i), \quad \text{for } k = 0,1,2,$$

play important roles. Define

$$s_{\theta,\eta}^{(k)}(t) = \mathbb{E}_{\theta,\eta}\, Y(t) W^k \exp(\theta W), \quad \text{for } k = 0,1,2,3,$$

so that due to the data being i.i.d., $s_{\theta,\eta}^{(k)}(t) = \mathbb{E}_{\theta,\eta}\, S_n^{(k)}(t,\theta)$. Recall that the conditional c.d.f. of the censoring time C given $W = w$ is $G(c\,|\,w)$, and that we define $\bar{G}(c\,|\,w) = 1 - G(c\,|\,w)$. Since T' and C are conditionally independent given W,

$$s_{\theta,\eta}^{(k)}(t) = \mathbb{E}\, \exp(-e^{\theta W} H_\eta(t)) \bar{G}(t\,|\,W) W^k \exp(\theta_0 W).$$

Write

$$s^{(k)} = s_{\theta_0,\eta_0}^{(k)}, \quad \text{and} \quad s_m^{(k)} = s_{\theta_0, T_m^{-1}\gamma_0}^{(k)},$$

as well as

$$v_{m,j} = \frac{\int_{V_{m,j}} s_m^{(2)}(s)\,\mathrm{d}s}{\int_{V_{m,j}} s_m^{(0)}(s)\,\mathrm{d}s} - \left(\frac{\int_{V_{m,j}} s_m^{(1)}(s)\,\mathrm{d}s}{\int_{V_{m,j}} s_m^{(0)}(s)\,\mathrm{d}s}\right)^2, \quad \text{and} \quad v(t) = \frac{s^{(2)}(t)}{s^{(0)}(t)} - \left(\frac{s^{(1)}(t)}{s^{(0)}(t)}\right)^2.$$

We make the following assumptions:

(cx1) $\mathbb{E}\, W^k \exp(\theta W) < \infty$ for $k = 0,1,2,3$, for all θ in a neighbourhood of θ_0;

(cx2) The baseline hazard η_0 is one time continuously differentiable and bounded below on $[0,1]$;

(cx3) In a neighbourhood U_{θ_0} of θ_0,

$$\sup_{t\in[0,1],\theta\in U} |S_n^{(k)}(t,\theta) - s_{\theta,\eta_0}^{(k)}(t)| = o_{P_0}(1),$$

for $k = 0,1,2$.

(cx4) For a neighbourhood U_{η_0} of η_0, the functions $(\theta,\eta) \mapsto s_{\theta,\eta}^{(k)}(t)$ for $k = 0,1,2,3$ are continuous on $U_{\eta_0} \times U_{\theta_0}$, uniformly in $t \in [0,1]$; they are bounded on $U_{\eta_0} \times U_{\theta_0} \times [0,1]$, and $s_{\theta,\eta}^{(0)}$ is bounded below on $U_{\eta_0} \times U_{\theta_0} \times [0,1]$, and $J = \int_0^1 v(t) s^{(0)}(t) \eta_0(t)\,\mathrm{d}t$ is positive.

Assumption (cx2), (cx3), and (cx4) are essentially Assumption A., B., and D. of the canonical Cox regression paper Andersen and Gill (1982, p. 1105), while condition (cx1) is implicitly made in that paper. The version of Rebolledo's central limit theorem presented in that paper (Andersen and Gill, 1982, Theorem I.2, p. 1115) is useful when working with counting process martingales. In our setting this theorem says the following: Let $N_1,\ldots,N_n$ be counting processes on $[0,1]$ with intensity processes $Y_i(t)\alpha_i(t)$, and assume that $\xi_{n,1},\ldots,\xi_{n,n}$ are sequences of locally bounded predictable processes, and form the locally square integrable martingales $Z_n(t) = n^{-1/2}\sum_{i=1}^n \int_0^t \xi_{n,i}(t)\{\mathrm{d}N_i(t) - Y_i(t)\alpha_i(t)\,\mathrm{d}t\}$. If $\langle Z_n, Z_n\rangle_t \to_p a(t)$ for each $t \in [0,1]$; and the Lindeberg condition holds, that is, for every $\varepsilon > 0$

$$\frac{1}{n}\sum_{i=1}^n \int_0^1 \xi_{n,i}(t)^2 Y_i(t)\alpha_i(t) I\{|\xi_{n,i}(t)| \geq \sqrt{n}\varepsilon\}\,\mathrm{d}t \to_p 0,$$

then $Z_n \rightsquigarrow Z$ in $D[0,1]$, where Z is a Gaussian martingale with $\mathrm{Cov}(Z(t),Z(s)) = a(t\wedge s)$. It is immediate from this theorem that $\sqrt{n}\bar{M}_n \rightsquigarrow Z$, where Z is a Gaussian martingale with $\langle Z, Z\rangle_t = \int_0^t s^{(0)}(s)\eta_0(s)\,\mathrm{d}s$.

