JSPI 167

ELSEVIER

Journal of Statistical Planning and Inference 000 (2000) 000–000

journal of
statistical planning
and inference

www.elsevier.com/locate/jspi

Application of convolution theorems in semiparametric models with non-i.i.d. data

Brad McNeney[a,b,c,*,1], Jon A. Wellner[c,2]

[a] *Statistics, North Carolina State University, Campus Box 8203, Raleigh, NC 27695-8203, USA*
[b] *National Institute of Statistical Sciences, 19 TW Alexander Dr., P.O. Box 14006, Research Triangle Park, NC 27709-4006, USA*
[c] *University of Washington, Statistics, Box 354322, Seattle, WA 98195-4322, USA*

Abstract

A useful approach to asymptotic efficiency for estimators in semiparametric models is the study of lower bounds on asymptotic variances via convolution theorems. Such theorems are often applicable in models in which the classical assumptions of independence and identical distributions fail to hold, but to date, much of the research has focused on semiparametric models with independent and identically distributed (i.i.d.) data because tools are available in the i.i.d. setting for verifying pre-conditions of the convolution theorems. We develop tools for non-i.i.d. data that are similar in spirit to those for i.i.d. data and also analogous to the approaches used in parametric models with dependent data. This involves extending the notion of the tangent vector figuring so prominently in the i.i.d. theory and providing conditions for smoothness, or differentiability, of the parameter of interest as a function of the underlying probability measures. As a corollary to the differentiability result we obtain sufficient conditions for equivalence, in terms of asymptotic variance bounds, of two models. Regularity and asymptotic linearity of estimators are also discussed. © 2000 Elsevier Science B.V. All rights reserved.

MSC: primary 60F05; 60F17; secondary 60J65; 60J70

Keywords: Convolution theorem; Asymptotic efficiency; Information bound; Semiparametric model; Local asymptotic normality; Differentiability; Regular estimator

1. Introduction

Given a number of choices for making inference from observed data or, more particularly, estimating parameters, an important goal is to identify the procedure that

[*] Correspondence address: North Carolina State University, National Institute of Statistical Sciences, Box 354322, Seattle, WA 98195-4322, USA.

E-mail addresses: mcneney@niss.org (B. McNeney), jaw@stat.washington.edu (J.A. Wellner).

[1] Supported in part by NIH grants DK-42654 and R01 CA-5196 and R01 CA-519622.

[2] Supported in part by National Science Foundation grant DMS-9532039 and NIAID grant 2R01 AI291968-04.

0378-3758/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(00)00193-2

makes the best use of available data. Unfortunately, even conceptually simple experiments often lead to statistical models in which it is extremely difficult to describe performance of estimators in finite samples. The simplification in the structure of the model obtained when the sample size tends to infinity is often the only way to obtain a tractable notion of optimality. The hope is that estimators determined to be optimal in an asymptotic sense may be expected to perform well in the finite samples obtained in practice.

A relatively well-studied concept of efficiency is based on what are commonly referred to as convolution theorems. The two key hypotheses of such a theorem are local asymptotic normality (LAN) and differentiability of the parameter of interest. The latter requires that this parameter, which could take values in an infinite-dimensional space, represents a smooth function of the probability measures in the underlying statistical model. Under these hypotheses, convolution theorems assert a minimum asymptotic variance among estimators that satisfy certain regularity conditions. Application to many interesting i.i.d. data models and the resulting characterization of efficient estimators has met with considerable success. See for example the monograph by Bickel et al. (1993) (hereafter referred to as BKRW) for applications to non- and semi-parametric models. The appeal of the i.i.d. theory is the availability of convenient sufficient conditions for the hypotheses. For example, it is known that certain "differentiability in quadratic mean" conditions imply LAN. These conditions introduce the concepts of *tangent vectors* and the *tangent space*. The geometry of the latter has proved to be particularly useful in characterizing efficient estimators. To establish differentiability when the parameter is an implicit function of the probability measures resulting from the parametrization of the model, rather than an explicit function, a result due to Van der Vaart (1991) gives necessary and sufficient conditions for differentiability. In this paper we explore analogous results that do not assume i.i.d. data. The results are illustrated on a series of examples.

A more detailed outline of the paper is as follows. In Section 2 we give more precise definitions of LAN, differentiability and regularity of an estimator, along with a statement of a convolution theorem due to Van der Vaart and Wellner (1996). Section 3 discusses LAN in more detail and describes a set of sufficient conditions that do not assume i.i.d. data. Particular emphasis is placed on stating these results in a way that resembles the i.i.d. theory as much as possible. Based on the definition of tangent vectors developed in Section 3, conditions for the smoothness of the parameter to be estimated are developed in Section 4 that parallel results in the i.i.d. theory. In Section 5 we discuss regularity of estimators in more detail along with a characterization of efficient estimators. We conclude with a discussion of the results and open problems.

2. Basic definitions and convolution theorem

To state the theorem, we first need a precise definition of local asymptotic normality (LAN), the differentiability hypothesis, and of the notion of regular estimators. The first definition introduces the tangent space $\mathscr{H}$; the second and third are relative to $\mathscr{H}$. All three are taken from Van der Vaart and Wellner (1996, p. 413).

Before giving the formal definition of LAN we describe the notation used in the definition. Let Θ denote a parameter space and $\mathscr{P}=\{P_\theta\colon \theta\in\Theta\}$ be a family of probability measures defined on a measurable space $(\Omega,\mathscr{F})$. Suppose we observe m_n random elements $(X_{n1},\dots,X_{nm_n})\sim P_{n,\theta}$ where $P_{n,\theta}=P_\theta|_{\mathscr{F}_n}$, the restriction of P_θ to the sigma algebra $\mathscr{F}_n=\sigma(X_{n1},\dots,X_{nm_n})$. The log likelihood ratio for two points θ_1 and θ_2 in Θ with n observations will be denoted by

$$\Lambda_n(\theta_1,\theta_2)=\log\frac{\mathrm{d}P_{n,\theta_1}}{\mathrm{d}P_{n,\theta_2}},$$

where $\mathrm{d}P_{n,\theta_1}/\mathrm{d}P_{n,\theta_2}$ denotes the Radon–Nikodym derivative of the absolutely continuous part of P_{n,θ_1} with respect to P_{n,θ_2}.

Definition 2.1 (*LAN*). Let $\mathscr{H}$ be a linear space with inner-product $\langle\cdot,\cdot\rangle$ and norm $||\cdot||$. We say the model is LAN at $\theta_0\in\Theta$ indexed by the *tangent space* $\mathscr{H}$ if for each $h\in\mathscr{H}$ there exists a sequence $\{P_{n,\theta_n(h)}\}$ of probability measures defined on $(\Omega,\mathscr{F})$ with

$$\Lambda_n(\theta_n(h),\theta_0)=\Delta_{n,h}-\tfrac{1}{2}||h||^2+\mathrm{o}_{\mathrm{P}_{n,0}}(1). \tag{2.1}$$

Here $\Delta_{n,h}:\Omega\to\mathbb{R}$ are measurable maps with

$$\mathscr{L}(\Delta_{n,h_1},\dots,\Delta_{n,h_d}|P_{n,0})\to \mathrm{N}_d(0,\langle h_i,h_j\rangle) \tag{2.2}$$

for every finite subset $h_1,\dots,h_d\in\mathscr{H}$.

Consider also the weaker condition

$$\mathscr{L}(\Delta_{n,h})\to \mathrm{N}_1(0,||h||^2)\quad \text{for every } h\in\mathscr{H}. \tag{2.3}$$

Note that if the maps $\Delta_{n,h}$ are linear in h then given any collection $(h_1,\dots,h_d)$ and any vector $a\in\mathbb{R}^d$ we have

$$\Delta_{n,\sum_{i=1}^d a_i h_i}=\sum_{i=1}^d a_i\Delta_{n,h_i}=(\Delta_{n,h_1},\dots,\Delta_{n,h_d})a^{\mathrm{T}}. \tag{2.4}$$

But, under (2.3),

$$\mathscr{L}(\Delta_{n,\sum_{i=1}^d a_i h_i})\to \mathrm{N}_1\left(0,\left|\left|\sum_{i=1}^d a_i h_i\right|\right|^2\right)\quad \text{where}\quad \left|\left|\sum_{i=1}^d a_i h_i\right|\right|^2=a^{\mathrm{T}}[\langle h_i,h_j\rangle]_d\, a.$$

We may thus conclude, after an application of the Cramér–Wold device, that (2.2) holds under the linearity assumption (2.4). This would continue to be the case if the $\Delta_{n,h}$ are only approximately linear; i.e. if

$$\Delta_{n,a_1h_1+a_2h_2}=a_1\Delta_{n,h_1}+a_2\Delta_{n,h_2}+\mathrm{o}_{\mathrm{P}_{n0}}(1).$$

With the tangent space defined, we may now give a statement of the differentiability condition and of regularity of an estimator.

Definition 2.2 (*Differentiability of a parameter*)**.** Let B be a Banach space and $v_n(P_{n,\theta})$ be B-valued "parameters". We say the sequence $\{v_n\}$ is *differentiable* if

$$R_n(v_n(P_{n,\theta_n(h)}) - v_n(P_{n,0})) \to \dot{v}(h) \quad \text{for every } h \in \mathscr{H} \tag{2.5}$$

for some sequence of linear maps $R_n : B \to B$ with $||R_n|| \to \infty$ and a continuous linear map $\dot{v} : \mathscr{H} \to B$.

Definition 2.3 (*Regular estimators*)**.** A sequence of maps $T_n : \mathscr{X}_n \to B$ is said to be *locally regular* for v_n if under $P_{n,\theta_n(h)}$,

$$R_n(T_n - v_n(P_{n,\theta_n(h)})) \Rightarrow \mathbb{Z} \quad \text{as } n \to \infty, \tag{2.6}$$

for every $h \in \mathscr{H}$, where $\mathbb{Z}$ is a Borel measurable tight random element in B which does not depend on $h \in \mathscr{H}$.

Theorem 2.4 (Convolution theorem)**.** *Suppose $\mathscr{P} = \{P_\theta\colon \theta \in \Theta\}$ is LAN at a point θ_0 indexed by a linear subspace $(\mathscr{H}, \langle\cdot,\cdot\rangle)$ of a Hilbert space. Further suppose $\{v_n\}$ is a differentiable sequence of parameters. Then if T_n is locally regular for v_n, there exist tight Borel measurable elements $\mathbb{Z}_0$ and $\mathbb{W}$ in B with*

(A) $P(\mathbb{Z}_0 \in \overline{\dot{v}(\mathscr{H})}) = 1$.
(B) $\mathscr{L}(\mathbb{Z}) = \mathscr{L}(\mathbb{Z}_0 + \mathbb{W})$.
(C) $\mathbb{Z}_0$ and $\mathbb{W}$ *are independent*.
(D) $\mathscr{L}(b^*\mathbb{Z}_0) = \mathrm{N}(0, ||\dot{v}^{\mathrm{T}}_{b^*}||^2)$ *for every* $b^* \in B^*$.

Here $\dot{v}^{\mathrm{T}}_{b^*}$ is the unique element of $\bar{\mathscr{H}}$ such that

$$(\dot{v}^{\mathrm{T}} b^*)h = \langle \dot{v}^{\mathrm{T}}_{b^*}, h\rangle \quad \textit{where } \dot{v}^{\mathrm{T}} : B^* \to \mathscr{H}^*$$

is the adjoint of $\dot{v}$.

Theorem 2.4 was established by Van der Vaart and Wellner (1991), and is Theorem 3.11.2, p. 414 of Van der Vaart and Wellner (1996).

Because a Hilbert space and its dual can always be identified, we often do not make a distinction between $\dot{v}^{\mathrm{T}} b^*$ and $\dot{v}^{\mathrm{T}}_{b^*}$. Note also that we do not require the adjoint to map to the dual of a closed, and hence complete, subspace. Although many interesting features of adjoints depend on completeness, we shall only use the most basic properties. For our applications we only need the dual to separate points of the space in question so that adjoints are well defined. This is certainly true for the (subspaces of) Banach spaces we will encounter in our applications.

3. Local asymptotic normality (LAN)

To interpret the LAN conditions it is helpful to specialize to the case where the parameter space is finite-dimensional. Then Definition 2.1 may be expressed as: The model is LAN at θ_0 if there are random vectors S_n and a positive-definite matrix K such that for all $t \in \mathbb{R}^k$

$$\Lambda_n(\theta_0 + \delta_n t, \theta_0) = t' S_n - \tfrac{1}{2} t' K t + R_n(\theta_0, t), \tag{3.1}$$

where $\mathscr{L}(S_n | P_{n,\theta_0}) \to \mathrm{N}(0, K)$ and $R_n(\theta_0, t) \to 0$ in P_{n,θ_0} probability.

It is a fact due to Le Cam (1960) that under the LAN conditions the sequences $P_{n,\theta_0+\delta_n t}$ and P_{n,θ_0} are mutually contiguous. However, the reasoning behind the terminology (locally asymptotically normal) is as follows. Consider a random vector X, distributed as $\mathrm{N}(Kt, K)$ under a measure Q_t, and distributed as $\mathrm{N}(0, K)$ under Q_0. Some algebra reveals that the log likelihood ratio of Q_t to Q_0 is indeed $t'X - (1/2)t'Kt$. Thus the likelihood ratios in (3.1) converge in distribution to the likelihood ratios of a Gaussian shift experiment where t indexes the shift. It is known (Le Cam, 1969) that this convergence in distribution of likelihood ratios is equivalent to a certain type of convergence of experiments; that is, the sequence $\mathscr{P}_n$ is approximated, in local (shrinking) neighborhoods of θ_0, by a Gaussian shift experiment. The vectors t in (3.1) can be thought of as directions in $\mathbb{R}^k$ from which θ is approached at rate δ_n. The path of approach represents a one-dimensional submodel. In problems where the parameter space is infinite-dimensional, such as in non- and semi-parametric models, we continue to look at one-dimensional submodels that satisfy a condition such as (3.1). Just as the t's index directions and the shift in the approximating experiment for parametric models, so do the h's in the more general context. These are the keys in the asymptotic expansion in a neighborhood of θ_0 and the geometry provided by the inner product is a natural extension of the form of the approximation for parametric models.

In certain examples LAN may be verified by direct calculation, as in the following.

Example 1. A particular case of the class of models studied by Pfanzagl (1993) can be described as follows. Consider sampling $X_1, X_2, \ldots$ independently from normal distributions $\mathrm{N}(\eta_1 + \theta, 1), \mathrm{N}(\eta_2 + \theta, 1), \ldots$ for an *unobserved* sequence $\boldsymbol{\eta} \equiv (\eta_1, \eta_2, \ldots)$ and common parameter $\theta \in \mathbb{R}$. The goal is to estimate θ. In the i.i.d. version, we envision $\boldsymbol{\eta}$ as a random sample according to some distribution G, say with mean 0 so that θ can be identified. Then the resulting observations are i.i.d. according to the measure defined by

$$P_{\theta,G}(X_i \leqslant x) = \int_{-\infty}^{\infty} \int_{-\infty}^{x} \phi(t - (\eta + \theta))\,\mathrm{d}t\,\mathrm{d}G(\eta),$$

where ϕ is the standard normal density. One possible non-i.i.d. case arises if $\boldsymbol{\eta}$ is thought of as a fixed, unknown, unobserved sequence which is centered at 0 in the sense that $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \eta_i = 0$. This is sometimes called a functional model. More generally, one could consider sampling $\boldsymbol{\eta}$ according to a measure Γ on $H^{\mathbb{N}}$ and then conditional on $\boldsymbol{\eta}$, sampling $X_1, X_2, \ldots$ from $P_{\theta,\eta_1}, P_{\theta,\eta_2}, \ldots$. The i.i.d. model corresponds to $\Gamma = G^{\mathbb{N}}$ while the functional model corresponds to $\Gamma = \delta_{\boldsymbol{\eta}}$, a point mass at $\boldsymbol{\eta}$. This more general setting allows the nuisance sequence to be specified, for instance, as a sample from a stationary process. However, for the remainder of this paper we consider two special cases where the nuisance sequence is chosen deterministically via a function on $[0,1]$. In both instances the resulting data are independent but not identically distributed.

For the first of these approaches, take a continuous function f_0 on the interval $[0,1]$. Since $[0,1]$ is compact, f_0 is actually uniformly continuous and bounded. Suppose further that f_0 is centered about 0 in that $\int f_0(s)\,\mathrm{d}s=0$ and denote the space of all such functions $C_b^0[0,1]$ equipped with the supremum norm $||f||_\infty=\sup_x|f(x)|$. Since $f_0\in C_b^0[0,1]$ it makes sense to specify an array of nuisance parameters by $\eta_{ni}=f_0(i/n)$, $i=1,\dots,n$.

In the second approach take the nuisance parameters to be generated by a function $\dot{f}_0\in L_2^0(\lambda)$ with λ Lebesgue measure on $[0,1]$. For given n take

$$\eta_{ni}=n\int_{(i-1)/n}^{i/n}\dot{f}_0\,\mathrm{d}\lambda,\quad i=1,\dots,n.$$

Then indeed

$$\frac{1}{n}\sum_{i=1}^{n}\eta_{ni}=\int_0^1\dot{f}_0\,\mathrm{d}\lambda=0.$$

To verify LAN, say for the model given by the first method of specifying the nuisance parameters, we must consider a sequence of points in the parameter space that tend toward (θ_0,f_0). This is achieved by first considering paths in the parameter space that pass through (θ_0,f_0), and then considering a sequence along the paths. In this case we take a path (θ_t,f_t) where $\theta_t\in\mathbb{R}$ with $\theta_t=\theta_0+ta$ and $f_t\in C_b^0[0,1]$ with $f_t(s)=f(s)+tg(s)$ for $g\in C_b^0[0,1]$. Then, for $t_n=n^{-1/2}$, the log likelihood ratio for a single observation X_{ni} in the nth row of the array is given by

$$-\tfrac{1}{2}(X_{ni}-(f_{t_n}(i/n)+\theta_{t_n}))^2+\frac{1}{2}(X_{ni}-(f_0(i/n)+\theta_0))^2.$$

It is then straightforward to verify that

$$\Lambda_n((\theta_{t_n},f_{t_n}),(\theta_0,f_0))=\Delta_{nh}-\sigma_n^2/2, \tag{3.2}$$

where h is given by $h(s)=g(s)+a$,

$$\sigma_n^2=\frac{1}{n}\sum_{i=1}^{n}(g(i/n)+a)^2$$

and

$$\Delta_{nh}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{[X_{ni}-(f_0(i/n)+\theta_0)](g(i/n)+a)\}\sim\mathrm{N}(0,\sigma_n^2).$$

In order to conclude LAN it is important that h be an element of an inner-product space. Here we choose $L_2(\lambda)$ where λ is Lebesgue measure on $[0,1]$. Thus we say $h=\dot{l}(a,g)=a+g$ for $\dot{l}:\mathbb{R}\times C_b^0[0,1]\to\mathscr{H}=L_2(\lambda)$. Since $\sigma_n^2=n^{-1}\sum_{i=1}^{n}(g(i/n)+a)^2$ converges to $\sigma^2\equiv\int_0^1 h(s)^2\mathrm{d}s=||h||_{L_2(\lambda)}^2$, $\mathscr{L}(\Delta_{nh}|P_{(\theta_0,f_0)})\to\mathrm{N}(0,\sigma^2)$. Conclude that

$$\Lambda((\theta_{t_n},f_{t_n}),(\theta_0,f_0))=\Delta_{nh}-\sigma^2/2+o_P(1).$$

Finally, note that Δ_{nh} is linear in h, while the operator $\dot{l}$ is linear so that its range, $\mathscr{H}$, is a linear space. With these final conditions satisfied, this model could be described as LAN indexed by $\mathscr{R}(\dot{l})$.

