Statistica Sinica **11**(2001), 863-960

Celebrating the New Millennium: Editors' Invited Article

INFERENCE FOR SEMIPARAMETRIC MODELS: SOME QUESTIONS AND AN ANSWER

Peter J. Bickel and Jaimyoung Kwon

University of California, Berkeley

Abstract: Non-and semi-parametric models have flourished during the last twenty five years. They have been applied widely and their theoretical properties have been studied extensively. We briefly review some of their development and list a few questions that we would like to see addressed. We develop an answer to one of these questions by formulating a 'calculus' similar to that of the i.i.d. case that enables us to analyze the efficiency of procedures in general semiparametric models when a nonparametric model has been defined. Our approach is illustrated by applying it to regression models, counting process models in survival analysis and submodels of Markov chains, which traditionally require elaborate special arguments. In the examples, the calculus lets us easily read off the structure of efficient estimators and check if a candidate estimator is efficient.

Key words and phrases: Asymptotic, efficiency, Markov chains, semiparametric models, survival analysis.

1. Introduction

Statistical inference necessarily is based on statistical models for data. During most of the history of the subject, these have been parametric; the mechanism generating the data could be identified by specifying a few real parameters. Nonparametric models, ones in which "as little as possible" was assumed, were not seriously considered for several reasons, as follows.

(a) Fitting them to the data was computationally intractable.
(b) The parameters of parametric models such as centers, measures of spread, measures of dependence, were readily interpretable as opposed to the curve needed to specify the mechanism in the nonparametric case.
(c) The performance or approximate performance of methods such as estimation, confidence bounds and testing could be computed, and relative optimality (as measured, say, by the Fisher information matrix) could be understood more easily in parametric models than in nonparametric ones.

During the last twenty-five years nonparametric and semiparametric models have flourished. The main reason has, of course, been the rise of computing power

permitting the fitting of such models to large data sets showing the inadequacies of parametric models. The deficiency in interpretability of nonparametric models was filled by the development of semiparametric models, ones partly but not fully characterized by some interpretable Euclidean parameters. Examples include the Cox proportional hazards model in survival analysis, econometric index models, the more classical additive models with non-Gaussian errors and a plethora of others described, for instance, in Bickel, Klaassen, Ritov and Wellner (1993)[BKRW]. We refer to BKRW as a general reference throughout. This is no way to be interpreted as slighting other general works such as Ibragimov and Has'minskii (1981) and Pfanzagl (1980), or works covering more specialized subareas such as Andersen, Borgan, Gill, and Keiding (1995)[ABGK] or the slightly abstract but almost all-encompassing and seminal work of Le Cam (see Le Cam and Yang (1990)). We also note a more recent comprehensive treatment in the spirit of our approach by van der Vaart (2000).

The main focus of research in this area has been the construction of such models and corresponding statistical procedures in response to particular types of data arising in various disciplines, primarily biostatistics and econometrics. In particular, situations in which the data one ideally wishes to observe is necessarily partly hidden through censoring (right, left, double, interval, etc.), truncation and other forms of coarsening at random.

The focus has been mainly on the i.i.d. case (see BKRW for a review) or a counting process framework (see ABGK), although many important examples involving dependence may be found in Ibragimov and Has'minskii (1981). Some important features have been these.

(i) There has been the development of "likelihood" based fitting procedures such as maximum nonparametric likelihood, partial likelihood, profile likelihood, etc., as well as less efficient but often more robust methods based on what are referred to variously as generalized M or moment or estimating equations.

(ii) Algorithms have been constructed to implement these procedures, often including some elements of the E-M algorithms (Dempster, Laird, and Rubin (1977)).

(iii) Assessment of variability/confidence regions is possible using the generalized delta method, or the nonparametric (Efron) bootstrap (see Efron and Tibshirani (1993)). Foolproof bootstraps (m out of n) have been proposed by Politis and Romano (1994) (see also Bickel, Götze and van Zwet (1996) and Politis, Romano and Wolf (1999)).

Corresponding theory has established the asymptotic behavior of the estimates of (i) and the confidence regions of (iii), and provided occasional proofs of convergence for the algorithms of (ii).