In the following we write $t_j = j\Delta_m$. Adding and subtracting we have

$$\int_{V_{m,j}} s_m^{(k)}(t)\,\mathrm{d}t = \int_{V_{m,j}} s^{(k)}(t)\,\mathrm{d}t + \int_{V_{m,j}} \{s_m^{(k)}(t) - s^{(k)}(t)\}\,\mathrm{d}t$$
$$= s^{(k)}(t_{j-1})\Delta_m + \int_{V_{m,j}} \{s^{(k)}(t) - s^{(k)}(t_{j-1})\}\,\mathrm{d}t + \int_{V_{m,j}} \{s_m^{(k)}(t) - s^{(k)}(t)\}\,\mathrm{d}t.$$

Then, for $k = 0, 1, 2$, by condition (cx4)

$$|\int_{V_{m,j}} \{s_m^{(k)}(t) - s^{(k)}(t)\}\,\mathrm{d}t| \le \sup_t |s_m^{(k)}(t) - s^{(k)}(t)|\Delta_m = o(\Delta_m).$$

By the same assumption, the functions $t \mapsto s^{(k)}(t)$ are Lipschitz for $k = 0, 1, 2$, with Lipschitz constants $\sup_t |s^{(k+1)}(t)|$, thus $|\int_{V_{m,j}}\{s^{(k)}(t) - s^{(k)}(t_{j-1})\}\,\mathrm{d}t| \le \sup_t |s^{(k+1)}(t)|\Delta_m^2$ for $k = 0, 1$. Consequently,

$$\max_{j \le k_m} |\int_{V_{m,j}} s_m^{(k)}(t)\,\mathrm{d}t - s^{(k)}(t_{j-1})\Delta_m| = o(\Delta_m), \quad \text{for } k = 0, 1, \tag{C.4}$$

as $m \to \infty$, and since $s^{(0)}$ is bounded below

$$\max_{j \le k_m} |\frac{\int_{V_{m,j}} s_m^{(1)}(t)\,\mathrm{d}t}{\int_{V_{m,j}} s_m^{(0)}(t)\,\mathrm{d}t} - \frac{s^{(1)}(t_{j-1})}{s^{(0)}(t_{j-1})}| = o(1), \quad \text{as } m \to \infty. \tag{C.5}$$

We start by verifying (3.4) and (3.5). Consider the submodels $\tau \mapsto p_{\theta_0 + a\tau, \eta_0 + b\tau}$ where $a \in \mathbb{R}$ and $b \in B \subset \mathcal{H}$. Differentiating with respect to τ, and evaluating in $\tau = 0$

$$\frac{\mathrm{d}}{\mathrm{d}\tau} p_{\theta + a\tau, \eta + b\tau}\big|_{\tau = 0} = a\dot{\ell}_{\theta,\eta} + Db,$$

where $\dot{\ell}_{\theta,\eta} = w(\delta - e^{w\theta}H_\eta(t))$ and $(Db)(x) = \delta b(t)\eta(t)^{-1} - e^{\theta w}\int_0^t b(s)\,\mathrm{d}s$. Let (γ) be a growing parameter sequence such that $\beta_m^{\mathrm{t}}\gamma \to b$ almost surely (at least along a subsequence). Comparing the two score function above with their parametric counterparts given in (5.6), we see that $\dot{\ell}_m(x) \to \dot{\ell}(x)$ and $\gamma^{\mathrm{t}}\dot{v}_m(x) \to (Db)(x)$ for μ-almost all x, by continuity. In terms of the counting process martingales, $\dot{\ell}_m(X) = WM^{(m)}(1)$ and $\dot{\ell}(X) = WM(1)$ are square integrable martingales under P_m and P_0, respectively. Then since $s_m^{(0)}$ and η_0 are bounded on $[0, 1]$,