For the version of this problem where the nuisance sequence is generated by an $L_2^0(\lambda)$ function $\dot{f}_0$, the paths through the parameter space are constructed analogously with $\dot{f}_t$ where $\dot{f}_t(s) = \dot{f}_0(s) + t\dot{g}(s)$ for $\dot{g} \in L_2^0(\lambda)$. In general, write

$$\eta_{ni}(\dot{f}) = n\int_{(i-1)/n}^{i/n} \dot{f}\,\mathrm{d}\lambda.$$

The path θ_t in R is as before. Again take $t_n = n^{-1/2}$. The log likelihood ratio for a single observation X_{ni} in the nth row is now

$$-\tfrac{1}{2}(X_{ni} - (\eta_{ni}(\dot{f}_{t_n}) + \theta_{t_n}))^2 + \tfrac{1}{2}(X_{ni} - (\eta_{ni}(\dot{f}_0) + \theta_0))^2.$$

From this we get a log-likelihood ratio of

$$\Lambda_n((\theta_{t_n}, \dot{f}_{t_n}), (\theta_0, \dot{f}_0)) = \Delta_{nh} - \sigma_n^2/2, \tag{3.3}$$

where $h(s) = \dot{g}(s) + a$,

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(\eta_{ni}(\dot{g}) + a)^2$$

and

$$\Delta_{nh} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}([X_{ni} - (\eta_{ni}(\dot{f}_0) + \theta_0)](\eta_{ni}(\dot{g}) + a)) \sim N(0, \sigma_n^2).$$

In this case, h is already an element of an inner-product space namely $L_2(\lambda)$. The score operator here is defined by $h = \dot{l}(a, g) = a + \dot{g}$ for $\dot{l}\colon \mathbb{R} \times L_2^0(\lambda) \to \mathscr{H} = L_2(\lambda)$. The functions

$$\phi_n(x) = n\int_{(i-1)/n}^{i/n} (\dot{g} + a)\,\mathrm{d}\lambda\, 1_{((i-1)/n, i/n]}(x)$$

are approximants that converge in $L_2(\lambda)$ to $\dot{g}+a$ by standard results in L_2 approximation theory (eg. Royden, 1988, pp. 128–129). Thus

$$\sigma_n^2 = ||\phi_n||^2 \to ||\dot{g} + a||^2 \equiv \sigma^2$$

so that $\mathscr{L}(\Delta_{nh}|P_{(\theta_0, \dot{f}_0)}) \to \mathrm{N}(0, \sigma^2)$, and

$$\Lambda((\theta_{t_n}, \dot{f}_{t_n}), (\theta_0, \dot{f}_0)) = \Delta_{nh} - \sigma^2/2 + o_P(1)$$

for the newly defined Δ_{nh} and σ^2. The map Δ_{nh} is still linear in h, and the score operator $\dot{l}$ is linear so this version of the model could be described as LAN indexed by $\mathscr{R}(\dot{l})$. □

In the above example the real-valued parameter θ is described as the parameter of interest, while the functions f_0 or $\dot{f}_0$ that generate the sequence $\eta_{n1}, \eta_{n2}, \ldots$ are described as nuisance parameters. However, estimation of f_0 or $\dot{f}_0$ could also be stated as goals of inference. These parameters are not expressed as explicit functions of the probability measures in the underlying model. Rather, they are defined implicitly via the parametrization. Establishing differentiability of implicitly defined parameters is taken up in Section 4. There we shall see that f is differentiable, while $\dot{f}$ is not.

The above LAN calculations rely on the assumed normality of the observations. Most often the task of establishing LAN is more difficult. However, in the case the experiments consist of independent observations, LAN is implied by a certain "differentiability in quadratic mean" condition. We first discuss these sufficient conditions, before describing analogous sufficient conditions that do not assume i.i.d. data.

3.1. Sufficient conditions for LAN in i.i.d. models

A related concept to LAN, one which plays a prominent role in the i.i.d. theory, is the tangent vector. This was first described by Koshevnik and Levit (1976) for sequences of probability measures. In its simplest form, a function $h \in L_2(P_{\theta_0})$ is the tangent vector at P_{θ_0} of a path $\eta \mapsto P_{\theta_\eta}$ through $\mathscr{P}$ with $P_{\theta_\eta} \ll P_{\theta_0}$ for all η if

$$\int \left[\eta^{-1} \left(\sqrt{\frac{\mathrm{d}P_{\theta_\eta}}{\mathrm{d}P_{\theta_0}}} - 1 \right) - \frac{1}{2} h \right]^2 \mathrm{d}P_{\theta_0} \to 0 \quad \text{as } \eta \downarrow 0. \tag{3.4}$$

Since the above is $L_2(P_{\theta_0})$ convergence, we may think of $h/2$ as the $L_2(P_{\theta_0})$, or quadratic mean derivative of $\sqrt{\mathrm{d}P_{\theta_\eta}/\mathrm{d}P_{\theta_0}}$ at $\eta = 0$. The absolute continuity condition can, in fact, be relaxed if the P_{θ_η}-measure of the set where P_{θ_0} places no mass (the singular part of P_{θ_0}) disappears fast enough. See for example the two (DQM$_0$) conditions in Le Cam and Yang (1990, p. 101). These two conditions are equivalent to

$$\int \left[\eta^{-1} \left(\sqrt{\frac{\mathrm{d}P_{\theta_\eta}}{\mathrm{d}\mu_\eta}} - \sqrt{\frac{\mathrm{d}P_{\theta_0}}{\mathrm{d}\mu_\eta}} \right) - \frac{1}{2} h \sqrt{\frac{\mathrm{d}P_{\theta_0}}{\mathrm{d}\mu_\eta}} \right]^2 \mathrm{d}\mu_\eta \to 0 \quad \text{as } \eta \downarrow 0, \tag{3.5}$$

where each μ_η is an arbitrary σ-finite measure dominating both P_{θ_η} and P_{θ_0}. In fact, the integral expression on the left-hand side is the same for all choices of μ_η so that this measure is often suppressed in the notation. In this form, $\frac{1}{2} h \sqrt{\mathrm{d}P_{\theta_0}/\mathrm{d}\mu_\eta}$ is called the Hellinger derivative since the Hellinger distance between two measures P_{θ_η} and P_{θ_0} is the square root of $\frac{1}{2} \int [\sqrt{\mathrm{d}P_{\theta_\eta}/\mathrm{d}\mu_\eta} - \sqrt{\mathrm{d}P_{\theta_0}/\mathrm{d}\mu_\eta}]^2 \, \mathrm{d}\mu_\eta$ (cf. Begun et al. (1983) for this terminology).

Under the above differentiability in quadratic mean condition, it can be shown that for $\eta_n = n^{-1/2} + \mathrm{o}(n^{-1/2})$,

$$\Lambda_n(\theta_{\eta_n}, \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(X_i) - \frac{1}{2} ||h||^2 + r_n, \tag{3.6}$$

where $r_n \to 0$ in $P_{\theta_0}^n$-probability and $|| \cdot ||$ is the $L_2(P_{\theta_0})$ norm. In the case of regular finite-dimensional parametric models, the tangents are given by $\dot{l}(\theta_0)^{\mathrm{T}} t$ for $t \in \mathbb{R}^k$ where $\dot{l}(\theta_0)$ is the score vector in the quadratic mean sense; that is, for points $\theta_0 + t/\sqrt{n}$ along a path through Θ,

$$\int \left[\sqrt{n} \left(\sqrt{\mathrm{d}P_{\theta_{\theta_0 + t/\sqrt{n}}}} - \sqrt{\mathrm{d}P_{\theta_{\theta_0}}} \right) - \frac{1}{2} \dot{l}(\theta_0)^{\mathrm{T}} t \sqrt{\mathrm{d}P_{\theta_{\theta_0}}} \right]^2 \to 0 \quad \text{as } n \to \infty.$$

Thus, the tangent space for a regular parametric model is given by the span of the score vector $\dot{l}(\theta_0)$ (which usually coincides with the score vector in the usual sense) as a subspace of $L_2(P_{\theta_0})$. This is actually a subspace of $L_2^0(P_{\theta_0})$, the set of $L_2(P_{\theta_0})$ functions with mean 0, since scores have mean 0 (BKRW, p. 15). Note also the role of the score vector as an operator from R^k to $\mathscr{H}$ that maps the derivative of a given path (indicated by t) to the corresponding tangent $(\dot{l}(\theta_0)^{\mathrm{T}}t)$. The analog for infinite-dimensional parameter spaces, usually assumed to be a Hilbert space, are score operators. As in the finite-dimensional case these can be used to determine the tangent space. Score operators in the current context will be discussed at the end of this section.

3.2. *Sufficient conditions for LAN without assuming i.i.d. data*

In general models we do not have a common density. However, we may still write the likelihood ratio as a product of "conditional densities". Such an approach was taken by Jeganathan (1982), in the context of parametric models, for verifying a more general asymptotic expansion of the likelihood ratios (local asymptotic mixed normality or LAMN). In the present context, we follow the approach outlined in Greenwood and Shiryayev (1985). As before Θ denotes the (possibly infinite dimensional) parameter space and $\mathscr{P} = \{P_\theta\colon \theta \in \Theta\}$ is a family of probability measures defined on a measurable space $(\Omega, \mathscr{F})$. For m_n observed random elements $(X_{n1}, \ldots, X_{nm_n})$ we have a non-decreasing family of sub-σ-algebras $\{\mathscr{F}_{nj}\colon j = 0, \ldots, m_n\}$ where $\mathscr{F}_{n0} = \{\emptyset, \Omega\}$, $\mathscr{F}_{nj} = \sigma(X_{n1}, \ldots, X_{nj})$ for $j \leqslant m_n$, and $\mathscr{F}_n = \mathscr{F}_{nm_n}$.

For given measures P and $\tilde{P}$ the above systems of sub-σ-algebras allow us to define

$$P_n = P|_{\mathscr{F}_n}, \quad \tilde{P}_n = \tilde{P}|_{\mathscr{F}_n}, \quad P_{nk} = P|_{\mathscr{F}_{nk}} = P_n|_{\mathscr{F}_{nk}}, \quad \tilde{P}_{nk} = \tilde{P}|_{\mathscr{F}_{nk}} = \tilde{P}_n|_{\mathscr{F}_{nk}}$$

and with $\mu_{nk} = (\tilde{P}_{nk} + P_{nk})/2$ and $\mu_n = (\tilde{P}_n + P_n)/2$,

$$\lambda_{nk} = \frac{\mathrm{d}P_{nk}}{\mathrm{d}\mu_{nk}}, \quad \tilde{\lambda}_{nk} = \frac{\mathrm{d}\tilde{P}_{nk}}{\mathrm{d}\mu_{nk}} \quad \text{and} \quad z_{nk} = \frac{\tilde{\lambda}_{nk}}{\lambda_{nk}}$$

with analogous definitions for λ_n, $\tilde{\lambda}_n$ and z_n. Since $\mathscr{F}_n = \mathscr{F}_{nm_n}$, $\lambda_n = \lambda_{nm_n}$, $\tilde{\lambda}_n = \tilde{\lambda}_{nm_n}$ and $z_n = z_{nm_n}$. We will also make use of $\alpha_{nk} = z_{nk}/z_{n(k-1)}$ with the conventions $a/0 = \infty$ if $a > 0$ and $0/0 = 0$. With $\mathscr{F}_{n0} = \{\emptyset, \Omega\}$, this implies $z_{n0} = 1$, since any probability measure restricted to this trivial σ-algebra is the same, so that $z_{nk} = \prod_{i=1}^k \alpha_{ni}$. In addition, define $\beta_{nk} = \lambda_{nk}/\lambda_{n(k-1)}$ and $\tilde{\beta}_{nk} = \tilde{\lambda}_{nk}/\tilde{\lambda}_{n(k-1)}$. The β_{nk} may be interpreted as conditional densities under P_n of X_{nk} given $X_{n1}, \ldots, X_{n(k-1)}$. Since $\alpha_{nk} = \tilde{\beta}_{nk}/\beta_{nk}$ we see that it is like a ratio of conditional likelihoods. Of course, if the observations are independent, then conditioning has no effect and the α_{nk} are the more familiar likelihood ratios corresponding to the observation X_{nk}.

When the measures $\tilde{P}$ and P are given by P_{θ_1} and P_{θ_2} respectively write $z_{nk}(\theta_1, \theta_2)$, $z_n(\theta_1, \theta_2)$, $\alpha_{nk}(\theta_1, \theta_2)$ and $\alpha_n(\theta_1, \theta_2)$. The log likelihood ratio is then given by

$$\Lambda_n(\theta_1, \theta_2) = \log z_n(\theta_1, \theta_2) = \sum_{k=1}^{m_n} \log \alpha_{nk}(\theta_1, \theta_2).$$

We are now ready to state a theorem providing a set of sufficient conditions for LAN in this more general setting.

Theorem 3.1. *Let $\mathscr{H}$ be a pre-Hilbert space and suppose that for each $h \in \mathscr{H}$ there exists a sequence $\{\theta_n\} \in \Theta$ and an array $\{h_{nk};\ k=1,\ldots,m_n;\ n=1,2,\ldots\}$ associated with h. Let E_n denote expectation under P_{n,θ_0} and $\tilde{E}_n$ denote expectation under P_{n,θ_n}. Suppose that with $\alpha_{nk} \equiv \alpha_{nk}(\theta_n, \theta_0)$,*

$$\sum_{k=1}^{m_n} \tilde{E}_n[1_{\{\alpha_{nk}=\infty\}}|X_{n1},\ldots,X_{n,k-1}] \xrightarrow{P_0} 0, \tag{3.7}$$

$$\frac{1}{m_n}\sum_{k=1}^{m_n} E_n\left[\sqrt{m_n}(\sqrt{\alpha_{nk}}-1)-\frac{1}{2}h_{nk}\right]^2 \to 0, \tag{3.8}$$

$$\frac{1}{m_n}\sum_{k=1}^{m_n} E_n[h_{nk}^2 1_{\{|h_{nk}|\geqslant\sqrt{n}\varepsilon\}}] \to 0 \quad \textit{for all } \varepsilon > 0, \tag{3.9}$$

$$\frac{1}{m_n}\sum_{k=1}^{m_n} E_n[h_{nk}^2|X_{n1},\ldots,X_{n,k-1}] \overset{P_{n,\theta_0}}{\to} ||h||^2, \tag{3.10}$$

$$\frac{1}{m_n}\sum_{k=1}^{m_n} E_n[h_{nk}^2] \to ||h||^2 \tag{3.11}$$

and

$$\Delta_{n,h} \equiv \frac{1}{\sqrt{m_n}}\sum_{i=1}^{m_n}(h_{nk}-E_n[h_{nk}|X_{n1},\ldots,X_{n,k-1}]) \tag{3.12}$$

is well defined and approximately linear in h; i.e.,

$$\Delta_{n,a_1h_1+a_2h_2} = a_1\Delta_{n,h_1} + a_2\Delta_{n,h_2} + \mathrm{o}_{P_0}(1).$$

Then

1. $\Lambda_n(\theta_n,\theta_0) = \Delta_{n,h} - \frac{1}{2}||h||^2 + r_n$ *where* $r_n \xrightarrow{P_0} 0$, *and*
2. $\mathscr{L}(\Delta_{n,h}|P_0) \to \mathrm{N}(0, ||h||^2)$.

The proof of the above theorem, given in Appendix A, follows Strasser (1985, Section 74) and Strasser (1989) who considered arrays of independent but not necessarily identically distributed observations. Our proof applies in the present more general setting that allows dependent observations. For a similar set of sufficient conditions for LAN in non-i.i.d. contexts when the parameter space is finite-dimensional, see also Ibragimov and Khas'minskii (1975).

Remark 3.2. The arrays $\{h_{nk};\ k=1,\ldots,m_n;\ n=1,2,\ldots\}$ are the real key to guaranteeing the right form of the asymptotic expansion of the log likelihood ratios. However the LAN definition requires that this expansion be indexed by elements h of a geometric (Hilbert) space. The connection between a tangent h and the corresponding array $\{h_{nk};\ k=1,\ldots,m_n;\ n=1,2,\ldots\}$ is specified by conditions (3.10)–(3.12). Because these conditions all involve approximations that improve as n tends to infinity, h may be thought of as a "feature" of the array $\{h_{nk}\}$ in the limit. This loose specification of the connection between tangents and the associated arrays leaves a certain amount of freedom in choosing the tangent space in applications. A concrete construction of tangents is given in Definition 3.5 and this construction is carried out in Examples 2 and 3 below.

Remark 3.3. Conditions (3.9)–(3.11) guarantee that

$$\{h_{nk} - E[h_{nk}|X_{n1},\dots,X_{n,k-1}] : k = 1,\dots,m_n\}$$

is a Martingale difference array which leads to the required asymptotic normality in conclusion 2.

Remark 3.4. Often the conditional expectations $E[h_{nk}|X_{n1},\dots,X_{n,k-1}]$ are identically zero and the support condition (3.7) is trivially satisfied. Then the above result asserts that if the conditional likelihood ratios along paths in the model are approximated in the sense of condition (3.8) by a Martingale difference array with properties (3.9)–(3.11) and each array is associated with an element of a Hilbert space such that (3.12) holds, then the model is LAN at θ_0.

For i.i.d. data, (3.7) becomes $nP_{\theta_n}\{p(X|\theta_0) = 0\} \to 0$, while (3.8) becomes (3.4). As described at the beginning of this section, these two conditions are equivalent to the more familiar Hellinger- or pathwise-differentiability with tangent h (Eq. (3.5)). Conditions (3.9)–(3.11) are satisfied if h is in $L_2(P_{\theta_0})$, and $\varDelta_{n,h}$ is taken to be $n^{-1/2}\sum_{i=1}^{n} h(X_i)$ which is clearly linear in h. In this situation the tangents really are tangents. Since each h is defined by (3.4) $\frac{1}{2}h$ is an $L_2(P_{\theta_0})$ tangent to the path indexed by η. Our h in general is just something constructed to verify LAN, but we are going to use it analogously.

As indicated in Remark 3.2, Theorem 3.1 leaves open the identification of $\mathscr{H}$. In the absence of any intuition about what form the tangents should take, a systematic approach is also available, based on Strasser's (1989) approach. With $h_{n0} = 0$ for all n, the step function $h_n(\cdot,t) = \sum_{k=0}^{n} h_{nk}(\cdot)1_{\{[nt]=k\}}$ is in $L_2(P_{n,\theta_0} \times \lambda)$, where λ is Lebesgue measure on $[0,1]$. If the sequence $\{h_n\}$ converges in $L_2(P_{\theta_0} \times \lambda)$, then the limit, call it h, could be used as a tangent vector since it belongs to a Hilbert space and has the property $||h||^2 = \lim_{n\to\infty}||h_n||^2 = \sigma^2$. The functions $\varDelta_{n,h}$ from the LAN definition are then given by

$$\varDelta_{n,h} = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}(h_{nk} - E[h_{nk}|\mathscr{F}_{n(k-1)}]).$$

We summarize the above in the following definition.

Definition 3.5. Let $h \in L_2(P_{\theta_0} \times \lambda)$. If there exists a sequence $\{\theta_n\} \in \Theta$ such that with $\alpha_{nk} = \alpha_{nk}(\theta_n, \theta_0)$ the conditions of Theorem 3.1 are satisfied by an array $\{h_{n1},\dots,h_{nm_n}\}$; $n = 1,2,\dots$ with each h_{nk} $\mathscr{F}_{nk}$-measurable, and furthermore that with

$$h_n(\cdot,t) = \sum_{k=1}^{n} h_{nk}(\cdot)1_{\{[nt]=k\}} \quad \text{we have } ||h_n - h||_{L_2(P_{\theta_0}\times\lambda)} \to 0, \tag{3.13}$$

then call h the tangent vector corresponding to $\{\theta_n(h)\} \equiv \{\theta_n\}$. The tangent set $\mathscr{H}^0 \subset L_2(P_{\theta_0} \times \lambda)$, is the collection of all h as in (3.13).

Although this definition does not describe the random variables $\Delta_{n,h}$ explicitly in terms of the tangent h (instead they are defined implicitly via the array $\{h_{nk}\}$ associated with h) the next proposition shows it is enough to guarantee (approximate) linearity of $\Delta_{n,h}$ in h over any linear subspace of $\mathscr{H}_0$. The implication is that the model is LAN indexed by such a subspace.