(iv) Asymptotic optimality theory has been developed to parallel that in parametric models using the ideas of L. LeCam and C. Stein (see Ibragimov and Has'minskii (1981) and Pfanzagl (1982) for the earliest general statements of such results). The focus is on procedures which converge uniformly to the same limit over shrinking neighborhoods of a fixed member of the model. In the i.i.d. case, a geometrically based "calculus" has been developed enabling us to read off optimality and degree of suboptimality of estimation procedures proposed in semiparametric models (see BKRW).

There have, of course, been many developments in the construction of estimates of objects like densities and regression functions which cannot be estimated at the classic $n^{-1/2}$ rate but for which minimax rates of convergence over various large models can be established (Donoho, Johnstone, and Picard (1995) for instance). The relation between these results and the optimality theory for regularly estimable objects of the type we discuss is subtle (see Bickel and Ritov (2000a) for one point of view).

Semiparametric models and fitting procedures are increasingly proposed for a wide varieties of non-i.i.d. data types, including independent non-identically distributed observations, time series, spatial and spatio-temporal data. We single out some questions that we find interesting and important and which remain to be addressed in the i.i.d. case and/or have to be re-addressed in general. These are far from exhaustive and are proposed to spur discussion. It is only the third one (C) which we develop further in this paper.

(A) It is already apparent in the i.i.d. case (adaptive estimation, Bickel (1982)) and for bivariate censoring (van der Laan (1996), for example) that in many situations efficient estimates have to be based on intermediate estimates of an object such as a density or derivative of a density which cannot be estimated at the $n^{-1/2}$ rate over the model of interest. Even worse, there are models, e.g., so-called partial spline regression (Engle, Granger, Rice, and Weiss (1986)), where it may be impossible to estimate parameters of interest at any rate without estimation of "irregular" parameters as above. It is possible to take the position (see Robins and Ritov (1997) and Bickel and Ritov (2000b)) that one should only consider parameters for which there are estimates which converge uniformly over bounded neighborhoods and redefine the notion of efficiency.

Alternatively if one is willing to postulate models in which irregular parameters converge at specified rates, and most people are, and we can, in principle, estimate regular parameters efficiently, we are left with the question always present in estimation of irregular parameters: how do we select bandwidth or some other regularization parameter? Some initial discussion

is in Bickel and Ritov (2000b). Choices of procedures when plug-in is desired to produce both good minimax rates and efficiency for estimates of regular parameters is another area—see Bickel and Ritov (2000a).

(B) The theory of the nonparametric bootstrap has been extended to stationary time series by Hall (1995), Carlstein (1986), Künsch (1989), via various versions of the "blockwise" bootstrap (see also Politis, Romano and Wolf (1999)). Semiparametric versions have been proposed by Kreiss and Franke (1992), Bickel and Bühlmann (1999), Rajarshi (1990), Paparoditis and Politis (1997), and others. A "nonparametric" alternative is being studied by Kwon (2000). All of these approaches involve the choice of several "bandwidth" parameters. These are usually selected through some standard model selection criterion, often AIC. As in (A), what the appropriate selection criterion is when one is interested in (say) variance of estimates of Euclidean parameters which may or may not themselves be regularly estimable is unclear.

(C) If we leave the i.i.d. world and consider asymptotics for regression models, time series and the like, it is possible to establish efficiency of procedures for important special cases, e.g., regression models, diffusions (Ibragimov and Has'minskii (1981)), Markov chains (Greenwood and Wefelmeyer (1995)), time series models (Drost, Klaassen and Werker (1994)) and counting process models (ABGK). However, each case requires elaborate special arguments.

Bickel (1993) proposed an approach, based on ideas of Levit (1978), to a calculus similar to that available in the i.i.d. case for situations where we have the analogue of a largest "nonparametric" model. In Section 3 we develop this approach further, linking it to the existence of an efficient estimate for an appropriate representation or "parametrization" of the "nonparametric" model, and we show how it suggests connections between the i.i.d. and counting process formulation of models and some generalizations of estimating equations. Our examples are drawn from the i.i.d. case with a new representation and from non- and semiparametric Markov models. Extension to semiparametric stationary ergodic models is suggested, but evidently difficult.