$$|\mathbb{E}_m\, \dot{\ell}_m(X)^2 - \mathbb{E}_0\, \dot{\ell}(X)^2| \lesssim W^2\{\int_0^1 |\beta_m(t)^{\mathrm{t}}\gamma_0 - \eta_0(t)|\,\mathrm{d}t + \int_0^1 |s_m^{(0)}(t) - s^{(0)}(t)|\,\mathrm{d}t\} \to 0,$$

as $m \to \infty$. Similarly,

$$|\mathbb{E}_m\,(\gamma^{\mathrm{t}}\dot{v}_m(X))^2 - \mathbb{E}_0\,((Db)(X))^2| = |\int_0^1 \frac{(\beta_m(t)^{\mathrm{t}}\gamma)^2}{\beta_m(t)^{\mathrm{t}}\gamma_0} s_m^{(0)}(t)\,\mathrm{d}t - \int_0^1 \frac{b(t)}{\eta_0(t)} s^{(0)}(t)\,\mathrm{d}t| \to 0,$$

as $m \to \infty$. The above results combined with the continuity of the mapping $\eta \mapsto p_{\theta_0,\eta}^{1/2}$ and Scheffé's lemma entails that (3.4) and (3.5) hold. Recall that the efficient score under $\mathcal{P}_m$ is

$$\tilde{\ell}_m(x) = \sum_{j=1}^{k_m} I_{V_{m,j}}(t)\Big(w - \frac{\int_{V_{m,j}} s_m^{(1)}(u)\,\mathrm{d}u}{\int_{V_{m,j}} s_m^{(0)}(u)\,\mathrm{d}u}\Big)\{\delta - e^{\theta w}H_m(t)\}.$$

Since $H_m(t) = \int_0^1 \beta_m(s)^{\mathrm{t}}\gamma_0\,\mathrm{d}t \to \int_0^1 \eta_0(s)\,\mathrm{d}s = H_0(t)$ by the $L^2(\mathrm{d}t)$ convergence, it suffices to find the limit of the sum. But by (C.5) and then the fact that $t_{j-1} = \lfloor t/\Delta_m \rfloor \Delta_m = t - \{t/\Delta_m\}\Delta_m \to t$, so that $s^{(k)}(t_{j-1}) \to s^{(k)}(t)$ by continuity,

$$\sum_{i=1}^{n} I_{V_{m,j}}(t)\frac{\int_{V_{m,j}} s_m^{(1)}(u)\,\mathrm{d}u}{\int_{V_{m,j}} s_m^{(0)}(u)\,\mathrm{d}u} = \sum_{i=1}^{n} I_{V_{m,j}}(t)\frac{s^{(1)}(t_{j-1})}{s^{(0)}(t_{j-1})} + o(1) = \frac{s^{(1)}(t)}{s^{(0)}(t)} + o(1).$$

This shows that

$$\tilde{\ell}_m(x) \to (w - \frac{s^{(1)}(t)}{s^{(0)}(t)})\{\delta - e^{\theta w} H(t)\} =: u(x), \quad \text{as } m \to \infty,$$

for μ-almost all x. The efficient information under $\mathcal{P}_m$ is

$$J_m = \sum_{j=1}^{k_m} \int_{V_{m,j}} v_{m,j} s_m^{(0)}(t) \beta_m(t)^{\mathrm{t}} \gamma_0 \, \mathrm{d}t.$$

By (C.5) we have that $\max_{j \le k_m} |v_{m,j} - v(t_{j-1})| = o(1)$. Therefore, using (cx1) and (cx4) and that $\beta_m(t)^{\mathrm{t}} \gamma_0 \to \eta_0$ in $L^2(\mathrm{d}t)$

$$\begin{aligned} J_m &= \sum_{j=1}^{k_m} v(t_{j-1}) \int_{V_{m,j}} s_m^{(0)}(t) \beta_m(t)^{\mathrm{t}} \gamma_0 \, \mathrm{d}t + o(1) \\ &= \sum_{j=1}^{k_m} v(t_{j-1}) s^{(0)}(t_{j-1}) \eta_0(t_{j-1}) \Delta_m + O(\Delta_m) + o(1) \\ &= \int_0^1 v(t) s^{(0)}(t) \eta_0(t) \, \mathrm{d}t + o(1) = \mathbb{E}_0 \, u(X)^2 + o(1), \end{aligned}$$