Proposition 3.6. *If $\mathscr{H}$ is a linear subspace of $\mathscr{H}^0$, then the model is LAN at θ_0 indexed by $\mathscr{H}$. In particular, if $\mathscr{H}^0$ itself is linear, the model is LAN at θ_0 indexed by $\mathscr{H}^0$.*

Proof. Let $h_1, h_2 \in \mathscr{H}$ and $a_1, a_2 \in R$. Then $a_1 h_1 + a_2 h_2 \in \mathscr{H}$ so that there exists a sequence $\tilde{h}_n$ and an array of elements $\tilde{h}_{nk}$ that satisfy (3.8)–(3.11), and give rise to the random variable $\Delta_{n,a_1h_1+a_2h_2}$ in the expansion. In addition there are sequences h_{1n} and h_{2n} with arrays h_{1nk} and h_{2nk} corresponding to h_1 and h_2, respectively. These are such that $a_1 h_{1n} + a_2 h_{2n} \to a_1 h_1 + a_2 h_2$ while $\tilde{h}_n \to a_1 h_1 + a_2 h_2$ by definition. Thus $||a_1 h_{1n} + a_2 h_{2n} - \tilde{h}_n|| \to 0$. But this convergence translates into the type of convergence of arrays in (3.8). In light of Remark A.4 (Appendix A) we see that the array $a_1 h_{1nk} + a_2 h_{2nk}$ satisfies the same expansion of the log likelihood ratio as $\tilde{h}_n$. Thus $\Delta_{n,a_1h_1+a_2h_2} = a_1 \Delta_{n,h_1} + a_2 \Delta_{n,h_2} + o_P(1)$. From this approximate linearity and the form of the expansion, which satisfies Eqs. (2.1) and (2.3) of the general LAN definition, we conclude the model is LAN at θ_0 indexed by $\mathscr{H}$. □

See Strasser (1989) for a more thorough treatment of tangent vectors in the case of independent but not identically distributed observations. One example of such a sampling scheme is the following.

Example 2 (*Bivariate three-sample model*)**.** In the bivariate three-sample model discussed in Van der Vaart and Wellner (1991), the first sample consists of pairs $(X_{11}, Y_{11}), \ldots, (X_{1n_1}, Y_{1n_1})$ from a bivariate distribution P. In the second sample we only observe the first margin, that is $X_{21}, \ldots, X_{2n_2}$, while in the third sample we observe the second margin, $Y_{31}, \ldots, Y_{3n_3}$. The parameter of interest can be taken to be the probability measure P itself – there is no parametric component. This model can be viewed as a missing data model where the Y's are missing in the second sample and the X's are missing in the third sample. It could, of course, be extended to a case where the "complete" observation was p-variate and we could have as many as $K = \sum_{i=1}^{p} \binom{p}{i} = 2^p - 1$ samples.

The computations here follow those in Example 4.2 of Van der Vaart and Wellner (1991). Consider first an i.i.d. non-parametric model $\mathscr{P}$ of measures on a measurable space $(\mathscr{X} \times \mathscr{Y}, \mathscr{A} \times \mathscr{B})$ and a point $P_0 \in \mathscr{P}$. We observe n_1 complete observations, n_2 observations on the first margin only and n_3 observations on the second margin only giving a total of n observations. It is supposed that each sequence n_i/n converges to a *known* constant λ_i with $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

It can be shown that for an arbitrary function $g \in G_0 = \{\text{all bounded } L_2^0(P_0) \text{ functions}\}$, g is a tangent in the i.i.d. sense corresponding to the path P_η given by

$$P_\eta(A \times B) = \int_{A \times B} (1 + \eta g) \, \mathrm{d}P_0$$

for $\eta \in R$ small enough so that $\mathrm{d}P_\eta / \mathrm{d}P_0 = 1 + \eta g \geqslant 0$. Let $P_{\eta,X}$ and $P_{\eta,Y}$ denote the marginal probability measures of P_η. To compute tangents in these marginal models we refer to Proposition A.5 in BKRW which states that the corresponding tangents for the model where we observe X only is $g_1(X) = E[g(X,Y)|X]$ and when we observe Y only is $g_2(Y) = E[g(X,Y)|Y]$.

To verify the conditions of Theorem 3.1 suppose the observations are arranged so that the n_1 complete observations come first, the n_2 observations with X only come second, and the n_3 observations with Y only come third.

For a total sample of size n sufficiently large, take η from the definition of the paths to be $n^{-1/2}$ and let the array of h_{nk} elements in the nth row be given by

$$h_{nk} = g1_{\{1,\dots,n_1\}}(k) + g_1 1_{\{n_1+1,\dots,n_1+n_2\}}(k) + g_2 1_{\{n_1+n_2+1,\dots,n\}}(k).$$

From the independence of the observations we have

$$\alpha_{nk} = \begin{cases} \mathrm{d}P_\eta / \mathrm{d}P_0 = 1 + \eta g, & 1 \leqslant k \leqslant n_1, \\ \mathrm{d}P_{\eta,X} / \mathrm{d}P_{0,X} = 1 + \eta g_1, & n_1 + 1 \leqslant k \leqslant n_1 + n_2, \\ \mathrm{d}P_{\eta,Y} / \mathrm{d}P_{0,Y} = 1 + \eta g_2, & n_1 + n_2 + 1 \leqslant k \leqslant n. \end{cases}$$

Note that $P_\eta \ll P_0$, $P_{\eta,X} \ll P_{0,X}$, $P_{\eta,Y} \ll P_{0,Y}$, so that any sequence along such a path satisfies (3.7). The above construction emphasizes the three i.i.d. subproblems. Using the i.i.d. theory on these subproblems allows straightforward verification of conditions (3.8) and (3.9).

The expression in condition (3.10) becomes

$$\frac{n_1}{n} \int g^2 \, \mathrm{d}P_0 + \frac{n_2}{n} \int g_1^2 \, \mathrm{d}P_{0,X} + \frac{n_3}{n} \int g_2^2 \, \mathrm{d}P_{0,Y}$$

which converges to $\lambda_1 ||g||^2 + \lambda_2 ||g_1||^2 + \lambda_3 ||g_3||^2 \equiv \sigma^2$. Thus (3.11) is also satisfied and hence all the necessary conditions for Definition 3.5 are met. The function $h_n(\cdot, t)$ from Definition 3.5 is

$$\sum_{k=1}^{n_1} g1_{\{[nt]=k\}} + \sum_{k=1}^{n_2} g_1 1_{\{[nt]=k+n_1\}} + \sum_{k=1}^{n_3} g_2 1_{\{[nt]=k+n_1+n_2\}}$$
$$= g1_{[0,n_1/n]}(t) + g_1 1_{(n_1/n,(n_1+n_2)/n]}(t) + g_2 1_{(n_1+n_2)/n,1]}(t).$$

From the convergence $n_i/n \to \lambda_i$ for all i, and the square integrability of the function g, the sequence $\{h_n\}$ can be seen to converge in $L_2(P_{\theta_0} \times \lambda)$ (or simply $L_2(P_0 \times \lambda)$) to the tangent $h(\cdot, t) = g1_{[0,\lambda_1]}(t) + g_1 1_{(\lambda_1,\lambda_1+\lambda_2]}(t) + g_2 1_{(\lambda_1+\lambda_2,1]}(t)$. Since the collection of functions G_0 is a linear space and g can be taken to be any member of G_0, the collection of all such h is a linear space. Thus the conclusion is that $\mathscr{H}_0$ is linear and hence the model is LAN at P_0 indexed by $\mathscr{H}_0$.

Another choice for the tangent, as in Van der Vaart and Wellner (1991), would be $h=(g,g_1,g_2)$ in a space where the inner product is defined by

$$\langle\tilde{h},h\rangle=\lambda_1\langle\tilde{g},g\rangle+\lambda_2\langle\tilde{g}_1,g_1\rangle+\lambda_3\langle\tilde{g}_2,g_2\rangle.$$

Here it is clear how the functions $\Delta_{n,h}$ depend on the tangent vector and it is easily seen that they are linear in h.

Example 3 (*Case-control data*)**.** Case-control designs are an effective tool for studying the relationship between exposures of interest and rare outcomes. Such a design is an example of a two-sample problem. Let the variable Y indicate case or control status of an individual with 0 being disease free and 1 being diseased. In addition, suppose a covariate vector $X=(X_1,\ldots,X_p)'$ of exposures and other factors related to disease is available. In a common form of the case-control design, covariate information for all or a random sample of subjects who develop disease in a specified “case accession” period of time are recorded, along with the information on a random sample of disease free individuals. Let n_0 and n_1 denote the number in each of the two samples and n the total number of observations. In this sampling scheme, we obtain realizations from the distribution of X conditional on case/control status, while the parameters of interest will typically be from a prospective model with

$$\Pr(Y=1|X=x)=\frac{\exp\{\alpha+X^{\mathrm{T}}\beta\}}{1+\exp\{\alpha+X^{\mathrm{T}}\beta\}}.$$

Here α is an intercept and β is a p-vector of regression, or odds-ratio parameters. The vector β is of primary interest in this problem. The marginal density $g(x)$ of X is the infinite-dimensional nuisance parameter in this model.

The parameter space can be described as $\Theta\times\mathscr{G}$ where $\Theta\subset\mathbb{R}^{p+1}$ corresponds to the regression parameters and $\mathscr{G}$ is the space of distributions for X. The usual approach for computing variance bounds in the estimation problem involving prospective models for case-control data is to alter the problem slightly so that it is an i.i.d. model. The i.i.d. modification is given by a two-stage sampling procedure. First, either a case or control is selected with probabilities λ_1 and λ_0, respectively, where these probabilities are assumed known. Second, the covariate X is sampled for the individual drawn at step 1. The sample sizes n_1 and n_0 are now regarded as random. See Breslow and Wellner (1997) for recent work establishing the efficiency of logistic regression for estimating regression parameters in this modified problem. It is widely believed that variance bounds obtained for this model are valid in the two-sample model that is used in practice. In Section 4 we show this to be true, at least when the proportions λ_1 and λ_0 are assumed known.

The approach for the two-sample version is similar to Example 2 (a three sample model) in that we use i.i.d. theory to find tangents within each of the two samples, and then form the tangents in the sense of Definition 3.5 based on these. The usual development is to first compute scores in the *prospective* model, where both disease outcome Y and the covariates X are considered random, and then relate these to the scores in the retrospective model. Following the calculations in BKRW Section 4.4, we obtain scores at a point $((\alpha,\beta)_0, g_0)$ of the form $\dot{l}(\alpha,\beta)^{\mathrm{T}}\xi + a$. Here

$$\dot{l}(\alpha,\beta)^{\mathrm{T}}\xi = (X^e(Y - E(Y|X;\alpha,\beta)))^{\mathrm{T}}\xi,$$

with $X^e = (1, X^{\mathrm{T}})^{\mathrm{T}}$, is the score along a path $\eta \mapsto (\alpha,\beta)_0 + \eta\xi$ (with $\xi \in \mathbb{R}^{p+1}$, $\eta \in \mathbb{R}$) evaluated at $\eta = 0$. The second component $a \in L_2(G_0)$ (with G_0 being the distribution function corresponding to the density g_0 of the covariates) satisfies

$$\int \left[\eta^{-1}(\sqrt{g_\eta} - \sqrt{g_0}) - \frac{1}{2}a\sqrt{g_0}\right]^2 \mathrm{d}\mu \to 0 \quad \text{as } \eta \downarrow 0$$

for a path g_η through the infinite-dimensional part of the parameter space.

Tangents for the retrospective model are then given by

$$h_i = \dot{l}(\alpha,\beta)^{\mathrm{T}}\xi + a - E(\dot{l}(\alpha,\beta)^{\mathrm{T}}\xi + a|Y = i) \equiv \Phi_i(\xi, a)$$

for $i = 0, 1$ corresponding to the control and case samples respectively (BKRW, pp. 116–117).

Now, suppose the data are arranged so that the controls are listed first and then the cases. Since the tangents h_0 and h_1 are chosen to satisfy a pathwise-differentiability condition like (3.5), the support condition (3.7) is automatically satisfied, as is condition (3.8). Verification of conditions (3.9)–(3.11) is straightforward and follows the calculations from Example 2 using the current h_0 and h_1. We conclude that the tangent in the sense of Definition 3.5 is

$$h(\cdot,t) = h_0 1_{[0,\lambda_0]}(t) + h_1 1_{(\lambda_0,1]}(t) = \Phi_0(\xi,a)1_{[0,\lambda_0]}(t) + \Phi_1(\xi,a)1_{(\lambda_0,1]}(t) \equiv \dot{l}_{\theta_0}(\xi,a).$$

The last two expressions emphasize that h is also a function of ξ and a which determine the path. Furthermore, the operator $\dot{l}_{\theta_0}$ that maps from ξ, a to h is linear. As a result, the collection of such h as ξ ranges over $\mathbb{R}^{p+1}$ and a ranges over $L_2^0(G_0)$ is a linear space which leads to the conclusion, via Proposition 3.6, that the model at $((\alpha,\beta)_0, g_0)$ is LAN indexed by this space. For future reference we also introduce the following notation:

$$\begin{aligned}\dot{l}_{\theta_0}(\xi,a) &= [\dot{l}(\alpha,\beta) - E(\dot{l}(\alpha,\beta)|Y=0)1_{[0,\lambda_0]}(t) + E(\dot{l}(\alpha,\beta)|Y=1)1_{(\lambda_0,1]}(t)]^{\mathrm{T}}\xi \\ &\quad + [a - E(a|Y=0)1_{[0,\lambda_0]}(t) + E(a|Y=1)1_{(\lambda_0,1]}(t)] \\ &\equiv \dot{l}_1^{\mathrm{T}}\xi + \dot{l}_2(a). \end{aligned} \tag{3.14}$$

As in Example 2, we note an alternative definition of the tangents could be based on h_0, h_1 and the relative sample sizes λ_0 and λ_1 as in the remarks at the end of Example 2 above.

The above examples all involve independent, but not identically distributed data. For an application of Theorem 3.1 to a model with dependent observations, see Breslow et al. (1998).

As illustrated by Example 3, when the model is specified by a parameterization (other than the probability measures themselves) the tangent set can often be conveniently described as the range of a linear operator. Because these so-called score operators play an important role in the differentiability conditions in Section 4 we need a more precise definition.

3.3. Score operators

As we have seen, a natural way to construct tangents at a point in the model is to introduce a linear operator that acts on tangents to paths in the parameter space. To describe paths in a general Θ that converge to a point θ_0, we at least need a topology, say τ, on Θ. For convenience let us also assume that Θ is a vector space and that the operations of addition and scalar multiplication are continuous under our topology; i.e. (Θ, τ) is a topological vector space. Most often in the study of efficiency Θ is even taken to be a (subset of) a Hilbert space. However, in the following definition we leave the structure of Θ open; all we require is that the dual space of (Θ, τ) separate points of Θ so that adjoints are well defined. In the case Θ is a product space with product topology this is true if the dual of each coordinate space separates points of that coordinate space (Lemma B.1). It is also true of any normed space.

The following is a formal definition of the concept of a score operator already used in the previous section. The notation $\overline{\mathrm{lin}(\Theta)}$ indicates the closed linear span of Θ.

Definition 3.7 (*Score operators*)**.** Let $t \mapsto \theta_t$ be a path in Θ converging to θ_0 as $t \downarrow 0$ with an element $\dot{\theta} \in \overline{\mathrm{lin}(\Theta)}$ such that $(\theta_t - \theta_0)/t \to \dot{\theta}$ in (Θ, τ) as $t \downarrow 0$. Let $\dot{\Theta}$ be the set of all $\dot{\theta}$ obtained in this way. We say $\dot{l}_{\theta_0} : \Theta \to \mathscr{H}$ is a score operator at θ_0 if for all $\dot{\theta} \in \dot{\Theta}$ there exists a sequence $c_n \downarrow 0$ such that θ_{c_n} forms a sequence $\theta_n(h)$ such that $P_{n,\theta_n(h)}$ has tangent h as in the LAN definition (Definition 2.1) and $\dot{l}_{\theta_0}(\dot{\theta}) = h$.

Linearity of $\dot{\Theta}$ and the score operator imply the image of $\dot{l}_{\theta_0}$ is a linear space. Thus, if in addition the random variables $\Delta_{n,h}$ are linear in h, the model is LAN indexed by the image space.

4. Differentiability

In non-parametric models where the parameter of interest is naturally stated as a function of the underlying probability measure, differentiability can be established directly, as in the following.

Example 2 (*Cont.*)**.** Here we continue with the calculations in Van der Vaart and Wellner (1991). In this example the parameter of interest is the probability measure P itself. To make this a parameter in a Banach space first consider a collection of square integrable functions $\mathscr{F}$. To simplify matters we might even suppose this collection of functions is also uniformly bounded since here we are thinking of, for example, the collection of indicators of measurable sets. Now we may take the Banach space B to be $l^{\infty}(\mathscr{F})$ – the space of all bounded real-valued functions on $\mathscr{F}$. Then the goal may be stated as estimation of $v_n(P_{n,0})$ defined by

$$v_n(P_{n,0}) = \int f(x_{11}, y_{11})\, \mathrm{d}P_0(x_{11}, y_{11}) = \frac{1}{n_1}\sum_{i=1}^{n_1} \int f(X_{1i}, Y_{1i})\, \mathrm{d}P_0.$$

Taking a sequence $\eta_n = n^{-1/2}$ along a path corresponding to a bounded, measurable, mean 0 function g as specified in Section 3, Example 2, this implies

$$\sqrt{n}(v_n(P_{n,\eta_n}) - v_n(P_{n,0})) = \int gf\, \mathrm{d}P_0 = \int g(f - P_0 f)\, \mathrm{d}P_0,$$

where $Pf = \int f\, \mathrm{d}P$ and the last equality follows from the fact that g has mean 0. Thus (2.5) holds with $R_n = \sqrt{n}$ and we can identify $\dot{v}(h)$ as

$$\dot{v}(h)f = \left\langle \frac{1}{\lambda_1}(f - P_0 f, 0, 0), (g, g_1, g_2) \right\rangle.$$

Here we are using the version of the tangent space used in Van der Vaart and Wellner (1991), rather than the version corresponding to Definition 3.5.

To identify the adjoint it suffices, because B is a function space, to consider the evaluation maps $\pi_f \in B^*$ defined by $\pi_f b = b(f)$ for all $b \in B$. Then the adjoint $\dot{v}^{\mathrm{T}}$ is defined by

$$\begin{aligned}\pi_f \dot{v}(h) = \dot{v}(h)f &= \left\langle \frac{1}{\lambda_1}(f - P_0 f, 0, 0), (g, g_1, g_2) \right\rangle \\ &= \left\langle \frac{1}{\lambda_1}\left(f - P_0 f - a_f - b_f, \frac{\lambda_1}{\lambda_2} a_f, \frac{\lambda_1}{\lambda_3} b_f \right), (g, g_1, g_2) \right\rangle \\ &= \langle \dot{v}^{\mathrm{T}} \pi_f, (g, g_1, g_2) \rangle, \end{aligned} \tag{4.1}$$

where a_f and b_f satisfy

$$E(\lambda_2 f^o(X, Y) - (\lambda_1 + \lambda_2) a_f(X) - \lambda_2 b_f(Y) | X = x) = 0$$

and

$$E(\lambda_3 f^o(X, Y) - \lambda_3 a_f(X) - (\lambda_1 + \lambda_3) b_f(Y) | Y = y) = 0$$

and $f^o(X, Y) \equiv f(X, Y) - Ef(X, Y)$. These last two conditions are derived from the fact that $\dot{v}^{\mathrm{T}}$ must map into $\mathscr{H}$ (which is identified with $\mathscr{H}^*$) so that the second element of $\dot{v}^{\mathrm{T}}\pi_f$ must be the conditional expectation of the first given X and the third element must be the conditional expectation of the first given Y. From these equations it follows that

$$\begin{aligned}\mathrm{Cov}(\mathbb{Z}(f), \mathbb{Z}(g)) &= \langle \dot{v}^{\mathrm{T}}\pi_f, \dot{v}^{\mathrm{T}}\pi_g \rangle \\ &= \frac{1}{\lambda_1} E(f^0 - a_f - b_f)(g^0 - a_g - b_g) + \frac{1}{\lambda_2} E(a_f a_g) + \frac{1}{\lambda_3} E(b_f b_g) \\ &= \frac{1}{\lambda_1} E(f^0 - a_f - b_f) g^0. \end{aligned}$$

The last equality follows from the fact that

$$-\frac{1}{\lambda_1} E[(f^0 - a_f - b_f) a_g] = -\frac{1}{\lambda_1} E\{E[(f^0 - a_f - b_f)|X] a_g\} = -\frac{1}{\lambda_1} E\left\{ \frac{\lambda_1}{\lambda_2} a_f a_g \right\}$$

resulting in some cancellation, with a similar calculation for $-1/\lambda_1 E[(f^0 - a_f - b_f)b_g]$ resulting in further cancellation.