(D) Semiparametric models have been written down in the i.i.d. case and, more generally, with little more justification than convenience and interpretability of parameters *if the models are valid*. There seems to have been little attention given to goodness-of-fit tests and diagnostics and/or these have

been proposed in an ad hoc manner, e.g., tests based on martingale residuals in survival analysis (see ABGK for instance). Of course, this is true even for goodness-of-fit of parametric models in the i.i.d. case. There has been a resurgence of interest in goodness-of-fit tests in the i.i.d. case (see, e.g., Rayner and Best (1989), Kallenberg and Ledwina (1995) and Fan (1996)). Bickel, Ritov and Stoker (1998) propose a unified framework for test construction in the i.i.d. case. Development and extension of these notions to the non-i.i.d. world seems worthwhile.

(E) Although there has been a huge amount of work on algorithms (MCMC) for Bayesian non- and semiparametric inference, the relations to frequentist inference are only poorly understood. There are negative results—see Freedman (1963, 1965, 1999) on consistency, and Cox(1993) on some higher order phenomena. But consistency rates and Bernstein-von Mises theorems, outside the nonparametric Dirichlet process domain for regular parameters, are only now being considered and much more is needed—see Wasserman (1998) and Ghosal, Ghosh, and van der Vaart (2000) and references therein.

In summary, we have indicated five areas in which we believe further theoretical understanding of inference in semiparametric models is needed. We develop some notions and results fully for one of these, (C), in Section 3, after introducing some general definitions and reviewing the i.i.d. case in Section 2. We leave consideration of other areas to the references and the future.

2. Basic Notions and Review of the i.i.d. Case

Disclaimer: We do not mention σ fields or questions of measurability in what follows, though these issues have to be taken into account by a careful treatment. We refer to van der Vaart and Wellner (1996), for example, for the additional assumptions, definitions and results needed to render our statements rigorous and, instead, focus on what we view as the essential conceptual issues.

In a very general sense we view a model as a set $\mathcal{P}$ of probabilities on the set $\mathcal{X}$ in which the total data $\mathbf{X}$, usually a vector of numbers coding for an object (e.g., a bit map, a time series, etc.) takes its values. We write $\mathbf{X} \sim P \in \mathcal{P}$.

A parametric model is one where the set $\mathcal{P}$ is smoothly describable as the image of a nice set Θ in R^d by a map $\theta \to P_\theta$, where smooth means (say) continuously differentiable in the Hellinger metric on $\mathcal{P}$. A nonparametric model is the set of all probabilities that we think the data could possibly have been generated under—or at least a dense subset of these. Since the set of all possible probability distributions on $\mathcal{X}$ is too large for discrimination on the basis of observation $\mathbf{X}$, the nonparametric models we consider are themselves subsets of

the set of all probability distributions, and parametrizable by $b \in \mathcal{M}$ which we take to be a subset of a Banach space $\mathcal{B}$. Semiparametric models are everything in between.

In all that follows we are interested in asymptotics, as $n \to \infty$, so that $\mathcal{X} = \mathcal{X}^{(n)}$ and $P = P^{(n)}$, where n is the sample size, the numbers of independent observations taken in a particular design, the length of time series observed, etc. However, the representing parameter space $\mathcal{M}$ and $\mathcal{B} \supset \mathcal{M}$ are fixed. A submodel $\mathcal{P}$ of the nonparametric model is then identified with a subset of $\mathcal{M}$. When we refer to elements of $\mathcal{P}$ as b, b corresponds to $P_b^{(n)}$.

We illustrate with the i.i.d. case. Here $\mathbf{X} = (X_1, \ldots, X_n)$ and the $X_i \in \mathcal{X}_1$ are i.i.d. $P^{(1)} \in \mathcal{P}^{(1)}$ so that $\mathcal{X} = \mathcal{X}^{(n)} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_1$, $\mathcal{P} = \{P^{(1)} \times \ldots \times P^{(1)} : P^{(1)} \in \mathcal{P}^{(1)}\}$. Nonparametric models are usually parametrized by $P^{(1)}$ viewed as an element of the Banach space of finite signed measures on $\mathcal{X}_1$ and $\mathcal{M}$ is the set of all probabilities on $\mathcal{X}_1$, or all absolutely continuous probabilities with respect to some σ-finite measure μ.

A parameter θ on $\mathcal{P}$ is a map from $\mathcal{P}$ to $\mathcal{T}$ which we take to be Euclidean or, more generally, Banach space. Thus, parameters to focus on are real, $\theta : \mathcal{P} \to R$, and $\mathcal{B}$-valued, $\theta : \mathcal{P} \to \mathcal{B}$, so $\theta(P_b^{(n)}) = b$.