as $m \to \infty$. From Corollary 3.4 we can now conclude that Assumption 4 holds, that $u = \tilde{\ell}$ almost surely, and that $J = \int_0^1 v(t) s^{(0}(t) \eta_0(t) \, \mathrm{d}t$, where $\tilde{\ell}$ and J are the efficient score and the efficient information under P_0. We now check Assumption 3. Recall from Section 5.2 that

$$\log \frac{\mathrm{d}P_m^n}{\mathrm{d}P_0^n} = \sqrt{n} \int_0^1 g_{m,n}(t) \, \mathrm{d}\bar{M}_n(t) - \tfrac{1}{2} \int_0^1 g_{m,n}(t)^2 \, \mathrm{d}\bar{N}_n(t) + r_{m,n}, \tag{C.6}$$

where $g_{m,n}(t) = h_{m,n}(t)/\eta_0(t)$ with $h_{m,n}(t) = \sqrt{n}(\beta_m(t)^{\mathrm{t}} \gamma_0 - \eta_0(t))$, the remainder is $r_{m,n} = \int_0^1 g_{m,n}^2(t) R(g_{m,n}(t)/\sqrt{n}) \, \mathrm{d}\bar{N}_n$, and R is the function defined via the Taylor expansion $\log(1 + a) = a - \frac{1}{2}a^2 + a^2 R(a)$, and $R(a) \to 0$ as $a \to 0$. In particular $R(a) = \frac{1}{2}\big(1 - 1/(1 + \tilde{a})^2\big)$, where $\tilde{a}$ is some point between 0 and a. Notice that $g_{m,n}(t)/\sqrt{n} = \beta_m(t)^{\mathrm{t}}(t) \gamma_0/\eta_0(t) - 1$, and that by condition (cx4), the ratio of hazard rates is bounded below, $\beta_m(t)^{\mathrm{t}}(t) \gamma_0/\eta_0(t) \ge a_{\min} > 0$ say. But this means that $g_{m_n,n}(t)/\sqrt{n} \ge a_{\min} - 1$, and and it follows that $|R(g_{m_n,n}(t)/\sqrt{n})| \le 1 + |1/a_{\min}|$ for all n and t. With the basis functions in (5.5), we know from the introduction that $h_{m_n,n} \to 0$ in $L^2(\mathrm{d}t)$ provided (m_n) is such that $\sqrt{n}\Delta_{m_n} \to 0$. But then

$$\mathbb{E}_0 \left(\sqrt{n} \int_0^1 g_{m_n,n}(t) \, \mathrm{d}\bar{M}_n(t) \right)^2 \le \int_0^1 g_{m_n,n}(t)^2 s_{m_n}^{(0)} \eta_0(t) \, \mathrm{d}t = o(1),$$

by the $L^2(\mathrm{d}t)$ convergence combined with (cx1) and (cx4). The random variable $\int_0^1 g_{m,n}(t)^2 \, \mathrm{d}\bar{N}_n(t)$ is nonnegative and has expectation $\mathbb{E}_0 \int_0^1 g_{m,n}(t)^2 \, \mathrm{d}\bar{N}_n(t) = \int_0^1 g_{m_n,n}(t)^2 s_{m_n}^{(0)} \eta_0(t) \, \mathrm{d}t = o(1)$. The remainder term is bounded by

$$|r_{m_n,n}| \le \sup_{t \in [0,1]} |R(g_{m_n,n}(t)/\sqrt{n})| \int_0^1 g_{m_n,n}^2(t) \, \mathrm{d}\bar{N}_n = o_{P_0^n}(1),$$

as $n \to \infty$, by the findings above and Markov's inequality. This shows that $\log(\mathrm{d}P_{m_n}^n/\mathrm{d}P_0^n) = o_{P_0^n}(1)$, and Assumption 3 holds. Finally, we prove that (4.1) holds. Recall that

$$A_{m,n}(h) = \frac{h}{\sqrt{n}} \sum_{i=1}^{n} u_{m,n,i} - \tfrac{1}{2} h^2 \sum_{j=1}^{k_m} B_{m,n,j} \{\bar{N}_n(t_j) - \bar{N}_n(t_{j-1})\}, \tag{C.7}$$