Although these equations characterize the adjoint and provide expressions for the covariances of influence functions, we cannot compute these covariances explicitly except in special cases. For instance, under independence, $E(a(X)|Y) = E(a(X))$ and $E(b(Y)|X)=E(b(Y))$. If we choose $a(X)=k_1E(f^o(X,Y)|X)$ and $b(Y)=k_2E(f^o(X,Y)|Y)$ with $k_1=\lambda_2/(\lambda_1+\lambda_2)$ and $k_2=\lambda_3/(\lambda_1+\lambda_3)$, then $E(a(X)|Y)=E(a(X))=0=E(b(Y))=E(b(Y)|X)$ and it is easily verified that the equations defining $\dot{v}^{\mathrm{T}}\pi_f$ hold.

Now we may make some calculations of relative efficiencies along the lines of those in Bickel et al. (1991). To simplify matters, take the sample space to be $[0,1]\times[0,1]$ and $f_{st}(X,Y) = 1_{[0,s]}(X)1_{[0,t]}(Y)$ so that $f^o_{st}(X,Y) = 1_{[0,s]}(X)1_{[0,t]}(Y) - st$. In this case we have $a(X) = k_1t(1_{[0,s]}(X) - s)$ and $b(Y) = k_2s(1_{[0,t]}(Y) - t)$. Direct calculation of the variance of $\dot{v}^{\mathrm{T}}\pi_f$ using the formula given above yields

$$\frac{1}{\lambda_1}E(f^{o2}_{st} - a_f f^o_{st} - b_f f^o_{st}) = \frac{1}{\lambda_1}(st(1-st) - k_1t^2s(1-s) - k_2s^2t(1-t)),$$

while the asymptotic variance of the estimator that uses only the complete data is given by

$$\frac{1}{\lambda_1}E(f^{o2}) = \frac{1}{\lambda_1}st(1-st).$$

Thus the asymptotic relative efficiency of the efficient estimator to the crude estimator is

$$\frac{E(f^{o2}_{st} - a_f f^o_{st} - b_f f^o_{st})}{E(f^{o2})} = \frac{(st(1-st) - k_1t^2s(1-s) - k_2s^2t(1-t))}{st(1-st)}.$$

With $\lambda_1 \approx 0$ and thus $k_1 \approx k_2 \approx 1$ this gives a relative efficiency of approximately $(1-s)(1-t)/(1-st)$ which agrees with Bickel et al. (1991).

When $\lambda_1=\lambda_2=\lambda_3=1/3$ we have $k_1=k_2=1/2$ and an ARE of $(1-(t+s)/2)/(1-st)$. This can be as small as 1/2 (when either s or t are 1) and as big as 1. At $t=s=1/2$ the ARE is 2/3 (cf. with 1/3 when $\lambda_1\approx 0$), and when $s=t$ in general we get an ARE of $(1+t)^{-1}$. We also see that when $\lambda_1 \approx 1$ we have $k_1 \approx k_2 \approx 0$ and an ARE of approximately 1.

Even when the parameter of interest is specified via a parametrization of the model, it is possible to verify differentiability using a "projection of scores" method.

Example 3 (*Cont.: Semiparametric models*). For semiparametric models when the finite-dimensional parameter, or some function q of this parameter, is of interest, one usually proceeds to calculate $\dot{v}$ via projection of scores. To describe this approach we adopt the more common notation from semiparametrics. Let now θ denote a k-dimensional parameter of interest and g the infinite-dimensional nuisance parameter. Corresponding to these are $\dot{\Theta}$ and $\dot{G}$. In Example 3, for instance, these were given by $\mathbb{R}^{p+1}$ and $L^0_2(G_0)$, respectively. The space $\dot{\Theta}\times\dot{G}$ replaces the space $\dot{\Theta}$ from the definition of score operators (Definition 3.7). In such situations the score operator also has two components. These were labeled $\dot{l}_1$ and $\dot{l}_2$ at the end of Example 3 (Eq. (3.14)), but we now use $\dot{l}_\theta$ and $\dot{l}_g$ to emphasize the point θ, g in the parameter space.

The estimation problem, as a function of the probability measures is $v(P_{n,(\theta,g)})=q(\theta)$, for $q : \mathbb{R}^k \to \mathbb{R}^m$. A derivative of the form $\dot{v}(\dot{l}_\theta^{\mathrm{T}}\xi + \dot{l}_g(a)) = \dot{q}(\theta)\xi$, where $\dot{q}$ is the $m \times k$ derivative matrix, satisfies (2.5) with $R_n = n^{1/2}$. We must, however, identify $\dot{v}$ as a function of $\dot{l}_\theta^{\mathrm{T}}\xi + \dot{l}_g(a)$.

First consider projection of the score function $\dot{l}_\theta$ onto the subspace $\overline{\dot{l}_g(\dot{G})}$. Let $\Pi(\cdot|\overline{\dot{l}_g(\dot{G})})$ be the projection operator. For example, we have from Theorem 2 of Appendix 2 of BKRW, that when $\dot{l}_g(\dot{G})$ is already closed, the projection of an element y onto $\dot{l}_g(\dot{G})$ is given by

$$\Pi(y|\dot{l}_g(\dot{G})) = \dot{l}_g(\dot{l}_g^{\mathrm{T}}\dot{l}_g)^{-}(\dot{l}_g^{\mathrm{T}}y).$$

where $(\dot{l}_g^{\mathrm{T}}\dot{l}_g)^{-}(\dot{l}_g^{\mathrm{T}}y)$ is a solution of $(\dot{l}_g^{\mathrm{T}}\dot{l}_g)x = (\dot{l}_g^{\mathrm{T}}y)$. When the inverse $(\dot{l}_g^{\mathrm{T}}\dot{l}_g)^{-1}$ exists, this is just $(\dot{l}_g^{\mathrm{T}}\dot{l}_g)^{-1}(\dot{l}_g^{\mathrm{T}}y)$.

Now let $l_\theta^* = \dot{l}_\theta - \Pi(\dot{l}_\theta|\overline{\dot{l}_g(\dot{G})})$, i.e. the projection of $\dot{l}_\theta$ onto the orthocomplement of $\overline{\dot{l}_g(\dot{G})}$, and suppose $I^* = \langle l_\theta^*, l_\theta^{*\prime}\rangle_{\mathscr{H}}$ is non-singular. Then we have

$$\begin{aligned} I^{*-1}\langle l_\theta^*, (\dot{l}_\theta'\xi + \dot{l}_g(\dot{g}))\rangle_{\mathscr{H}} &= I^{*-1}\langle l_\theta^*, l_\theta^{*\prime}\xi + \Pi(\dot{l}_\theta|\overline{\dot{l}_g(\dot{G})})'\xi + \dot{l}_g(\dot{g})\rangle_{\mathscr{H}} \\ &= I^{*-1}\langle l_\theta^*, l_\theta^{*\prime}\xi\rangle_{\mathscr{H}} + I^{*-1}\langle l_\theta^*, \Pi(\dot{l}_\theta|\overline{\dot{l}_g(\dot{G})})'\xi + \dot{l}_g(\dot{g})\rangle_{\mathscr{H}} \\ &= I^{*-1}I^*\xi + \mathbf{0} = \xi, \end{aligned}$$

where the $\mathbf{0}$ term follows from the fact that $\Pi(\dot{l}_\theta|\overline{\dot{l}_g(\dot{G})})'\xi + \dot{l}_g(\dot{g})$ is in $\overline{\dot{l}_g(\dot{G})}$ while l_θ^* is a vector of elements in $\overline{\dot{l}_g(\dot{G})}^{\perp}$. Thus a candidate for $\dot{v}(\cdot)$ is $\langle \dot{q}I^{*-1}l_\theta^*, \cdot\rangle_{\mathscr{H}}$. The only concern is that $\dot{q}I^{*-1}l_\theta^*$ be a vector of elements in $\mathscr{H}$. This follows from the calculation

$$\dot{q}I^{*-1}l_\theta^* = \dot{q}I^{*-1}\dot{l}_\theta - \dot{q}I^{*-1}\Pi(\dot{l}_\theta|\overline{\dot{l}_g(\dot{G})}),$$

so that the first term on the right-hand side is a vector of elements in $[\dot{l}_\theta]$ and the second term is a vector of linear combinations of elements of $\overline{\dot{l}_g(\dot{G})}$ and hence is a vector of elements of $\overline{\dot{l}_g(\dot{G})}$. Conclude that

$$\dot{q}I^{*-1}l_\theta^* \in [\dot{l}_\theta] + \overline{\dot{l}_g(\dot{G})} = \mathscr{H}.$$

From this point, computation of the adjoint $\dot{v}^{\mathrm{T}}$ is straightforward and we find that for b^* in $\mathbb{R}^m$, $\dot{v}^{\mathrm{T}}b^* = b^{*\prime}I^{*-1}l_\theta^*$ which has norm $b^{*\prime}I^{*-1}b^*$.

This entire development is completely analogous to the i.i.d. case and when one begins calculating scores and projections, the similarity with the i.i.d. formulation of Example 3 becomes apparent. In fact, the information bounds are identical and therefore logistic regression is still efficient. This will be shown more formally in the next section via a corollary to the differentiability theorem (Theorem 4.1).

Besides the above projection-of-scores approach, another useful method for verifying differentiability and identifying the adjoint $\dot{v}^{\mathrm{T}}$ in the i.i.d. framework is via a theorem due to Van der Vaart (1991). The next section describes how this carries over to the present context.

4.1. Differentiability of implicitly defined functionals

Returning to the general notation, we consider estimation of functions of $P_{n,\theta}$ of the form

$$v_n(P_{n,\theta}) = \psi_n(\theta) \tag{4.2}$$

for a sequence of maps ψ_n from the set Θ to a Banach space B. Assume there exists a continuous linear map $\dot{\psi}: \dot{\Theta} \to B$ with

$$R_n(\psi_n(\theta_{c_n}) - \psi_n(\theta)) - \dot{\psi}(\dot{\theta}) \to 0 \tag{4.3}$$

for c_n corresponding to a $\dot{\theta}$ as in Definition 3.7. This occurs, for example, if there exists a map $\psi : \Theta \to B$ such that

$$R_n(\psi(\theta_{c_n}) - \psi(\theta)) - \dot{\psi}(\dot{\theta}) \to 0 \quad \text{as } n \to \infty$$

and $\{\psi_n\}$ converges to ψ in the sense that

$$R_n(\psi_n(\theta_{c_n}) - \psi(\theta_{c_n})) \to 0 \quad \text{as } n \to \infty.$$

Then

$$\begin{aligned} &\lim_{n\to\infty} R_n(\psi_n(\theta_{t_n}) - \psi_n(\theta)) \\ &\quad = \lim_{n\to\infty} \{R_n(\psi_n(\theta_{t_n}) - \psi(\theta_{t_n})) - R_n(\psi_n(\theta) - \psi(\theta)) + R_n(\psi(\theta_{c_n}) - \psi(\theta))\} \end{aligned}$$

and the convergence assumption ensures the limits of the first two parts in the right-hand side of the above are 0.

We further assume there is an $N \geqslant 0$ such that for $n \geqslant N$, v_n is well defined. That is, if $P_{n,\theta_1} = P_{n,\theta_2}$, then $\psi_n(\theta_1) = \psi_n(\theta_2)$.

For this discussion we can specialize Definition 2.2 and define the sequence $\{v_n\}$ to be differentiable if, for any sequence $\{c_n\}$ as in Definition 3.7 (Section 3),

$$R_n(v_n(P_{n,\theta_{c_n}}) - v_n(P_{n,\theta})) - \dot{v}(\dot{l}(\dot{\theta})) \ \to \ 0, \tag{4.4}$$

where $\dot{v}: \mathscr{H} \to B$ is continuous and linear. Thus, we wish to identify conditions under which $\dot{v}(\dot{l}(\dot{\theta})) = \dot{\psi}(\dot{\theta})$. The theorem to provide such conditions in the i.i.d. case was given in Van der Vaart (1991). With the above notation, the argument at the heart of his theorem applies to the present set-up.

Theorem 4.1. *Suppose the topology on the space Θ is chosen so that the dual separates points (making adjoints well defined). Let $\dot{\Theta}$ be the closed subset of Θ described above and suppose $\mathscr{R}(i) \subset \mathscr{H} \subset \overline{\mathscr{R}(i)}$. Then the sequence $\{v_n\}$ as in (4.2) is differentiable in the sense of (2.5) if and only if*

$$\mathscr{R}(\dot{\psi}^{\mathrm{T}}) \subset \mathscr{R}(\dot{l}^{\mathrm{T}}).$$

The adjoint $\dot{v}^{\mathrm{T}}$ *is then uniquely specified by the relation*

$$\dot{\psi}^{\mathrm{T}} b^* = \dot{l}^{\mathrm{T}} \dot{v}^{\mathrm{T}} b^*.$$

Proof. Proceed as in Van der Vaart (1991). First suppose $\dot{v}$ exists. By (4.2)–(4.4) we have that for all $\dot{\theta}$ in $\dot{\Theta}$

$$\dot{v}(\dot{l}(\dot{\theta})) = \lim_{n\to\infty} R_n(v_n(P_{n,\theta_{c_n}}) - v_n(P_{n,\theta})) = \lim_{n\to\infty} R_n(\psi_n(\theta_{c_n}) - \psi_n(\theta)) = \dot{\psi}(\dot{\theta}).$$

Hence, for any element b^* in the dual B^*,

$$b^* \dot{v}(\dot{l}(\dot{\theta})) = b^* \dot{\psi}(\dot{\theta}) = \langle b^*, \dot{\psi}(\dot{\theta})\rangle = \langle \dot{\psi}^{\mathrm{T}} b^*, \dot{\theta}\rangle.$$

But, we also have

$$b^* \dot{v}(\dot{l}(\dot{\theta})) = \langle b^*, \dot{v}(\dot{l}(\dot{\theta}))\rangle = \langle \dot{v}^{\mathrm{T}} b^*, \dot{l}(\dot{\theta})\rangle = \langle \dot{l}^{\mathrm{T}} \dot{v}^{\mathrm{T}} b^*, \dot{\theta}\rangle.$$

Since this holds for all $\dot{\theta}$, we conclude that $\dot{\psi}^{\mathrm{T}} b^* = \dot{l}^{\mathrm{T}} \dot{v}^{\mathrm{T}} b^*$ and this is true for all b^*. Hence $\mathscr{R}(\dot{\psi}^{\mathrm{T}}) \subset \mathscr{R}(\dot{l}^{\mathrm{T}})$. Note that this also implies $\mathscr{R}(\dot{\psi}^{\mathrm{T}})^{\perp} \supset \mathscr{R}(\dot{l}^{\mathrm{T}})^{\perp}$.

Now to prove the converse suppose $\mathscr{R}(\dot{\psi}^{\mathrm{T}}) \subset \mathscr{R}(\dot{l}^{\mathrm{T}})$. We may tentatively define $\dot{v}$ on $\mathscr{R}(\dot{l})$ via

$$\dot{v}(\dot{l}(\dot{\theta})) = \dot{\psi}(\dot{\theta}).$$

That this is well defined can be seen as follows. If $\dot{l}(\dot{\theta}_1) = \dot{l}(\dot{\theta}_2)$, then $\dot{\theta}_1 - \dot{\theta}_2 \in \mathscr{N}(\dot{l}) = \mathscr{R}(\dot{l}^{\mathrm{T}})^{\perp} \subset \mathscr{R}(\dot{\psi}^{\mathrm{T}})^{\perp} = \mathscr{N}(\dot{\psi})$ which implies $\dot{\psi}(\dot{\theta}_1) = \dot{\psi}(\dot{\theta}_2)$. (The equalities involving the annihilators are familiar in the case $\dot{\Theta}$ is a Banach space – cf. Theorem 4.12 of Rudin (1991) – but it is straightforward to show they extend to more general topological vector spaces given more general definitions of adjoints and annihilators. See, for example, the definitions in Kelley and Namioka (1963).) We now must show $\dot{v}$ is continuous and linear. Using the fact that B is a Banach space and Lemma B.2 in Appendix B, it suffices to show it is weakly continuous and linear; that is, for all $b^* \in B^*$, $b^* \dot{v}$ is a continuous linear functional.

By hypothesis, $\dot{\psi}^{\mathrm{T}} b^* \in \mathscr{R}(\dot{l}^{\mathrm{T}})$. Let $h^* \in \mathscr{H}^*$ be such that $\dot{\psi}^{\mathrm{T}} b^* = \dot{l}^{\mathrm{T}} h^*$. Then

$$b^* \dot{v}(\dot{l}(\dot{\theta})) = b^* \dot{\psi}(\dot{\theta}) = \langle b^*, \dot{\psi}(\dot{\theta})\rangle = \langle \dot{\psi}^{\mathrm{T}} b^*, \dot{\theta}\rangle = \langle \dot{l}^{\mathrm{T}} h^*, \dot{\theta}\rangle = \langle h^*, \dot{l}(\dot{\theta})\rangle = h^* \dot{l}(\dot{\theta}).$$

Continuity and linearity of h^* implies the functional $b^* \dot{v}$ is continuous and linear. This concludes the proof that $\dot{v}$ is continuous and linear on $\mathscr{R}(\dot{l})$ and it may be continuously extended to $\mathscr{H}$ if necessary.

To show uniqueness of $\dot{v}^{\mathrm{T}}$ under the condition $\mathscr{R}(\dot{l}) \subset \mathscr{H} \subset \overline{\mathscr{R}(\dot{l})}$, suppose $\dot{v}_1^{\mathrm{T}}$ and $\dot{v}_2^{\mathrm{T}}$ are two solutions. Then

$$\dot{\psi}^{\mathrm{T}} b^* = \dot{l}^{\mathrm{T}} \dot{v}_1^{\mathrm{T}} b^* = \dot{l}^{\mathrm{T}} \dot{v}_2^{\mathrm{T}} b^*$$

so

$$\dot{v}_1^{\mathrm{T}} b^* - \dot{v}_2^{\mathrm{T}} b^* \in \mathscr{N}(\dot{l}^{\mathrm{T}}) = \mathscr{R}(\dot{l})^{\perp} = \overline{\mathscr{R}(\dot{l})}^{\perp} \subset \mathscr{H}^{\perp}.$$

But $\dot{v}_2^{\mathrm{T}} b^*$ and $\phi \dot{v}_1^{\mathrm{T}} b^*$ are both elements of $\bar{\mathscr{H}}$ and therefore so is the difference. Thus $\dot{v}_2^{\mathrm{T}} b^* - \dot{v}_1^{\mathrm{T}} b^* = 0$ in $\mathscr{H}$ for all $b^* \in B^*$ so that $\dot{v}_2^{\mathrm{T}} = \dot{v}_1^{\mathrm{T}}$. □

Remark 4.2. If the space $\dot{\Theta}$ can be assumed to be a subspace of a Hilbert space, the conclusion of the theorem can be restated to provide the influence functions $\dot{v}^{\mathrm{T}}_{b^*}$ which specify the finite-dimensional distributions of the optimal limit random variable in the convolution theorem. Specifically, when $\dot{\Theta}$ is Hilbert, $\dot{\psi}^{\mathrm{T}}$ can be thought of as a map from B^* to $\dot{\Theta}$ itself rather than its dual. Similarly, we can think of $\dot{v}$ as a map from B^* to $\mathscr{H}$ and interpret $\dot{l}^{\mathrm{T}}$ as a map from $\mathscr{H}$ to $\dot{\Theta}$. Then the influence functions $\dot{v}^{\mathrm{T}}_{b^*}$ are determined by the equation $\dot{\psi}^{\mathrm{T}}_{b^*} = \dot{l}^{\mathrm{T}}\dot{v}^{\mathrm{T}}_{b^*}$ where $\dot{\psi}^{\mathrm{T}}_{b^*}$ is the element of $\dot{\Theta}$ such that $\dot{\psi}^{\mathrm{T}} b^* h = \langle \dot{\psi}^{\mathrm{T}}_{b^*}, h\rangle$ for all $h \in \mathscr{H}$. This is how Van der Vaart's theorem for i.i.d. data is stated.

Example 1 (*Cont*)**.** This example provides a simple illustration of the theorem. Here the tangents were not scores in the quadratic mean sense, as is usually the case, but nevertheless the theorem can be used to check differentiability and identify $\dot{v}^{\mathrm{T}}$. Recall that in this problem the tangent space was given by $\mathscr{H} = \mathscr{R}(\dot{l})$; the image of the score operator $\dot{l} : \mathbb{R} \times C^0_b[0,1] \to \mathscr{H}$ with $\dot{l}(a,g) = a + g \equiv h$ for $a \in \mathbb{R}$ and $g \in C^0_b[0,1]$, where the latter is equipped with the supremum norm. Here we take $\mathscr{H}$ to be a subspace of $L_2(\lambda)$. It is interesting to note that $\mathscr{H}$ is *not* a closed subspace of $L_2(\lambda)$ and hence it is not a Hilbert space. Represent $\dot{\Theta}^* = (R \times C^0_b[0,1])^*$ as $\mathbb{R} \times C^0_b[0,1]^*$; that is, for $\dot{\theta} = (a,g)$, $\dot{\theta}^*\dot{\theta} = \dot{\theta}^*_1 a + \dot{\theta}^*_2 g$ where $\dot{\theta}^*_1 \in \mathbb{R}$ and $\dot{\theta}^*_2 \in C^0_b[0,1]^*$.

By direct calculation, we find the adjoint $\dot{l}^{\mathrm{T}} : L_2(\lambda)^* \to \dot{\Theta}^*$, defined via

$$\langle h^*, \dot{l}(\dot{\theta})\rangle_{\mathscr{H}} = \langle \dot{l}^{\mathrm{T}} h^*, \dot{\theta}\rangle_{\dot{\Theta}}$$

for $h^* \in \mathscr{H}^*$, to be given by $\dot{l}^{\mathrm{T}} h^* = (h^*1, h^* - h^*1)$.

To determine if either $\psi_\theta(\theta, f) = \theta$ or $\psi_f(\theta, f) = f$ are differentiable functionals we must compute the relevant derivatives $\dot{\psi}$ and their adjoints. Begin with the functional $\psi_\theta(\theta, f) = \theta$. This has derivative $\dot{\psi}_\theta(a,g) = a$. In this case, the adjoint $\dot{\psi}^{\mathrm{T}}_\theta : \mathbb{R} \to \dot{\Theta}^*$ is given by $\dot{\psi}^{\mathrm{T}}_\theta b^* = (b^*, 0)$ for all $b^* \in \mathbb{R}$. Thus the range of $\dot{\psi}^{\mathrm{T}}_\theta$ is $(\mathbb{R} \times \{0\})$. That this is contained in the range of $\dot{l}^{\mathrm{T}}$ can be seen as follows. For any $x \in \mathbb{R}$ we can find an $h^*_x \in \mathscr{H}^*$ with $h^*_x 1 = x$ by taking $h^*_x(t) = x1_{[0,1]}(t)$. We also have $h^*_x - \int h^*_x = 0$ on all of $[0,1]$ so that, of course,

$$\int \left(h^*_x - \int h^* \,\mathrm{d}\lambda\right) g \,\mathrm{d}\lambda = 0 \quad \text{for } g \in C^0_b[0,1].$$

Thus $(\mathbb{R} \times \{0\}) \subset \mathscr{R}(\dot{l}^{\mathrm{T}})$ and we conclude from Theorem 4.1 that ψ is differentiable. To compute the lower bound in the convolution theorem we need the adjoint $\dot{v}^{\mathrm{T}} : B^* \to \mathscr{H}^*$ which, according to Theorem 4.1, is given by $\dot{\psi}^{\mathrm{T}} b^* = \dot{l}^{\mathrm{T}}\dot{v}^{\mathrm{T}} b^*$. Thus, take $\dot{v}^{\mathrm{T}} b^*$ to be $(\dot{v}^{\mathrm{T}} b^*)h = b^* \int h \,\mathrm{d}\lambda$. We might also consider $\dot{v}^{\mathrm{T}}$ as a map from B^* to $\mathscr{H}$ in which case we would say $\dot{v}^{\mathrm{T}} b^* = \dot{v}^{\mathrm{T}}_{b^*}$ where $\dot{v}^{\mathrm{T}}_{b^*}(t) = b^* 1_{[0,1]}(t)$ since then $\int \dot{v}^{\mathrm{T}}_{b^*}(t) h \,\mathrm{d}\lambda = b^* \int h \,\mathrm{d}\lambda$.

From these calculations and the conclusion of Theorem 2.4, it follows that

$$\mathrm{Cov}(b^*_1\mathbb{Z}, b^*_2\mathbb{Z}) = \langle \dot{v}^{\mathrm{T}}_{b^*_1}, \dot{v}^{\mathrm{T}}_{b^*_2}\rangle = b^*_1 b^*_2;$$

that is, the optimal limit random variable has variance 1. This lower bound is achieved by the sample mean $\bar X$.

If the parameter of interest is $\psi_f(\theta, f) = f$, another direct calculation shows that, for $b^* \in B^* = C_b^0[0,1]^*$, $\dot\psi^{\mathrm T}_f b^* = (0, b^*)$. Thus this operator has range $0 \times C_b^0[0,1]^*$. However, the form of $\dot l^{\mathrm T}$ is $\dot l_1^{\mathrm T} + \dot l_2^{\mathrm T}$ with $(\dot l^{\mathrm T} h^*)_2\, g = \int h^* g \,\mathrm d\lambda$. The range of $\dot l_2^{\mathrm T}$ is a strict subset of $C_b^0[0,1]^*$ since, for example, the evaluation maps $\pi_t(f) = f(t)$ are in $C_b^0[0,1]^*$ but cannot be represented as $\int h^* f \,\mathrm d\lambda$ for any h^*. We therefore conclude that $\psi_f(\theta, f) = f$ is not a differentiable function. This is to be expected since it is well known (Stone, 1982) that even if smoothness of f is imposed, say by assuming an infinite number of derivatives, it is not possible to obtain a convergence rate of $n^{-1/2}$ for estimating an element of $(C[0,1], ||\cdot||_\infty)$.

When we use the version of the problem where the nuisance sequence is generated by the $L_2^0(\lambda)$ function $\dot f_0$, the score operator $\dot l$ is specified by $\dot l(a, \dot g) = \dot g + a$, now considered as a map from $\mathbb R \times L_2^0(\lambda)$ to $L_2(\lambda)$. Calculation of the adjoint $\dot l^{\mathrm T}$ in this version of the model is as before and we find that $\dot l^{\mathrm T}$ is given by $\dot l_1^{\mathrm T} + \dot l_2^{\mathrm T}$ with $(\dot l^{\mathrm T} h^*)_1\, a = a \int h^* \,\mathrm d\lambda$ and $(\dot l^{\mathrm T} h^*)_2\, g = \int (h^* - \int h^* \,\mathrm d\lambda) g \,\mathrm d\lambda$.

For estimating $\psi_{\dot f_0}(\theta, \dot f_0) = \dot f_0$, the derivative $\dot\psi_{\dot f_0}$ is a map from $\mathbb R \times L_2^0(\lambda)$ to $L_2^0(\lambda)$. To compute this derivative, it is useful to back up to the definition of $\psi_{\dot f}$. Following Eq. (4.2), we define

$$v_n(P_{n,(\theta,\dot f)}) = \psi_n(\theta, \dot f) = \phi_n(\dot f),$$

where $\psi_n(\dot f)(x) = n \int_{(i-1)/n}^{i/n} \dot f \equiv \eta_{ni}(\dot f)$ if $x \in ((i-1)/n, i/n]$. This is reasonable because we can really only hope to estimate $\phi_n(\dot f)$ based on $X_{n1}, \ldots, X_{nn}$. It is only in the limit that $\psi_{\dot f}$ enters. Now, from (4.3), using the special choices $\mathbb R_n = \sqrt n$ and $c_n = n^{-1/2}$, the derivative is defined to be the continuous linear map such that

$$\sqrt n(\psi_n(\theta_{c_n}, \dot f_{c_n}) - \psi_n(\theta_0, \dot f_0)) - \dot\psi_{\dot f_0}(a, \dot g) \to 0. \tag{4.5}$$

If we let $\dot\psi_{\dot f_0}(a, \dot g) = \dot g$, the norm of the left-hand side of this expression is

$$\begin{aligned}
||\sqrt n(\phi_n(\dot f_{c_n}) - \phi_n(\dot f_0)) - \dot g||_{L_2(\lambda)} &= ||\sqrt n(\phi_n(\dot f_{c_n} - \dot f_0)) - \dot g||_{L_2(\lambda)} \\
&= ||\sqrt n(\phi_n(\dot g/\sqrt n) - \dot g||_{L_2(\lambda)} \\
&= ||\phi_n(\dot g) - \dot g||_{L_2(\lambda)} \to 0,
\end{aligned}$$

where, as noted previously, the convergence to 0 follows from standard results in L_2 approximation theory (Royden, 1988). The adjoint may then be computed to be $\dot\psi^{\mathrm T}_{\dot f_0} g^* = (0, g^*)$ for $g^* \in L_2^0(\lambda)^*$, so that $\mathscr R(\dot\psi^{\mathrm T}_{\dot f_0}) = (\{0\} \times L_2^0(\lambda))^*$. On the other hand, the range of $\dot l^{\mathrm T}$ is $(\mathbb R \times L_2^0(\lambda))^*$. To see this, note that every functional in $(\mathbb R \times L_2^0(\lambda))^*$ is based on an $x \in \mathbb R$ and an $f \in L_2^0(\lambda)$. Let $h^* = f + x$. For this choice, $(\dot l^{\mathrm T} h^*)_1 a = xa$ and $(\dot l^{\mathrm T} h^*)_2 g = \int fg \,\mathrm d\lambda$. We conclude that $\mathscr R(\dot\psi^{\mathrm T}_{\dot f_0}) \subset \mathscr R(\dot l^{\mathrm T})$ and hence by Theorem 4.1, $\psi_{\dot f_0}$ is differentiable. The lower bound from the convolution theorem can be calculated by first computing the adjoint $\dot v^{\mathrm T}$ given by the solution to $\dot\psi^{\mathrm T} g^* = \dot l^{\mathrm T} \dot v^{\mathrm T} g^*$ for all

$g^* \in L_2^0(\lambda)^*$. Taking $\dot{v}^{\mathrm{T}}$ to be the identity map from L_2^0 into L_2 gives

$$\dot{l}^{\mathrm{T}}\dot{v}^{\mathrm{T}}g^* = \dot{l}^{\mathrm{T}}g^* = (0,g^*) = \dot{\psi}^{\mathrm{T}}_{\dot{f}_0}g^*$$

as required. Thus the influence functions are simply $\dot{v}^{\mathrm{T}}_{g^*} = g^*$. From Theorem 2.4 we conclude that the optimal limit random variable would have covariance function given by

$$\mathrm{Cov}(g_1^*\mathbb{Z}, g_2^*\mathbb{Z}) = \langle \dot{v}^{\mathrm{T}}_{g_1^*}, \dot{v}^{\mathrm{T}}_{g_2^*}\rangle_{L_2(\lambda)} = \int g_1^* g_2^*\,\mathrm{d}\lambda.$$

Unfortunately, a continuous Gaussian process with these marginal distributions does not exist, since, via Sudakov's inequality (Van der Vaart and Wellner, 1996, Proposition A.2.5), the index set for any continuous process must be totally bounded. This is not the case for $L_2(\lambda)$. Furthermore, according to Proposition 5.5 of Millar (1982, p. 724) there are actually *no* $\sqrt{n}$-consistent estimators in this problem. This illustrates the fact that differentiability of a parameter alone does not guarantee that it can be estimated at $\sqrt{n}$ rate. Differentiability only allows calculation of the lower bound which asserts a minimum variance among the class of regular, $\sqrt{n}$-consistent estimators. It does not rule out the possibility that this class is empty.

As an alternative to the regularity as in Definition 2.3, we might instead define an estimator T_n of a B-valued parameter to be *weakly efficient* if

$$b^*(\sqrt{n}(T_n - \psi_n(\theta_0, \dot{f}_0))) \to_{\mathrm{d}} \mathrm{N}(0, ||\dot{v}^{\mathrm{T}}_{b^*}||^2)$$

for all $b^* \in B^*$. This is equivalent to the usual notion of efficiency if B is a finite-dimensional space. Such an estimator does exist in this problem, namely $\hat{\dot{f}}_{0_n}(x) = \sum_{i=1}^n X_{ni} 1_{((i-1)/n,i/n]}(x)$. To see why, first recall the notation

$$\phi_n(f)(x) = \sum_{i=1}^n \eta_{ni}(f)\ 1_{((i-1)/n,i/n]}(x) = \sum_{i=1}^n n\int_{(i-1)/n}^{i/n} f\,\mathrm{d}\lambda\, 1_{((i-1)/n,i/n]}(x).$$

Then for any bounded linear functional g^* and the function $g^* \in L_2^0(\lambda)$ that represents it, we have

$$\begin{aligned}
&g^*(\sqrt{n}(\hat{\dot{f}}_{0_n} - \phi_n(\dot{f}_0)))\\
&\quad= \sqrt{n}g^*(\hat{\dot{f}}_{0_n} - \phi_n(\dot{f}_0)) = \sqrt{n}\left(\int g^*\hat{\dot{f}}_{0_n}\,\mathrm{d}\lambda - \int g^*\phi_n(\dot{f}_0)\,\mathrm{d}\lambda\right)\\
&\quad= \sqrt{n}\left(\sum_{i=1}^n \frac{\eta_{ni}(g^*)}{n}(X_{ni} - \eta_{ni}(\dot{f}_0))\right)\\
&\quad= \sum_{i=1}^n \frac{\eta_{ni}(g^*)}{\sqrt{n}}(X_{ni} - \eta_{ni}(\dot{f}_0)) \overset{\mathrm{d}}{=} \mathrm{N}(0,\sigma_n^2),
\end{aligned}$$

where $\sigma_n^2 = \sum_{i=1}^n \eta_{ni}(g^*)^2/n$. But then

$$\begin{aligned}
\sigma_n^2 &= \sum_{i=1}^n \eta_{ni}(g^*)^2 \int_{(i-1)/n}^{i/n} 1\,\mathrm{d}\lambda = \sum_{i=1}^n \int_{(i-1)/n}^{i/n} \eta_{ni}(g^*)^2\,\mathrm{d}\lambda\\
&= \int \phi_n^2(g^*)\,\mathrm{d}\lambda = ||\phi_n(g^*)||^2 \to ||g^*||^2.
\end{aligned}$$

As a result,

$$g^*(\sqrt{n}(\hat{\dot{f}}_{0_n} - \phi_n(\dot{f}_0))) \to_{\mathrm{d}} \mathrm{N}(0, ||g^*||^2),$$

so that it is weakly efficient. Note however that $\hat{\dot{f}}_{0_n}$ is not even consistent for $\phi_n(\dot{f}_0)$ in $L_2(\lambda)$. We have

$$||\hat{\dot{f}}_{0_n} - \phi_n||^2 = \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} (\hat{\dot{f}}_{0_n} - \phi_n)^2 \,\mathrm{d}\lambda = \frac{1}{n} \sum_{i=1}^{n} (X_{ni} - \eta_{ni})^2 \to_{\mathrm{p}} 1$$

by the WLLN since $X_{ni} - \eta_{ni} \sim \mathrm{N}(0,1)$ so the squares are independent χ_1^2 with mean 1.

As in Van der Vaart (1991), the differentiability theorem may be specialized to the case of semiparametric models. The results and proofs are identical, but we include the following corollary for completeness. We use the standard semiparametrics notation, introduced earlier in this section, with a k-dimensional parameter of interest θ and infinite-dimensional nuisance parameter g, along with the corresponding score operators $\dot{l}_\theta$ and $\dot{l}_g$. We consider estimating $q(\theta)$ with $q: \mathbb{R}^k \to \mathbb{R}^m$ and write $\dot{q}(\theta)$ for the $m \times k$ derivative matrix. As before, the efficient score for θ is defined to be $l_\theta^* = \dot{l}_\theta - \Pi(\dot{l}_\theta | \bar{\dot{l}}_g(\dot{G}))$, the projection of $\dot{l}_\theta$ onto the orthocomplement of $\bar{\dot{l}}_g(\dot{G})$, and the efficient information matrix is $I^* = \langle l_\theta^*, l_\theta^{*'} \rangle_{\mathscr{H}}$. In this and the following we use the notation a' for vector transpose of a column vector a to avoid confusion with adjoints.

Corollary 4.3. *The sequence* $v_n(P_{n,\theta,g}) = q(\theta)$ *is differentiable if and only if* $\mathscr{N}(I^*) \subset \mathscr{N}(\dot{q}(\theta)^{\mathrm{T}})$.

Proof. First define $\psi_n(\theta, g) \equiv \psi(\theta, g) = q(\theta)$. Let $\dot{\theta} \in \mathbb{R}^k$ and $\dot{g} \in \dot{G}$, where $\dot{G}$ is the space of derivatives in the infinite-dimensional part of the parameter space and $\dot{g}$ corresponds to a path $\{g_t: t > 0\}$ with $g = g_0$. Then

$$\dot{\psi}(\dot{\theta}, \dot{g}) = \lim_{t \downarrow 0} (\psi(\theta + t\dot{\theta}, g_t) - \psi(\theta, g)) = \lim_{t \downarrow 0} (q(\theta + t\dot{\theta}, g_t) - q(\theta)) = \dot{q}(\theta)\dot{\theta}.$$

Thus we have, for any $b^* \in B^* = \mathbb{R}^m$, $\langle \dot{\psi}(\dot{\theta}, \dot{g}), b^* \rangle_B = \langle (\dot{\theta}, \dot{g}), \dot{\psi}^{\mathrm{T}} b^* \rangle_{\mathbb{R}^k \times \dot{G}}$ so that $\dot{\psi}^{\mathrm{T}} b^* = (b^{*'} \dot{q}(\theta), 0)$. In other words, $\mathscr{R}(\dot{\psi}^{\mathrm{T}}) = (\mathscr{R}(\dot{q}(\theta)^{\mathrm{T}}), \{0\})$.

The score operator $\dot{l}$ is given by $\dot{l}(\dot{\theta}, \dot{g}) = \dot{l}'_\theta \dot{\theta} + \dot{l}_g \dot{g}$ so that for any $h \in \mathscr{H}$,

$$\langle \dot{l}(\dot{\theta}, \dot{g}), h \rangle_{\mathscr{H}} = \langle \dot{l}'_\theta \dot{\theta} + \dot{l}_g \dot{g}, h \rangle_{\mathscr{H}} = \langle \dot{l}'_\theta \theta, h \rangle_{\mathscr{H}} + \langle \dot{l}_g \dot{g}, h \rangle_{\mathscr{H}} = \langle \dot{l}'_\theta, h \rangle_{\mathscr{H}} \dot{\theta} + \langle \dot{g}, \dot{l}_g^{\mathrm{T}} h \rangle_{\dot{G}}$$

and hence

$$\dot{l}^{\mathrm{T}} h = (\dot{l}_\theta^{\mathrm{T}} h, \dot{l}_g^{\mathrm{T}} h) = (\langle \dot{l}'_\theta, h \rangle_{\mathscr{H}}, \dot{l}_g^{\mathrm{T}} h).$$

Thus, according to Theorem 4.1, the parameter is differentiable if and only if

$$\mathscr{R}(\dot{q}(\theta)^{\mathrm{T}}) \subset \{\langle \dot{l}'_\theta, h \rangle_{\mathscr{H}}: h \in \mathscr{N}(\dot{l}_g) = \mathscr{R}(\dot{l}_g)^{\perp}\}.$$

Since $\mathscr{R}(\dot{l}_g)^{\perp}$ is spanned by the efficient score l^*_θ, any $h \in \mathscr{R}(\dot{l}_g)^{\perp}$ is of the form $h = l^{*'}_\theta a$, and consequently

$$\langle \dot{l}'_\theta, h \rangle_{\mathscr{H}} = \langle l^*_\theta, h \rangle_{\mathscr{H}} = \langle l^*_\theta, l^{*'}_\theta \rangle_{\mathscr{H}} = I^* a.$$

Thus the parameter is differentiable if and only if $\mathscr{R}(\dot{q}(\theta)^{\mathrm{T}}) \subset \mathscr{R}(I^*)$. Finally, because the two ranges in question are finite-dimensional and hence closed, this last statement is equivalent to the condition of the corollary. □

Note that for estimating all of θ, q is the identity, and the corollary states that θ is differentiable if and only if the efficient score matrix is non-singular. This was the condition used in the projection-of-scores approach earlier in this section.