where

$$u_{m,n,i} = \sum_{j=1}^{k_m} \int_{V_{m,j}} \Big(W_i - \frac{\int_{V_{m,j}} S_n^{(1)}(s,\theta_0)\,\mathrm{d}s}{\int_{V_{m,j}} S_n^{(0)}(s,\theta_0)\,\mathrm{d}s}\Big)\,\mathrm{d}M_i^{(m)}(t).$$

and

$$B_{m,n,j} = \Big\{ \frac{\int_{V_{m,j}} S_n^{(2)}(s,\xi_n)\,\mathrm{d}s}{\int_{V_{m,j}} S_n^{(0)}(s,\xi_n)\,\mathrm{d}s} - \Big(\frac{\int_{V_{m,j}} S_n^{(1)}(s,\xi_n)\,\mathrm{d}s}{\int_{V_{m,j}} S_n^{(0)}(s,\xi_n)\,\mathrm{d}s} \Big)^2 \Big\}.$$

First,

$$\begin{aligned} D_{m,n} &:= \frac{1}{\sqrt{n}} \sum_{i=1}^n u_{m,n,i} - \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_m(X_i) \\ &= -\sum_{j=1}^{k_m} \Big(\frac{\int_{V_{m,j}} S_n^{(1)}(s,\theta_0)\,\mathrm{d}s}{\int_{V_{m,j}} S_n^{(0)}(s,\theta_0)\,\mathrm{d}s} - \frac{\int_{V_{m,j}} s_m^{(1)}(s)\,\mathrm{d}s}{\int_{V_{m,j}} s_m^{(0)}(s)\,\mathrm{d}s} \Big) \sqrt{n}\{\bar{M}_n^{(m)}(t_j) - \bar{M}_n^{(m)}(t_{j-1})\}, \end{aligned}$$

where $\bar{M}_n^{(m)}(t) = n^{-1}\sum_{i=1}^n M_i^{(m)}(t)$. We have that

$$\max_{j\le k_m} \Big| \int_{V_{m,j}} S_n^{(k)}(s,\theta_0)\,\mathrm{d}s - \int_{V_{m,j}} s_m^{(k)}(s)\,\mathrm{d}s \Big| \le \sup_{s\in[0,1]} |S_n^{(k)}(s,\theta_0) - s_m^{(k)}(s)|\Delta_m,$$

where, by conditions (cx3) and (cx4) and the triangle inequality $|S_n^{(k)}(s,\theta_0) - s_m^{(k)}(s)| \le |S_n^{(k)}(s,\theta_0) - s^{(k)}(s)| + |s_m^{(k)}(s) - s^{(k)}(s)|$, the left hand side is $o_{P_0^n}(1)\Delta_m$ as $m,n\to\infty$. This entails that

$$\max_{j\le k_m} \Big| \frac{\int_{V_{m,j}} S_n^{(1)}(s,\theta_0)\,\mathrm{d}s}{\int_{V_{m,j}} S_n^{(0)}(s,\theta_0)\,\mathrm{d}s} - \frac{\int_{V_{m,j}} s_m^{(1)}(s)\,\mathrm{d}s}{\int_{V_{m,j}} s_m^{(0)}(s)\,\mathrm{d}s} \Big| = o_{P_0^n}(1)\Delta_m,$$

as $m,n\to\infty$. Then $|D_{m,n}| \le o_{P_0^n}(1)\Delta_m \sum_{j=1}^{k_m} |\sqrt{n}(\bar{M}_n^{(m)}(t_j) - \bar{M}_n^{(m)}(t_{j-1}))|$, where by Hölder's inequality followed by Jensen's inequality

$$\begin{aligned} \mathbb{E}_m \sum_{j=1}^{k_m} |\bar{M}_n^{(m)}(t_j) - \bar{M}_n^{(m)}(t_{j-1})| &\le \sum_{j=1}^{k_m} \Big(\mathbb{E}_m \Big(\int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t)\Big)^2\Big)^{1/2} \\ &= \sum_{j=1}^{k_m} \Big(\int_{V_{m,j}} s_m^{(0)}(t)\beta_m(t)^{\mathrm{t}}\gamma_0\,\mathrm{d}t\Big)^{1/2} \\ &= k_m^{1/2}\Big(\int_0^1 s_m^{(0)}(t)\beta_m(t)^{\mathrm{t}}\gamma_0\,\mathrm{d}t\Big)^{1/2} = k_m^{1/2} O(1). \end{aligned}$$