In addition to the above specialization of Theorem 4.1 we have the following corollary giving sufficient conditions for equivalence of information bounds in different experiments. This can be used, for example, to show that the variance bounds are the same for the case-control example (Example 3) as in its randomized i.i.d. version.

Corollary 4.4. *Consider two different LAN experiments that involve the same parameter space and for which the goal is estimation of the same parameter in a Banach space B. Suppose the first experiment is LAN indexed by* $\mathscr{H}_1$ *and the second LAN indexed by* $\mathscr{H}_2$. *Let* $\dot{l}_1 : \dot{\Theta} \to \mathscr{H}_1$ *denote the score operator for the first experiment and* $\dot{l}_2 : \dot{\Theta} \to \mathscr{H}_2$ *the score operator for the second where* $\mathscr{R}(\dot{l}_1) = \mathscr{H}_1$ *and* $\mathscr{R}(\dot{l}_2) = \mathscr{H}_2$. *If the map* $\phi : \mathscr{H}_1 \to \mathscr{H}_2$ *defined by* $\dot{l}_2\dot{\theta} = \phi \dot{l}_1 \dot{\theta}$ *is a (Hilbert space) isomorphism then the parameter is differentiable in one experiment if and only if it is differentiable in the other. If the parameter is differentiable, the information bounds are the same in the two experiments.*

Proof. The function ψ of the parameter is hypothesized to be the same for both experiments and hence so is $\dot{\psi}^{\mathrm{T}}$. From the definition of ϕ and the fact that it is an isomorphism (which means that ϕ is a one-to-one mapping of $\mathscr{H}_1$ onto $\mathscr{H}_2$ which satisfies $\langle \phi g, \phi h \rangle_{\mathscr{H}_2} = \langle g, h \rangle_{\mathscr{H}_1}$ for all $g, h \in \mathscr{H}_1$; see e.g. Rudin (1966, p. 86) we find that

$$\langle \dot{\theta}, \dot{l}_1^{\mathrm{T}} h_1 \rangle_{\dot{\Theta}} = \langle \dot{l}_1 \dot{\theta}, h_1 \rangle_{\mathscr{H}_1} = \langle \phi^{-1} \dot{l}_2 \dot{\theta}, h_1 \rangle_{\mathscr{H}_1} = \langle \dot{l}_2 \dot{\theta}, \phi h_1 \rangle_{\mathscr{H}_2} = \langle \dot{\theta}, \dot{l}_2^{\mathrm{T}} \phi h_1 \rangle_{\dot{\Theta}}.$$

Thus $\dot{l}_1^{\mathrm{T}} = \dot{l}_2^{\mathrm{T}} \phi$ and similarly $\dot{l}_2^{\mathrm{T}} = \dot{l}_1^{\mathrm{T}} \phi^{-1}$. Suppose $\theta^* \in \mathscr{R}(\dot{l}_1^{\mathrm{T}})$. Then $\theta^* = \dot{l}_1^{\mathrm{T}} h_1$ for some $h_1 \in \mathscr{H}_1$. Setting $h_2 = \phi h_1$, we have

$$\dot{l}_1^{\mathrm{T}} h_1 = \dot{l}_1^{\mathrm{T}} \phi h_2 = \dot{l}_2^{\mathrm{T}} h_2$$

so that $\theta^* \in \mathscr{R}(\dot{l}_2^{\mathrm{T}})$. Hence $\mathscr{R}(\dot{l}_1^{\mathrm{T}}) \subset \mathscr{R}(\dot{l}_2^{\mathrm{T}})$. A similar argument shows the reverse inclusion so that $\mathscr{R}(\dot{l}_1^{\mathrm{T}}) = \mathscr{R}(\dot{l}_2^{\mathrm{T}})$. By Theorem 4.1 we conclude that a parameter will be differentiable in one model if and only if it is differentiable in the other.

Next let $\dot{v}_i : \mathscr{H}_i \to B$ and $\dot{v}_i^{\mathrm{T}} : B^* \to \mathscr{H}_i$, $i = 1, 2$ be the derivatives of the functionals to be estimated and their adjoints for the two experiments. That $\dot{v}_2^{\mathrm{T}} = \phi \circ \dot{v}_1^{\mathrm{T}}$ follows in a similar manner to the proof of uniqueness of $\dot{v}^{\mathrm{T}}$ in Theorem 4.1. In particular, we have from Theorem 4.1 that for any $b^* \in B^*$

$$\dot{\psi}^{\mathrm{T}} b^* = \dot{l}_1^{\mathrm{T}} \dot{v}_1^{\mathrm{T}} b^* = \dot{l}_2^{\mathrm{T}} \phi \dot{v}_1^{\mathrm{T}} b^* = \dot{l}_2^{\mathrm{T}} \dot{v}_2^{\mathrm{T}} b^*.$$

Thus,

$$\dot{v}_2^{\mathrm{T}} b^* - \phi \dot{v}_1^{\mathrm{T}} b^* \in \mathcal{N}(\dot{l}_2^{\mathrm{T}}) = \mathcal{R}(\dot{l}_2)^{\perp} = \mathcal{H}_2^{\perp}.$$

We conclude that $\dot{v}_2^{\mathrm{T}} b^* - \phi \dot{v}_1^{\mathrm{T}} b^* = 0$ in $\mathcal{H}_2$ for all $b^* \in B^*$ so that $\dot{v}_2^{\mathrm{T}} = \phi \dot{v}_1^{\mathrm{T}}$, or equivalently, $\dot{v}_1^{\mathrm{T}} = \phi^{-1} \dot{v}_2^{\mathrm{T}}$. Finally, using again the fact that ϕ is an isomorphism,

$$\langle \dot{v}_1^{\mathrm{T}} b_1^*, \dot{v}_1^{\mathrm{T}} b_2^* \rangle_{\mathcal{H}_1} = \langle \dot{v}_2^{\mathrm{T}} b_1^*, \dot{v}_2^{\mathrm{T}} b_2^* \rangle_{\mathcal{H}_2}$$

and the optimal limit $\mathbb{Z}_0$ has the same covariance function under either experiment. □

Example 3 (*Cont*)**.** In either the example or its randomized version the goal is still the estimation of the regression parameters β. Compare the score operator

$$\dot{l}_{\theta_0}(\xi, a)(\cdot, t) = (\dot{l}_1^{\mathrm{T}} \xi)(\cdot, t) + (\dot{l}_2(a))(\cdot, t) = \Phi_0(\xi, a) 1_{[0, \lambda_0]}(t) + \Phi_1(\xi, a) 1_{(\lambda_0, 1]}(t)$$

given in Section 3 to the score operator for the i.i.d. version specified in BKRW on p. 116 as $\dot{l}_{\text{i.i.d.}}(\xi, a)(\cdot, i) = \Phi_i(\xi, a)$. Label the former as experiment 1 and the second as experiment 2. Then it is easily verified that the map ϕ is an isomorphism since

$$\begin{aligned} \langle \dot{l}_{\theta_0}(\xi_1, a_1), \dot{l}_{\theta_0}(\xi_2, a_2) \rangle_{\mathcal{H}_1} &= \int \int_0^1 \dot{l}_{\theta_0}(\xi_1, a_1) \dot{l}_{\theta_0}(\xi_2, a_2) \, \mathrm{d}t \, \mathrm{d}P_{\theta_0} \\ &= \int \sum_{i=0}^{1} \lambda_i \Phi_i(\xi_1, a_1) \Phi_i(\xi_2, a_2) \, \mathrm{d}P_{\theta_0}, \end{aligned}$$

while

$$\langle \dot{l}_{\text{i.i.d.}}(\xi_1, a_1), \dot{l}_{\text{i.i.d.}}(\xi_2, a_2) \rangle_{\mathcal{H}_2} = \int \sum_{i=0}^{1} \lambda_i \Phi_i(\xi_1, a_1) \Phi_i(\xi_2, a_2) \, \mathrm{d}P_{\theta_0}$$

as well.

Thus, the information bound is the same in Example 3 as in the i.i.d. version. As shown in Breslow and Wellner (1997) logistic regression is an efficient estimator for the regression parameters in the latter and hence in the two sample version of Example 3 as well.

The above program has also been carried out for response-selective sampling designs; see Breslow et al. (1998).

5. Regularity and linearity of estimators

In this section, the goal is to develop the notions of regularity and linearity of estimators, and to connect these with the geometry of the tangent space. The first task is to establish an analog to Proposition 3.3.1 of BKRW, part of which states that an asymptotically linear and regular estimator with influence function in the tangent space is efficient. For this we need an analogous definition of linearity of an estimator. One possibility is to say an estimator T_n is asymptotically linear if for all $b^* \in B^*$,

$$\sqrt{n} b^*(T_n - v_n(P_{n,\theta_0})) = \Delta_{n,h_{b^*}} + o_P(1)$$

for some h_{b^*}. The problem here is that $\Delta_{n,h}$ is, in general, only defined on $\mathscr{H}$. Note that in the i.i.d. setting, the map $\Delta_{n,h}$ is defined to be

$$\Delta_{n,h} = n^{-1/2} \sum_{i=1}^{n} h(X_i)$$

for $h \in \dot{\mathscr{P}} \subset L_2(P_{\theta_0})$. In this case it is clear how to extend $\Delta_{n,h}$ to all of $L_2(P_{\theta_0})$ and we obtain a useful definition of asymptotic linearity. Without being able to extend $\Delta_{n,h}$ in some meaningful way, the resulting definition would declare an estimator to be asymptotically linear only if it had influence function in the tangent space, in which case asymptotically linear and regular would be a synonym for efficient. This is only the case in the i.i.d. theory when treating a fully nonparametric model.

There are several possibilities for extending $\Delta_{n,h}$. For instance, following Bickel (1993, p. 67) one could define a tangent space and maps $\Delta_{n,h}$ on a “largest possible” model of interest $\boldsymbol{M}$. Then the $\Delta_{n,h}$ described in this paper would be the restriction to the tangent space $\mathscr{H}$ of particular interest. This largest model corresponds to a fully nonparametric model in the i.i.d. theory.

Alternatively, we could proceed based on the results of Theorem 3.1 and in particular the condition (3.12) which states

$$\Delta_{n,h} \equiv \frac{1}{\sqrt{m_n}} \sum_{i=1}^{m_n} (h_{nk} - E_n[h_{nk} | X_{n1}, \dots, X_{n,k-1}])$$

is well defined and approximately linear in h; i.e., $\Delta_{n,a_1h_1+a_2h_2} = a_1\Delta_{h_1} + a_2\Delta_{h_2} + o_{P_0}(1)$. The array $\{h_{nk}\}$ above is required to approximate the array of conditional log-likelihood ratios $\{\alpha_{nk}\}$, but $\Delta_{n,h}$ remains perfectly well defined for arbitrary arrays, say subject to the constraint that each h_{nk} is square integrable. This is still a rather vague definition. To go one step further, the definition could be based on the particular form of tangent vectors h defined in Definition 3.5 based on building step functions

$$h_n(\cdot, t) = \sum_{k=1}^{n} h_{nk}(\cdot) 1_{\{[nt]=k\}} \quad \text{and requiring } ||h_n - h||_{L_2(P_{\theta_0} \times \lambda)} \to 0.$$

The tangent vectors constructed in this way are also in the spirit of Bickel (1992); see his Example 2.5.2, p. 68.

For now, the simplest approach appears to be to specify a largest model of interest $\boldsymbol{M}$. Corresponding to $\boldsymbol{M}$, let $\mathscr{H}_{\boldsymbol{M}}$ denote the tangent space and suppose $\Delta_{n,h}$ is well defined and linear on $\mathscr{H}_{\boldsymbol{M}}$. In what follows assume $\mathscr{H}_{\boldsymbol{M}}$ is closed.

Definition 5.1 (*Asymptotic linearity*). Let $v_n : \mathscr{P}_n \to B$ be a sequence of parameters taking values in a Banach space B and T_n be a sequence of estimators for v_n. We say

T_n is *asymptotically linear* if for all $b^* \in B^*$,

$$\sqrt{n} b^*(T_n - v_n(P_{n,\theta_0})) = \Delta_{n,h_{b^*}} + o_P(1)$$

for $h_{b^*} \in \mathscr{H}_{\boldsymbol{M}}$ and the associated array $\Delta_{n,h_{b^*}}$ as in the LAN definition (Definition 2.1) for the largest model of interest $\boldsymbol{M}$. In particular, if $v_n(P_{n,\theta})$ is an m-dimensional parameter ($m < \infty$), a sequence $\{T_n\}$ is asymptotically linear if

$$\sqrt{n}(T_n - v(P_{\theta_0})) = \Delta_{n,\tilde{h}} + o_P(1)$$

for $\tilde{h} = (\tilde{h}_1, \dots, \tilde{h}_m) \in \mathscr{H}^m_{\boldsymbol{M}}$ and the associated $\Delta_{n,\tilde{h}} = (\Delta_{n,\tilde{h}_1}, \dots, \Delta_{n,\tilde{h}_m})$.

Remark 5.2. Even in the i.i.d. theory there are several possible definitions of asymptotic linearity for estimators of a general parameter (all of which coincide when the parameter is finite-dimensional). See, for example, Definition 5.2.5 of BKRW (p. 180). The above choice corresponds to their *weakly* asymptotically linear.

The first proposition uses the specialized form of the asymptotic linearity definition for finite-dimensional parameters. The second is a result for the general case. In view of Remark 5.2, it seems helpful to have separate results for the finite-dimensional and general cases.

Proposition 5.3. *Suppose the model is LAN at a point θ_0 indexed by a subspace $\mathscr{H}$ of a Hilbert space and that the maps $\Delta_{n,h}$ of Definition 2.1 are linear in h. Let $v_n(P_{n,\theta})$ be a sequence of m-dimensional parameters that are pathwise differentiable with derivative represented by $\dot{v} \in \bar{\mathscr{H}}^m$. If $\{T_n\}$ is asymptotically linear in $\Delta_{n,\tilde{h}}$ then it is* regular *if and only if $(\tilde{h} - \dot{v}) \perp \mathscr{H}^m$. If $\tilde{h} \in \bar{\mathscr{H}}^m$ then $\{T_n\}$ is regular if and only if $\tilde{h} = \dot{v}$ in which case T_n is efficient.*

Proof. Let $P_{n,0} \equiv P_{n,\theta_0}$. For an $h \in \mathscr{H}$, consider its LAN sequence $\{\theta_n(h)\}$. Then by linearity of $\Delta_{n,h}$ in h on all of $\mathscr{H}_M$, LAN, and the Cramér–Wold device it follows that under $P_{n,0}$

$$\begin{pmatrix} \sqrt{n}(T_n - v_n(P_{n,0})) \\ \Lambda_n(\theta_n(h), \theta_0) \end{pmatrix} = \begin{pmatrix} \Delta_{n,\tilde{h}} + o_P(1) \\ \Delta_{n,h} - \frac{1}{2}\Sigma_{22} + o_P(1) \end{pmatrix} \rightarrow_{\mathrm{d}} \mathrm{N}\left(\begin{pmatrix} 0 \\ -\Sigma_{22}/2 \end{pmatrix}, \Sigma \right),$$

where $\Sigma = [\Sigma_{ij}]$, $\Sigma_{11} = [\langle \tilde{h}_i, \tilde{h}_j \rangle_{\mathscr{H}_M}]$, $\Sigma_{12} = \langle h, \tilde{h} \rangle_{\mathscr{H}_M}$, and $\Sigma_{22} = ||h||_{\mathscr{H}_M}$.

Then, by Le Cam's third lemma (see, for example, Van der Vaart and Wellner, 1996, pp. 404–405)

$$\sqrt{n}(T_n - v_n(P_{n,0})) \rightarrow_{\mathrm{d}} \mathrm{N}(\Sigma_{12}, \Sigma_{11})$$

under $P_{n,\theta_n(h)}$. Also,

$$\sqrt{n}(v_n(P_{n,\theta_n(h)}) - v_n(P_{n,0})) \rightarrow \dot{v}(h) = \langle h, \dot{v} \rangle_{\mathscr{H}_M} \quad \text{for } \dot{v} \in \bar{\mathscr{H}}^m \subset \bar{\mathscr{H}}^m_M.$$

Combining these two facts with

$$\sqrt{n}(T_n - v_n(P_{n,\theta_n(h)})) = \sqrt{n}(T_n - v_n(P_{n,0})) - \sqrt{n}(v_n(P_{n,\theta_n(h)}) - v_n(P_{n,0}))$$

implies that under $P_{n,\theta_n(h)}$

$$\sqrt{n}(T_n - \nu_n(P_{n,\theta_n(h)})) \to_{\mathrm{d}} \mathrm{N}(\Sigma_{12} - \dot{\nu}(h), \Sigma_{11}) = \mathrm{N}(\langle h, \tilde{h} - \dot{\nu}\rangle_{\mathscr{H}_M}, \Sigma_{11}).$$

Hence T_n is regular if and only if $\langle h, \tilde{h} - \dot{\nu}\rangle_{\mathscr{H}_M}$ is constant for all h. Since $0 \in \mathscr{H}$ it must be that this constant is 0, which implies that $(\tilde{h} - \dot{\nu}) \perp \mathscr{H}$ (coordinatewise). If, in addition, $\tilde{h} \in \bar{\mathscr{H}}^m$, then $\tilde{h} - \dot{\nu}$ is in $\bar{\mathscr{H}}^m$, while also being orthogonal to $\mathscr{H}$ and $\bar{\mathscr{H}}$. This implies that $\tilde{h} - \dot{\nu} = 0$; i.e. $\tilde{h} = \dot{\nu}$. Thus the asymptotic variance of $\sqrt{n}(T_n - \nu_n(P_{n,}))$ is that of $\Delta_{n,\dot{\nu}}$ which is $\langle \dot{\nu}^{\mathrm{T}}, \dot{\nu}\rangle$; the information bound. In other words, T_n is efficient. □

Proposition 5.4. *Suppose the model is LAN at a point θ_0 indexed by a subspace $\mathscr{H}$ of a Hilbert space and that the maps $\Delta_{n,h}$ of Definition* 2.1 *are linear in h. Let $\nu_n(P_{n,\theta})$ be a sequence of B-valued parameters that are pathwise differentiable with derivative $\dot{\nu}$ such that $b^*\dot{\nu}$ can be represented by a $\dot{\nu}_{b^*} \in \bar{\mathscr{H}}$ for all $b^* \in B^*$. If $\{T_n\}$ is asymptotically linear in the sense of Definition* 5.1, *with all $h_{b^*} \in \mathscr{H}$, then $\{b^*T_n\}$ is regular if and only if $(h_{b^*} - \dot{\nu}_{b^*}) \perp \mathscr{H}$. If, in addition, $\sqrt{n}(T_n - \nu_n(P_{n,\theta_n(h)}))$ converges weakly under $\{P_{n,\theta_n(h)}\}$ to a tight limit in B for each $\{\theta_n(h)\}$, then $\{T_n\}$ is regular.*

When $h_{b^*} \in \mathscr{H}$, then $\{b^*T_n\}$ is regular if and only if $h_{b^*} = \dot{\nu}_{b^*}$ in which case $\{b^*T_n\}$ is efficient for $b^*\nu_n(P_{n,0})$. Similarly, if $\sqrt{n}(T_n - \nu_n(P_{n,\theta_n(h)}))$ converges weakly under $\{P_{n,\theta_n(h)}\}$ to a tight limit in B for each $\{\theta_n(h)\}$, then $\{T_n\}$ is regular and efficient.

Proof. The assertions regarding regularity and efficiency of b^*T_n follow from Proposition 5.3. If b^*T_n is regular and efficient for all $b^* \in B$, then with the additional assumption of weak convergence of $\sqrt{n}(T_n - \nu_n(P_{n,\theta_n(h)}))$ under $\{P_{n,\theta_n(h)}\}$, conclude that this limit must be the same for all $\{\theta_n(h)\}$ so that $\{T_n\}$ is regular.

When b^*T_n is regular and efficient for all $b^* \in B$, then the additional assumption of weak convergence implies $\sqrt{n}(T_n - \nu_n(P_{n,\theta_n(h)}))$ has the same covariance function as the optimal limit random element. That is, T_n is regular and efficient. □

Example 2 (*Cont.*). In the bivariate three sample model, the natural way to obtain a largest possible model of interest would be to drop the restriction that the second sample is from the first margin (X) of the bivariate distribution $P_0 \equiv P_{01}$ and that the third sample is from the second margin (Y). The second and third samples could instead be drawn from measures P_{02} and P_{03} where $P_{0i} \ll P_{01}$, $i = 2, 3$. This larger model would lead to tangents of the form (g_1, g_2, g_3) in a space $\mathscr{H}_M$ with inner-product

$$\langle h, \tilde{h}\rangle_{\mathscr{H}_M} = \sum_{i=1}^{3} \lambda_i \langle g_i, \tilde{g}_i\rangle_{L_2(P_{01})},$$

where now g_2 need not be $E(g_1|X)$ and g_3 need not be $E(g_1|Y)$; these are the restrictions that hold true on the subspace $\mathscr{H}$ corresponding to the model of interest. It can be shown that the maps $\Delta_{n,h}$ on the larger space $\mathscr{H}_M$ are of the form

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{3} \sum_{i=1}^{n_j} g_j(X_{ji}, Y_{ji}).$$

The naive estimator of P_{01} just uses the n_1 observations from the first sample; i.e.

$$\hat{P}_{01} = \mathbb{P}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} \delta(X_{1i}, Y_{1i}).$$

For the purpose of estimation, P_{01} is considered to be an element of $\ell^\infty(\mathscr{F})$ where $\mathscr{F}$ is a given collection of P_{01}-square-integrable functions. Because this is a function space, it suffices to consider only $\pi_f \in \ell^\infty(\mathscr{F})^*$ for a given $f \in \mathscr{F}$, where $\pi_f(b) = b(f)$ for all $b \in B \equiv \ell^\infty(\mathscr{F})$. By definition $\mathbb{P}_{n_1}$ is asymptotically linear if for all π_f

$$\begin{aligned}\sqrt{n}(\pi_f\mathbb{P}_{n_1} - \pi_f P_{01}) = \sqrt{n}(\mathbb{P}_{n_1} f - P_{01}f) &= \sqrt{n}\frac{1}{n_1}\sum_{i=1}^{n_1}(f(X_{i1}, Y_{1i}) - P_{01}f)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n_1}\frac{n}{n_1}(f(X_{i1}, Y_{1i}) - P_{01}f) = \Delta_{n,h_{\pi_f}}\end{aligned}$$

for some $h_{\pi_f} \in \mathscr{H}_M$, at least up to an $o_P(1)$ term. This is true for $h_{\pi_f} = \lambda_1^{-1}(f - P_{01}f, 0, 0)$.

From Van der Vaart and Wellner's calculations, as described earlier in Section 4, the element of $\mathscr{H}$ that represents the derivative $\pi_f \dot{v}$ is

$$\dot{v}_{\pi_f} = \frac{1}{\lambda_1}\left(f - P_{01}f - a_f - b_f, \frac{\lambda_1}{\lambda_2}a_f, \frac{\lambda_1}{\lambda_3}b_f\right),$$

where a_f and b_f satisfy

$$E(\lambda_2 f^o(X, Y) - (\lambda_1 + \lambda_2)a_f(X) - \lambda_2 b_f(Y)|X = x) = 0$$

and

$$E(\lambda_3 f^o(X, Y) - \lambda_3 a_f(X) - (\lambda_1 + \lambda_3)b_f(Y)|Y = y) = 0$$

with $f^o(X, Y) = f(X, Y) - Ef(X, Y)$. Recall that these two conditions were derived from the fact that the second element of $\dot{v}_{\pi_f}$ must be the conditional expectation of the first given X and the third element must be the conditional expectation of the first given Y.

From the definition of $\dot{v}_{\pi_f}$ and the calculations in Eq. (4.1) we see that

$$\langle h_{\pi_f}, (g_1, g_2, g_3)\rangle_{\mathscr{H}_M} = \langle \dot{v}_{\pi_f}, (g_1, g_2, g_3)\rangle_{\mathscr{H}_M} \quad \text{or} \quad \langle h_{\pi_f} - \dot{v}_{\pi_f}, (g_1, g_2, g_3)\rangle_{\mathscr{H}_M} = 0$$

for any $(g_1, g_2, g_3) \in \mathscr{H}$. Hence $\pi_f\mathbb{P}_{n_1} = \mathbb{P}_{n_1}f$ is a regular (but not efficient) estimator of $\pi_f P_{01} = P_{01}f$. Finally, if we assume $\mathscr{F}$ is such that $\sup_{f\in\mathscr{F}}||P_{n,\theta_n}f^2|| = O(1)$, then Theorem 3.10.12 of Van der Vaart and Wellner (1996, p. 407) implies $\sqrt{n}(\mathbb{P}_{n_1} - P_{n,\theta_n})$ converges weakly under P_{n,θ_n}. Thus Proposition 5.4 implies that $\mathbb{P}_{n_1}$ is regular. (Note that Theorem 3.10.12 of Van der Vaart and Wellner (1996) may be used to show regularity of $\mathbb{P}_{n_1}$ without Proposition 5.4. However, the above calculation provides an illustration of the geometric content of this proposition.)

6. Discussion of open problems

Although several results used in application of convolution theorems for i.i.d. data have been extended to non-i.i.d. models, many others remain. For instance, a nice conclusion to Section 4 would be a version of Theorem 3 in Van der Vaart (1995) that provides sufficient conditions for efficiency of M-estimators. This remains to be done. Even a version of this theorem for efficiency of maximum likelihood as a special case would be useful. Others that remain to be extended include Theorem 2.1 from Van der Vaart (1991) stating that the existence of regular $n^{1/2}$ consistent estimators of a functional implies its differentiability. This also appears as Theorem 5.2.3 in BKRW.

Although the theory developed so far is enough to deal with all current examples, there are two additional extensions that come to mind that may allow treatment of additional examples that do not fit into this framework. The first would be to consider different rates. In the proof of Theorem 3.1 (specifically in Lemma A.3), the array $\{2(\sqrt{\alpha_{nk}}-1)\}$ is approximated by an array of the form $\{h_{nk}/\sqrt{n}\}$ and we impose conditions on this second array (conditions (B.I.S), (B.II), and (C) from Appendix A). Instead we could replace the rate $\sqrt{n}$ by an arbitrary rate c_n and consider approximation of $\{2(\sqrt{\alpha_{nk}}-1)\}$ by $\{c_n^{-1}h_{nk}\}$. Such an approach would be useful for the study of estimators of parameters for which information accumulates at rates different than n.

A second possible extension of the results would be to allow the second term in the stochastic expansion of the log likelihood ratios to remain random. This is a so-called locally asymptotically quadratic (LAQ) condition. An interesting class of models that fit into this framework are the cointegrated time-series models studied by, for example, Park and Phillips (1988) or Phillips (1991). Although the examples in these papers involve parametric models, one could easily imagine semiparametric analogs. Convolution theorems for finite-dimensional parameters under LAMN (locally asymptotically mixed normal), a special case of LAQ, have been given by Jeganathan (1982) and more recently a convolution theorem under general LAQ has been given by Van den Heuvel (1996). In addition, a convolution theorem under LAMN when there are possibly infinite-dimensional nuisance parameters has been given by Schick (1988). Extensions of these LAQ conditions, LAMN conditions and convolution theorems for more general parameter spaces remain open problems.

Appendix A. Proof of Theorem 3.1

The proof of Theorem 3.1 consists of a series of lemmas beginning with Lemma A.1 below. To simplify notation we replace m_n with n throughout, but the conclusions remain valid for general m_n. As noted in Section 3, the treatment follows Strasser (1985, Section 74; 1989) who considered arrays of independent but not necessarily identically distributed observations. Here we work in the more general setting which allows dependent observations. Throughout the development we assume condition (3.7), but we begin with weaker conditions than (3.8)–(3.12). The next result is taken from Greenwood and Shiryayev (1985) although we do not use their full level of generality. They consider the process (in t) $\sum_{k=0}^{[nt]}\log\alpha_{nk}$ and conditions for convergence to a Gaussian process with drift. We use the special case when only $t=1$ is of interest.

Greenwood and Shiryayev rely on the following conditions:

$\overset{P_0}{\rightarrow}$(I) $\sum_{k=1}^n E_n[(\sqrt{\alpha_{nk}}-1)^2 1_{\{|\sqrt{\alpha_{nk}}-1|\geqslant\varepsilon\}}|\mathscr{F}_{n(k-1)}] \overset{P_{n,\theta_0}}{\rightarrow} 0$ for all $\varepsilon > 0$,

(II) $\sum_{k=1}^n E_n[(\sqrt{\alpha_{nk}}-1)^2|\mathscr{F}_{n(k-1)}] \overset{P_{n,\theta_0}}{\rightarrow} \frac{\sigma^2}{4}$,

(III) $\sum_{k=1}^n (\sqrt{\alpha_{nk}}-1)^2 \overset{P_{n,\theta_0}}{\rightarrow} \frac{\sigma^2}{4}$.

Lemma A.1. *Assume* (3.7) *holds. Then the following three pairs of conditions are equivalent*:

1. (I, II),
2. (I, III),
3. (a) $\mathscr{L}(\Lambda(\theta_n,\theta_0)|P_{n,\theta_0}) \to \mathrm{N}(0,\sigma^2)) - \frac{1}{2}\sigma^2$,
   (b) $\mathscr{L}(\Lambda(\theta_n,\theta_0)|P_{n,\theta_n}) \to \mathrm{N}(0,\sigma^2) + \frac{1}{2}\sigma^2$.

Proof. With the addition of (3.7), this follows from Theorems 8 and 9 and Remark 5 in Greenwood and Shiryayev (1985). The additional condition seems necessary to make their proof work. Lemma B.4, which is used in the proof of Lemma A.2 below, gives a corrected version of the part of Greenwood and Shiryayev's argument that appears to be in error. □

Part (a) of the last assertion looks very much like Eqs. (2.1) and (2.3) of the LAN definition. Our next goal is to use the above result to identify the functions h and $\Delta_{n,h}$ and show that the latter is linear in h. Such expansions of the log likelihood ratio are well known in the case of independent observations. For the first step in this direction, we will need something a little stronger than (I), namely:

$$\text{(I.S)}\quad \sum_{k=1}^n E_n[(\sqrt{\alpha_{nk}}-1)^2 1_{\{|\sqrt{\alpha_{nk}}-1|\geqslant\varepsilon\}}] \to 0 \quad \text{for all } \varepsilon > 0.$$

Lemma A.2. *Assume* (3.7) *holds. Then under* (I.S) *and either* (II) *or* (III), *and with the notation* $g_{nk} = 2(\sqrt{\alpha_{nk}}-1)$, *the following expansion is valid*:

$$\Lambda_n(\theta_n,\theta_0) = \sum_{k=1}^n (g_{nk} - E_n[g_{nk}|\mathscr{F}_{n(k-1)}]) - \frac{1}{2}\sum_{k=1}^n E_n[g_{nk}^2|\mathscr{F}_{n(k-1)}] + r_n, \tag{A.1}$$

where $r_n \overset{P_{n,\theta_0}}{\rightarrow} 0$. *Furthermore*,

$$\mathscr{L}\left(\sum_{k=1}^n g_{nk} - E_n[g_{nk}|\mathscr{F}_{n(k-1)}]|P_{n,\theta_0}\right) \to \mathrm{N}(0,\sigma^2). \tag{A.2}$$

Proof. Recall that $\Lambda_n(\theta_n,\theta_0) = \sum_{k=1}^n \log\alpha_{nk}(\theta_n,\theta_0)$. Let

$$g_{nk}(\theta_n,\theta_0) = 2(\sqrt{\alpha_{nk}(\theta_n,\theta_0)}-1) \quad \text{so that } \log\alpha_{nk} = 2\log(\tfrac{1}{2}g_{nk}+1).$$

From a Taylor series expansion we have that $2\log(x/2+1) = x - \frac{1}{4}x^2 r(x)$ for $x > -2$ with $r(x)$ such that $r(0)=1$ and $|r(x_1)-r(x_2)| \leqslant C|x_1-x_2|$ for a constant C if $|x_i| < 1$,

$i = 1, 2$. In particular, we have $|1 - r(x)| \leqslant C|x|$ if $|x| < 1$. Thus,

$$\Lambda_n(\theta_n, \theta_0) = 2\sum_{k=1}^{n} \log\left(\frac{1}{2}g_{nk} + 1\right) = \sum_{k=1}^{n} g_{nk} - \frac{1}{4}\sum_{k=1}^{n} g_{nk}^2 r(g_{nk}).$$

This in turn can be written as

$$\sum_{k=1}^{n} g_{nk} - \frac{1}{4}\sum_{k=1}^{n} g_{nk}^2 + \frac{1}{4}\sum_{k=1}^{n} g_{nk}^2(1 - r(g_{nk})).$$

The goal now is to show the last term of the above converges to 0 in P_{n,θ_0}-probability. Since $|\sum_{k=1}^{n} g_{nk}^2(1-r(g_{nk}))| \leqslant \max_{1\leqslant k\leqslant n}|1-r(g_{nk})|\sum_{k=1}^{n} g_{nk}^2$ and $\sum_{k=1}^{n} g_{nk}^2 = \mathrm{O}_{\mathrm{P}_{n,\theta_0}}(1)$ by (III), it suffices, from the properties of the remainder term, to show $\max_{1\leqslant k\leqslant n}|g_{nk}| \overset{P_{n,\theta_0}}{\rightarrow} 0$. For this we have

$$P_{n,\theta_0}\left(\max_{1\leqslant k\leqslant n}|g_{nk}| > \varepsilon\right) \leqslant \sum_{k=1}^{n} P_{n,\theta_0}(|g_{nk}| > \varepsilon) \leqslant \sum_{k=1}^{n} \frac{1}{\varepsilon^2}E_n[g_{nk}^2 1_{\{|g_{nk}|>\varepsilon\}}] \rightarrow 0$$

for all $\varepsilon > 0$, where the asserted convergence follows from (I.S). Thus, we arrive at the expansion

$$\Lambda_n(\theta_n, \theta_0) = \sum_{k=1}^{n} g_{nk} - \frac{1}{4}\sum_{k=1}^{n} g_{nk}^2 + r_n \quad \text{where } r_n \overset{P_{n,\theta_0}}{\rightarrow} 0.$$

We also have, by the equivalence of (II) and (III) under (I) or (I.S), and the definition of g_{nk} that

$$\sum_{k=1}^{n} g_{nk}^2 - \sum_{k=1}^{n} E_n[g_{nk}^2|\mathscr{F}_{n(k-1)}] \overset{P_{n,\theta_0}}{\rightarrow} 0,$$

so we could also write

$$\Lambda_n(\theta_n, \theta_0) = \sum_{k=1}^{n} g_{nk} - \frac{1}{4}\sum_{k=1}^{n} E_n[g_{nk}^2|\mathscr{F}_{n(k-1)}] + r_n.$$

From the identity $2(\sqrt{\alpha} - 1) = (\alpha - 1) - (\sqrt{\alpha} - 1)^2$ we have

$$\begin{aligned}\sum_{k=1}^{n} E_n[g_{nk}|\mathscr{F}_{n(k-1)}] &= \sum_{k=1}^{n} E_n[(\alpha_{nk} - 1)|\mathscr{F}_{n(k-1)}] - \sum_{k=1}^{n} E_n[(\sqrt{\alpha_{nk}} - 1)^2|\mathscr{F}_{n(k-1)}] \\ &= \sum_{k=1}^{n} E_n[(\alpha_{nk} - 1)|\mathscr{F}_{n(k-1)}] - \frac{1}{4}\sum_{k=1}^{n} E_n[g_{nk}^2|\mathscr{F}_{n(k-1)}].\end{aligned}$$

The first term on the right-hand side of the above converges to 0 in P_{n,θ_0}-probability by Lemma B.4. Thus it is also true that

$$\Lambda_n(\theta_n, \theta_0) = \sum_{k=1}^{n}(g_{nk} - E_{P_n}[g_{nk}|\mathscr{F}_{n(k-1)}]) - \frac{1}{2}\sum_{k=1}^{n} E_n[g_{nk}^2|\mathscr{F}_{n(k-1)}] + r_n,$$

where $r_n \overset{P_{n,\theta_0}}{\rightarrow} 0$. Since $\sum_{k=1}^{n} E_n[g_{nk}^2|\mathscr{F}_{n(k-1)}] - \sigma^2 = \mathrm{o}_{\mathrm{P}_{n,\theta_0}}(1)$, Lemma A.1 implies that

$$\sum_{k=1}^{n}(g_{nk} - E_n[g_{nk}|\mathscr{F}_{n(k-1)}])$$

converges in distribution (under P_{n,θ_0}) to a standard normal with variance σ^2. □

One approach to verifying (I.S) and (II) is to find a different array which satisfies these conditions and also approximates the original array in a way that implies (I.S) and (II) hold for the original. This is the idea behind the next lemma.

Lemma A.3. *Suppose* (3.7) *holds and there exists an array* $\{h_{n1},\ldots,h_{nn}\}$, $n=1,2,\ldots$ *where* h_{nk} *is* $\mathscr{F}_{nk}$*-measurable for all* k *and* n, *such that*

(A) $\frac{1}{n}\sum_{k=1}^{n} E_n\left[\sqrt{n}(\sqrt{\alpha_{nk}}-1)-\frac{1}{2}h_{nk}\right]^2 \to 0,$

(B.I.S) $\frac{1}{n}\sum_{k=1}^{n} E_n[h_{nk}^2 1_{\{|h_{nk}|\geqslant\sqrt{n}\varepsilon\}}] \to 0$ *for all* $\varepsilon>0,$

(B.II) $\frac{1}{n}\sum_{k=1}^{n} E_n[h_{nk}^2|\mathscr{F}_{n(k-1)}] \overset{P_{n,\theta_0}}{\to} \sigma^2,$

(C) $\overline{\lim}_n \frac{1}{n}\sum_{k=1}^{n} E_n[h_{nk}^2] < \infty.$

Then

(a) *conditions* (B.I.S), (B.II) *and* (C) *are true with* h_{nk} *replaced by* $\sqrt{n}g_{nk}=\sqrt{n}2(\sqrt{\alpha_{nk}}-1)$. *In particular, expansion* (*A*.1) *of Lemma A*.2 *is valid.*

(b) *In* (A.1), g_{nk} *can be replaced with* $h_{nk}/\sqrt{n}$.

Proof. (a) Note that for a vector of functions $f=(f_1,\ldots,f_n)$, the function $\|f\| \equiv \{(1/n)\sum_{k=1}^{n} E_n[f_k^2]\}^{1/2}$ is indeed a (pseudo-) norm. Thus,

$$\left[\frac{1}{n}\sum_{k=1}^{n} ng_{nk}^2\right] \leqslant \left(\left\{E_n\left[\frac{1}{n}\sum_{k=1}^{n} h_{nk}^2\right]\right\}^{1/2} + \left\{E_n\left[\frac{1}{n}\sum_{k=1}^{n}(\sqrt{n}g_{nk}-h_{nk})^2\right]\right\}^{1/2}\right)^2.$$

Hence by (C) and (A), $\overline{\lim}_n E_n[(1/n)\sum_{k=1}^{n} ng_{nk}^2] < \infty$.