This shows that $\sum_{j=1}^{k_m} |\sqrt{n}(\bar{M}_n^{(m)}(t_j) - \bar{M}_n^{(m)}(t_{j-1}))| = O_{P_m^n}(\Delta_m^{-1/2})$ as $n\to\infty$, since k_m is proportional to Δ_m^{-1}. If we choose (m_n) so that $\sqrt{n}\Delta_{m_n} \to 0$ then P_0^n and $P_{m_n}^n$ are mutually contiguous, as seen above, so we can swap $o_{P_{m_n}^n}(1)$ and $o_{P_0^n}(1)$ terms, hence

$$|D_{m_n,n}| \le o_{P_{m_n}^n}(1)\Delta_{m_n} O_{P_{m_n}^n}(\Delta_{m_n}^{-1/2}) = o_{P_{m_n}^n}(\Delta_{m_n}^{1/2}),$$

as $n\to\infty$. This shows that the profile likelihood score is only $o_{P_{m_n}^n}(\Delta_{m_n}^{1/2})$ away from the efficient score under $\mathcal{P}_m$. Next, we show that the second term in (C.7) converges to J_{m_n}. By the same arguments as above $\max_{j\le k_m} |B_{m,n,j} - v_{m,j}| = o_{P_m^n}(1)\Delta_m$, as $m,n\to\infty$. We decompose the counting process under P_m and write

$$\sum_{j=1}^{k_m} B_{m,n,j} \int_{V_{m,j}} \mathrm{d}\bar{N}_n(t) = \sum_{j=1}^{k_m} B_{m,n,j} \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t) + \sum_{j=1}^{k_m} B_{m,n,j} \int_{V_{m,j}} S_n^{(0)}(t,\theta_0)\beta_m(t)^{\mathrm{t}}\gamma_0\,\mathrm{d}t.$$

For the first term on the right, write

$$\sum_{j=1}^{k_m} B_{m,n,j} \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t) = \sum_{j=1}^{k_m} v_{m,j} \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t) + \sum_{j=1}^{k_m} \left(B_{m,n,j} - v_{m,j}\right) \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t),$$

where

$$\mathbb{E}_m \left(\sum_{j=1}^{k_m} v_{m,j} \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t) \right)^2 = \frac{1}{n} \sum_{j=1}^{k_m} \int_{V_{m,j}} v_{m,j}^2 s_m^{(0)}(t) \beta_m(t)^{\mathrm{t}} \gamma_0 \,\mathrm{d}t = O(1/n),$$

which shows that $\sum_{j=1}^{k_m} v_{m,j} \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t) = o_{P_m^n}(1)$ as $m, n \to \infty$. For the second term on the right,

$$\begin{aligned} \Big|\sum_{j=1}^{k_m} (B_{m,n,j} - v_{m,j}) \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t)\Big| &\leq \max_{j \leq k_m} |B_{m,n,j} - v_{m,j}| \sum_{j=1}^{k_m} \Big| \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t)\Big| \\ &= o_{P_m^n}(1) \Delta_m \sum_{j=1}^{k_m} \Big|, \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t)\Big| \end{aligned}$$

and as we saw above $\sum_{j=1}^{k_m} \big| \int_{V_{m,j}} \mathrm{d}\bar{M}_n^{(m)}(t)\big| = O_{P_m^n}(\Delta_m^{-1/2})$. The second term is therefore $o_{P_m^n}(1)\Delta_m O_{P_m^n}(\Delta_m^{-1/2}) = o_{P_m^n}(\Delta_m^{1/2})$ as $m, n \to \infty$. This shows that

$$\sum_{j=1}^{k_m} B_{m,n,j} \int_{V_{m,j}} \mathrm{d}\bar{N}_n(t) - J_m = \sum_{j=1}^{k_m} B_{m,n,j} \int_{V_{m,j}} S_n^{(0)}(t, \theta_0) \beta_m(t)^{\mathrm{t}} \gamma_0 \,\mathrm{d}t - J_m + o_{P_m^n}(1).$$

where we recall that $J_m = \sum_{j=1}^{k_m} v_{m,j} \int_{V_{m,j}} s_m^{(0)}(t) \beta_m(t)^{\mathrm{t}} \gamma_0 \,\mathrm{d}t$. Considering one summand at the time, the difference on the right is bounded by