To show (I.S) holds, or equivalently, for any $\varepsilon>0$, $\sum_{k=1}^{n} E_n[g_{nk}^2 1_{\{|g_{kn}|\geqslant\varepsilon\}}] \to 0$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n} E_n[ng_{nk}^2 1_{\{|g_{nk}|\geqslant\varepsilon\}}]\\
&\quad\leqslant \left(\left\{\frac{1}{n}\sum_{k=1}^{n} E_n[(\sqrt{n}g_{nk}-h_{nk})^2 1_{\{|g_{kn}|\geqslant\varepsilon\}}]\right\}^{1/2} + \left\{\frac{1}{n}\sum_{k=1}^{n} E_n[h_{nk}^2 1_{\{|g_{nk}|\geqslant\varepsilon\}}]\right\}^{1/2}\right)^2\\
&\quad\leqslant \frac{2}{n}\sum_{k=1}^{n} E_n[\sqrt{n}g_{nk}-h_{nk}]^2 + \frac{2}{n}\sum_{k=1}^{n} E_n[h_{nk}^2 1_{\{|g_{nk}|\geqslant\varepsilon\}}]. \qquad (A.3)
\end{aligned}$$

The first term on the right-hand side of (A.3) converges to 0 by (A). Ignoring the factor of 2, the second term can be bounded above by

$$\frac{1}{n}\sum_{k=1}^{n} E_n[h_{nk}^2 1_{\{|h_{nk}|\geqslant\sqrt{n}\varepsilon/2\}}] + \sum_{k=1}^{n} E_n\left[\frac{h_{nk}^2}{n} 1_{\{|\sqrt{n}g_{nk}-h_{nk}|\geqslant\sqrt{n}\varepsilon/2\}}\right]. \qquad (A.4)$$

The first term of (A.4) also tends to 0 by (B.I.S). Then note that

$$\frac{h_{nk}^2}{n} 1_{\{|\sqrt{n}g_{nk}-h_{nk}|\geqslant\sqrt{n}\varepsilon/2\}} \leqslant \begin{cases} h_{nk}^2/n & \text{if } |h_{nk}|\geqslant\sqrt{n},\\ 1_{\{\sqrt{n}g_{nk}-h_{nk}|\geqslant\sqrt{n}\varepsilon/2\}} & \text{if } |h_{nk}|\leqslant\sqrt{n}.\end{cases}$$

Hence the second term of (A.4) is bounded by

$$\sum_{k=1}^{n} E_n\left[\frac{h_{nk}^2}{n} 1_{\{|h_{nk}|\geqslant\sqrt{n}\}}\right] + \sum_{k=1}^{n} E_n[1_{\{\sqrt{n}g_{nk}-h_{nk}|\geqslant\sqrt{n}\varepsilon/2\}} 1_{\{|h_{nk}|\geqslant\sqrt{n}\}}].$$

The first term of this expression also converges to 0, again by (B.I.S), while the second is less than or equal to

$$\sum_{k=1}^{n} E_n[1_{\{|\sqrt{n}g_{nk}-h_{nk}|\geqslant\sqrt{n}\varepsilon/2\}}] \leqslant \sum_{k=1}^{n} \frac{4}{n\varepsilon^2} E_n[(\sqrt{n}g_{nk}-h_{nk})^2] \to 0$$

by (A). Thus (I.S) is proved.

To obtain (II) we must show $(1/n)\sum_{k=1}^{n} E_n[ng_{nk}^2|\mathscr{F}_{n(k-1)}]$ converges in P_{n,θ_0}-probability to σ^2. This expression can be written as

$$\frac{1}{n}\sum_{k=1}^{n} E_n[ng_{nk}^2 - h_{nk}^2|\mathscr{F}_{n(k-1)}] + \sigma^2 + o_{\mathrm{P}_{n,\theta_0}}(1)$$

by (B.II). Thus it would suffice to show the first term converges in P_{n,θ_0}-probability to 0. In fact we can show the $L_1(P_{n,\theta_0})$ norms converge to 0 since,

$$\begin{aligned} E_n\left|\frac{1}{n}\sum_{k=1}^{n} E_n[ng_{nk}^2 - h_{nk}^2|\mathscr{F}_{n(k-1)}]\right| &\leqslant \frac{1}{n}\sum_{k=1}^{n} E_n\left|ng_{nk}^2 - h_{nk}^2\right| \\ &= \frac{1}{n}\sum_{k=1}^{n} E_n|(\sqrt{n}g_{nk}-h_{nk})(\sqrt{n}g_{nk}+h_{nk})| \\ &\leqslant \frac{1}{n}\sum_{k=1}^{n}\{E_n(\sqrt{n}g_{nk}-h_{nk})^2\}^{1/2}\{E_n(\sqrt{n}g_{nk}+h_{nk})^2\}^{1/2}, \end{aligned}$$

where the last inequality is due to the Cauchy–Schwarz inequality. Another application of Cauchy–Schwarz treating the summation as an integral bounds the above by

$$\left\{\frac{1}{n}\sum_{k=1}^{n} E_n(\sqrt{n}g_{nk}-h_{nk})^2\right\}^{1/2}\left\{\frac{1}{n}\sum_{k=1}^{n} E_n(\sqrt{n}g_{nk}+h_{nk})^2\right\}^{1/2}.$$

The first term converges to 0 by (A) so that we obtain the desired convergence if, for example, the second term is bounded. This is the case since

$$\left\{\frac{1}{n}\sum_{k=1}^{n} E_n(\sqrt{n}g_{nk}+h_{nk})^2\right\}^{1/2} \leqslant \left\{\frac{1}{n}\sum_{k=1}^{n} E_n(\sqrt{n}g_{nk})^2\right\}^{1/2} + \left\{\frac{1}{n}\sum_{k=1}^{n} E(h_{nk})^2\right\}^{1/2}.$$

Now condition (C) and the conclusion following Eq. (A.3) imply boundedness.

(b) The preceding argument also shows that

$$\sum_{k=1}^{n} E_n[g_{nk}^2|\mathscr{F}_{n(k-1)}] \quad \text{and} \quad \sum_{k=1}^{n} E_n[h_{nk}^2/n|\mathscr{F}_{n(k-1)}]$$

are equal up to an $o_{\mathrm{P}_{n,\theta_0}}(1)$ remainder term. Thus all that remains is to show g_{nk} can be replaced by $h_{nk}/\sqrt{n}$ in the first term of (A.1). Using Lemma B.3 we have

$$\begin{aligned} &E_n\left[\left(\sum_{k=1}^{n} g_{nk} - E_n[g_{nk}|\mathscr{F}_{n(k-1)}]\right) - \left(\sum_{k=1}^{n} h_{nk}/\sqrt{n} - E_n[h_{nk}/\sqrt{n}|\mathscr{F}_{n(k-1)}]\right)\right]^2 \\ &\quad = E_n\left[\sum_{k=1}^{n}(g_{nk} - h_{nk}/\sqrt{n} - E_n[g_{nk} - h_{nk}/\sqrt{n}|\mathscr{F}_{n(k-1)}])\right]^2 \\ &\quad \leqslant \sum_{k=1}^{n} E_n[g_{nk} - h_{nk}/\sqrt{n}]^2 \to 0 \end{aligned}$$

by (A), from which the result follows. □

Remark A.4. The proof of (a) was not specific to the array involving $\sqrt{n}g_{nk}$ and is more of a criteria for a certain type of equivalence of arrays. Similarly, the arguments in (b) can be used to show that given an array that satisfies the expansion of Lemma A.2 (g_{nk} in this case) the array involving terms $h_{nk}/\sqrt{n}$ also satisfies the expansion if (A) holds. In the case of independent data the conditional expectations are unconditional and we obtain the results in Strasser (1985, Section 74), upon which the above proof was based. See also Van der Vaart (1988, Proposition A.8).

The above results essentially prove Theorem 3.1. Condition (A) is (3.8), (B.I.S) is (3.9), (B.II) is (3.10), and (3.11) implies (C). In fact, (3.11) is only *required* for the approach to tangent vectors summarized in Definition 3.5, but it is a reasonable condition to impose in general. Among other things, it implies that (for sufficiently large n) each $h_{nk} \in L_2(P_{n,\theta_0}) \subset L_2(P_{\theta_0})$. Also note that when the data are independent, the conditional expectations in condition (B.II) are unconditional and we obtain (C) anyway. All that remains is the concept of a tangent vector h associated with the array $\{h_{nk}\}$ and this is provided by (3.8).

Appendix B. Some technical lemmas

Lemma B.1. *Let X be a topological vector space and let $X^{\mathbb{N}}$ be the product space of sequences of points in X with its product topology. If X^* separates points of X, $(X^{\mathbb{N}})^*$ separates points of $X^{\mathbb{N}}$.*

Proof. Given $x^* \in X^*$, the functional $x^*\pi_i$ is a continuous linear functional on $X^{\mathbb{N}}$ so is in $(X^{\mathbb{N}})^*$. If $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, then there exists j such that $x_{1j} \neq x_{2j}$ and there exists $x^* \in X^*$ such that $x^*x_{1j} \neq x^*x_{2j}$ since X^* separates points of X. Thus $x^*\pi_i\boldsymbol{x}_1 \neq x^*\pi_i\boldsymbol{x}_2$ so that $(X^{\mathbb{N}})^*$ separates points of $X^{\mathbb{N}}$. □

Lemma B.2. *Let ϕ be a map from a normed linear space X to a complete normed linear space Y such that $y' \circ \phi \in X^*$ for every y' in a closed subspace Y' of Y^* satisfying $\|y\| = \sup\{y'(y)\colon \|y'\| \leqslant 1\}$ for every $y \in Y$. Then ϕ is continuous and linear.*

Proof. See Van der Vaart (1991, Lemma A.2).

Lemma B.3. *Let $\mathscr{F}_1 \subset \cdots \subset \mathscr{F}_n$ be a sequence of σ-algebras and $X_1,\ldots,X_n$ be random variables such that each X_i is $\mathscr{F}_i$-measurable and $EX_i^2 < \infty$. Then*

$$E\left(\sum_{i=1}^{n} X_i - E(X_i|\mathscr{F}_{i-1})\right)^2 = \sum_{i=1}^{n} EX_i^2 - \sum_{i=1}^{n} E[E(X_i|\mathscr{F}_{i-1})]^2 \leqslant \sum_{i=1}^{n} EX_i^2.$$

Proof. This follows from standard martingale theory. □

Lemma B.4. *Let $\mathscr{F}_{nk}$ and $\alpha_{nk}=\alpha_{nk}(\theta_n,\theta_0)=$ be defined as in Section* 3. *Let E_n denote expectation under θ_0 and $\tilde{E}_n$ denote expectation under θ_n. If*

$$\sum_{i=1}^{n} \tilde{E}_n(1_{\{\alpha_{nk}=\infty\}}|\mathscr{F}_{n,k-1}) \overset{P_{n,\theta_0}}{\to} 0, \tag{B.1}$$

then

$$\sum_{k=1}^{n} E_n[(1-\alpha_{nk})|\mathscr{F}_{n,k-1}] \overset{P_{n,\theta_0}}{\to} 0.$$

Proof. Start by noting that

$$\alpha_{nk} = \alpha_{nk}1_{\{\alpha_{nk}<\infty\}} + \alpha_{nk}1_{\{\alpha_{nk}=\infty\}} = \alpha_{nk}1_{\{\alpha_{nk}<\infty\}} \quad P_{\theta_0}\text{-a.s.},$$

since the set $\{\alpha_{nk}=\infty\}$ has P_{θ_0}-probability 0. Thus,

$$\begin{aligned} E(\alpha_{nk}|\mathscr{F}_{n,k-1}) &= E(\alpha_{nk}1_{\{\alpha_{nk}<\infty\}}|\mathscr{F}_{n,k-1}) \\ &= 1_{\{1/\alpha_{n,k-1}<\infty\}}\tilde{E}\left(\frac{\alpha_{nk}}{\alpha_{nk}}1_{\{\alpha_{nk}<\infty\}}|\mathscr{F}_{n,k-1}\right) \\ &\quad +E(\alpha_{nk}1_{\{\alpha_{nk}<\infty\}}1_{\{1/\alpha_{nk}=\infty\}}|\mathscr{F}_{n,k-1}), \end{aligned} \tag{B.2}$$

where the last equality follows from Lemma 3 of Greenwood and Shiryayev (1985, pp. 17–18). The last term on the right-hand side of (B.2) can be taken to be 0 since

$$\alpha_{nk}1_{\{\alpha_{nk}<\infty\}}1_{\{1/\alpha_{nk}=\infty\}} = \alpha_{nk}1_{\{\alpha_{nk}=0\}} = 0.$$

In the first term of the right-hand side of (B.2) we note that

$$\alpha_{nk}/\alpha_{nk}1_{\{\alpha_{nk}<\infty\}} = 1_{\{0<\alpha_{nk}<\infty\}},$$

but the set $\{\alpha_{nk}=0\}$ has P_{θ_n}-probability 0 so that a version of (B.2) is

$$\begin{aligned} &1_{\{1/\alpha_{n,k-1}<\infty\}}\tilde{E}(1_{\{\alpha_{nk}<\infty\}}|\mathscr{F}_{n,k-1}) \\ &\quad =1_{\{1/\alpha_{n,k-1}<\infty\}} + 1_{\{1/\alpha_{n,k-1}<\infty\}}[\tilde{E}(1_{\{\alpha_{nk}<\infty\}}|\mathscr{F}_{n,k-1}) - 1] \\ &\quad =1_{\{1/\alpha_{n,k-1}<\infty\}} - 1_{\{1/\alpha_{n,k-1}<\infty\}}\tilde{E}(1_{\{\alpha_{nk}=\infty\}}|\mathscr{F}_{n,k-1}). \end{aligned}$$

From this it follows that

$$\begin{aligned} &\sum_{k=1}^{n} E_n[(1-\alpha_{nk})|\,\mathscr{F}_{n,k-1}] \\ &\quad = \sum_{k=1}^{n}\left\{1 - 1_{\{1/\alpha_{n,k-1}<\infty\}} + 1_{\{1/\alpha_{n,k-1}<\infty\}}\tilde{E}(1_{\{\alpha_{nk}=\infty\}}|\mathscr{F}_{n,k-1})\right\} \\ &\quad = \sum_{k=1}^{n} 1_{\{1/\alpha_{n,k-1}=\infty\}} + \sum_{k=1}^{n} 1_{\{1/\alpha_{n,k-1}<\infty\}}\tilde{E}(1_{\{\alpha_{nk}=\infty\}}|\mathscr{F}_{n,k-1}) \\ &\quad \leqslant \sum_{k=1}^{n} 1_{\{1/\alpha_{n,k-1}=\infty\}} + \sum_{k=1}^{n} \tilde{E}(1_{\{\alpha_{nk}=\infty\}}|\mathscr{F}_{n,k-1}). \end{aligned}$$

The first term on the right of this last display is Eq. (5.18), p. 100 of Greenwood and Shiryayev (1985). The proof that it converges in probability to 0 is a key part of the proof (pp. 100–103) of their Theorem 8. The second term converges in probability to 0 by assumption (B.1), whence the result.

7. Uncited References

Bickel and Ritov (1995); Blackwell (1951); Breslow and Holubkov (1997); Hájek (1970); Lawless et al. (1997); Le Cam (1986); Strasser, 1996; Strasser, 1998.

References

Bickel, P.J., 1993. Estimation in semiparametric models. In: Rao, C.R. (Ed.), Multivariate Analysis: Future Directions. North-Holland, Amsterdam.

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

Bickel, P.J., Ritov, Y., Wellner, J.A., 1991. Efficient estimation of linear functionals of a probability measure P with known marginal distributions. Ann. Statist. 19, 1316–1346.

Bickel, P.J., Ritov, Y., 1995. Inference in hidden Markov models I: local asymptotic normality in the stationary case. Bernoulli 2, 199–228.

Blackwell, D., 1951. Comparison of experiments. Proceedings of the Second Berkeley Symposium on Mathematics Statistics and Probability, pp. 93–102.

Breslow, N.E., Holubkov, R., 1997. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome dependent sampling. J. Roy. Statist. Soc. Ser. B 59, 447–461.

Breslow, N.E., McNeney, B., Wellner, J.A., 1998. Large sample theory for semiparametric regression models with two-phase outcome dependent sampling. Manuscript in progress. University of Washington, Seattle.

Breslow, N.E., Wellner, J.A., 1997. On the semiparametric efficiency of logistic regression under case-control sampling. Technical Report, Department of Statistics, University of Washington, Seattle.

Greenwood, P.E., Shiryayev, A.N., 1985. Contiguity and the Statistical Invariance Principle. Gordon and Breach, London.

Hájek, J., 1970. A characterization of limiting distributions of regular estimates. Z. Wahrscheinlichkeitstheorie verw. Geb. 14, 323–330.

Ibragimov, I.A., Khas'minskii, R.Z., 1975. Local Asymptotic normality for non-identically distributed observations. Thoeoret. Probab. Appl. 20, 246–260.

Jeganathan, P., 1982. On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. Sankhyā, Ser. A Part 1 44, 66–87.

Kelley, J.L., Namioka, I., 1963. Linear Topological Spaces. Van Nostrand, Princeton, NJ.

Koshevnik, Yu. A., Levit, Ya., 1976. On an non-parametric analogue of the information matrix. Theory Probab. Appl. 21, 738–753.

Lawless, J.F., Wild, C.J., Kalbfleisch, J.D., 1997. Semiparametric methods for response-selective and missing data problems in regression. J. Roy. Statist. Soc. Ser. B., submitted for publication.

Le Cam, L., 1960. Locally asymptotically normal families of distributions. Univ. California Publ. Statist. 3, 37–98.

Le Cam, L., 1969. Théorie Asymptotique de la Décision Statistique. University of Montréal Press, Montreal.

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

Le Cam, L., Yang, G.L., 1990. Asymptotics in Statistics. Some Basic Concepts. Springer, New York.

Millar, P.W., 1982. Optimal estimation of a general regression function. Ann. Statist. 10, 717–740.

Park, J.Y., Phillips, P.C.B., 1988. Statistical inference in regression with integrated processes: Part I. Econometric Theory 4, 468–498.

Phillips, P.C.B., 1991. Optimal inference in cointegrated systems. Econometrica 59, 283–306.

Pfanzagl, J., 1993. Incidental versus random nuisance parameters. Ann. Statist. 21, 1663–1691.

Royden, H.L., 1988. Real Analysis, 3rd Edition. Macmillan, New York.

Rudin, W., 1966. Real and Complex Analysis. McGraw-Hill, New York.

Rudin, W., 1991. Functional Analysis, 2nd Edition. McGraw-Hill, New York.

Schick, A., 1988. On estimation in LAMN families when there are nuisance parameters present. Sankhyā, Series A 50, 249–268.

Stone, C.J., 1982. Optimal global rates of convergence for nonparametric regression. Ann. Statist. 10, 1040–1053.

Strasser, H., 1985. Mathematical Theory of Statistics. De Gruyter, Berlin.

Strasser, H., 1989. Tangent vectors for models with independent but not identically distributed observations. Statistics and Decisions 7, 127–152.

Strasser, H., 1996. Asymptotic efficiency of estimates for models with incidental nuisance parameters. Ann. Statist. 24, 879–901.

Strasser, H., 1998. Perturbation invariant estimates and incidental nuisance parameters. Math. Methods Statist. 7, 1–26.

Van den Heuvel, E.R., 1996. Bounds for statistical estimation in semiparametric models. Ph.D. Thesis, University of Amsterdam.

Van der Vaart, A.W., 1988. Statistical Estimation in Large Parameter Spaces. CWI Tract, 44, Centrum voor Wiskunde en Informatica, Amsterdam.

Van der Vaart, A.W., 1991. On differentiable functionals. Ann. Statist. 19, 178–204.

Van der Vaart, A.W., 1995. Efficiency of infinite-dimensional M-estimators. Statist. Neerlandica 49, 9–30.

Van der Vaart, A.W., Wellner, J.A., 1991. Prohorov and continuous mapping theorems in the Hoffmann-Jørgensen weak convergence theory, with applications to convolution and asymptotic minimax theorems. Technical Report, University of Washington, Department of Statistics.

Van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes with Applications to Statistics. Springer, New York.