$$\begin{aligned} &\big|B_{m,n,j} \int_{V_{m,n,j}} S_n^{(0)}(t, \theta_0) \beta_m(t)^{\mathrm{t}} \gamma_0 \,\mathrm{d}t - v_{m,j} \int_{V_{m,j}} s_m^{(0)}(t) \beta_m(t)^{\mathrm{t}} \gamma_0 \,\mathrm{d}t\big| \\ &\leq \big|B_{m,n,j} - v_{m,j}\big| \int_{V_{m,n,j}} S_n^{(0)}(t, \theta_0) \beta_m(t)^{\mathrm{t}} \gamma_0 \,\mathrm{d}t + v_{m,j} \int_{V_{m,j}} \big|S_n^{(0)}(t, \theta_0) - s_m^{(0)}(t)\big| \beta_m(t)^{\mathrm{t}} \gamma_0 \,\mathrm{d}t \\ &= o_{P_m^n}(\Delta_m) O_{P_m^n}(1) + o_{P_m^n}(\Delta_m) O(1) = o_{P_m^n}(\Delta_m), \end{aligned}$$

as $m, n \to \infty$, and these bounds are uniform in j as $m \to \infty$, i.e., they hold when taking $\max_{j \leq k_m}$ outside the absolute value. Therefore, $\sum_{j=1}^{k_m} B_{m,n,j} \int_{V_{m,j}} S_n^{(0)}(t, \theta_0) \beta_m(t)^{\mathrm{t}} \gamma_0 \,\mathrm{d}t - J_m = o_{P_m^n}(1)$. We conclude that $\sum_{j=1}^{k_m} B_{m,n,j} \int_{V_{m,j}} \mathrm{d}\bar{N}_n(t) - J_m = o_{P_m^n}(1)$. This means that we have now shown that the process $A_{m,n}(h)$ in (C.7) is $A_{m,n}(h) = n^{-1/2} \sum_{i=1}^n \tilde{\ell}_m(X_i) - \frac{1}{2} h^2 J_{m_n} + o_{P_m^n}(1)$, for each h, as $m, n \to \infty$. In other words, $A_{m_n,n}(h)$ satisfies (4.1) for any sequence (m_n) tending to infinity with n. From Proposition 4.1 we can now conclude that, provided (m_n) is such that $\sqrt{n}\Delta_{m_n} \to 0$, the maximiser of $\mathrm{pl}_{m_n,n}(\theta)$ is best regular under P_0.

References

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

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: pointwise and uniform results. *Journal of Econometrics*, 186:345–366.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer, New York.

Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley & Sons, New York.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics*, 32(2):189–218.

Dudley, R. (2002). *Real Analysis and Probability.* Cambridge University Press, Cambridge.

Feinberg, E. A., Kasyanov, P. O., and Zgurovsky, M. Z. (2016). Uniform Fatou's lemma. *Journal of Mathematical Analysis and Applications*, 444:550–567.

Gill, R. D. (1984). Understanding Cox's regression model: A martingale approach. *Journal of the American Statistical Association*, 79:441–447.

Gut, A. (1992). The weak law of large numbers for arrays. *Statistics & Probability Letters*, 14:49–52.

Hjort, N. L. and Pollard, D. B. (1993). Asymptotics for minimisers of convex processes. Technical report, Department of Mathematics, University of Oslo.

Hoesch, L., Lee, A., and Mesters, G. (2024). Locally robust inference for non-gaussian SVAR models. *Quantitative Economics*, 15:523–570.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer, New York.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer, New York.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95:449–465.

Mykland, P. A. and Zhang, L. (2009). Inference for continuous semimartingales observed at high frequency. *Econometrica*, 77:1403–1445.

Powell, M. J. D. (1981). *Approximation Theory and Methods.* Cambridge University Press.

Riesz, F. (1928). Sur la convergence en moyenne. *Acta Sci. Math. (Szeged)*, 4:58–64.

Royden, H. L. and Fitzpatrick, P. (2010). *Real Analysis. Fourth Edition.* Pearson Education Asia Limited and China Machine Press.

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics*, 25:2555–2591.

Strasser, H. (1985). *Mathematical Theory of Statistics.* de Gruyter, Berlin, New York.

Szegő, G. (1975). *Orthogonal Polynomials.* American Math. Soc: Colloquium publ. American Mathematical Society.

van der Vaart, A. W. (1988). *Statistical Estimation in Large Parameter Spaces.* Number 44 in CWI Tracts. Centrum voor Wiskunde en Informatica, Amsterdam.

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

Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press, Cambridge.