The theory of efficient estimation in i.i.d. models can be thought of as follows. We identify b with $P^{(1)}$. Let $\hat{P}_n$ be the empirical distribution.

(A) If $\mathcal{P} = \mathcal{M} = \{\text{all } P^{(1)}\}$ then $\hat{P}_n$ is an efficient estimate of $P^{(1)}$, at least in the sense that for all (continuous, linear) $\theta(P^{(1)}) = \int \omega dP^{(1)}$, ω bounded, the linear estimate,

$$\theta(\hat{P}_n) = \int \omega d\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \omega(X_i)$$

is "asymptotically efficient".

(B) If $\theta(\cdot)$ is smoothly differentiable, i.e., locally approximable by a continuous linear parameter, the optimality of estimation holds for $\theta(\hat{P}_n)$ and is characterized by its local linear approximations.

(C) Optimality of $\theta(\hat{P}_n)$ extends to suitable Euclidean- or Banach-valued parameters $\theta(P)$ by viewing them as collections of real parameters.

(D) If $\mathcal{P}$ is a submodel of $\mathcal{M}$, a calculus is developed for geometrically characterizing a local linear approximation that an efficient estimate of θ must have.

In Section 3 we argue that, in general, given how to efficiently estimate $\theta(b) = b$ in $\mathcal{M}$, conclusions (B), (C) and particularly (D) extend quite generally.

We next make all these notions more precise in the i.i.d. case.

A *regular parametric one dimensional model (or curve) through* P_0 is defined as $\mathcal{Q} = \{P_t : |t| < 1\}$ where:

(i) $t \to P_t$ is $1-1$.

(ii) $P_t \ll \mu$ for a σ-finite μ for all t, and if $p(\cdot,t) \equiv dP_t/d\mu$ then the map $r_0 : (-1,1) \to L_2(P_0)$ defined by $r_0(t) = 2\left((p(\cdot,t)/p(\cdot,0))^{1/2} - 1\right) 1(p(\cdot,0) > 0)$ is continuously differentiable at 0, and its derivative $\dot{l}(\cdot,0) \in L_2(P_0)$ is nonzero.

The notation $\dot{l}$ reflects that, "pointwise", $\dot{l}(x,0) = \partial l/\partial t(x,0)$ where $l(x,t) = \log p(x,t)$ and $\|\dot{l}(X_1,0)\|_0^2$, the squared norm of $\dot{l}$ in $L_2(P_0)$, is the Fisher information.

The *tangent space* $\dot{\mathcal{P}}(P_0)$ at $P_0 \in \mathcal{P}$ is the closed linear span of $\dot{\mathcal{P}}^0(P_0) \equiv \{h \in L_2(P_0) : h = \dot{l}(\cdot,0)$ for some curve $\mathcal{Q} \subset \mathcal{P}\}$.

The tangent space of a nonparametric model is $L_2^0(P_0) := \{h \in L_2(P_0) : \int h dP_0 = 0\}$ itself. An *estimate* $\hat{\theta}_n$ of a real parameter $\theta(P)$ is *regular* at $P_0 \in \mathcal{P}$ if *for every curve* $\mathcal{Q}$ through P_0, $\mathcal{L}_{t_n}(\sqrt{n}(\hat{\theta}_n - \theta(P_{t_n}))) \to \mathcal{L}_0$. Here $\mathcal{L}_t$ denotes distribution under $P_t \in \mathcal{Q}$ and $|t_n| \le Mn^{-1/2}$ for some $M < \infty$ with $\mathcal{L}_0$ not dependent on $\{t_n\}$. Note that regularity is implied by uniform convergence of $\mathcal{L}_t(\sqrt{n}(\hat{\theta}_n - \theta(P_t)))$ on some compact neighborhood of θ. An *estimate* $\hat{\theta}_n$ of $\theta(P)$ real is *asymptotically linear* at P_0 if

$$\hat{\theta}_n = \theta(P_0) + \int \psi(x,P_0) d\hat{P}_n(x) + o_{P_0}(n^{-1/2}),$$

where $\psi(\cdot,P_0) \in L_2^0(P_0)$. Then $\psi(\cdot,P_0)$ is called the influence function of $\hat{\theta}_n$. It is necessarily unique. A regular asymptotically linear estimate $\hat{\theta}_n^*$ which minimizes the asymptotic variance of $\sqrt{n}\hat{\theta}_n$ (which is just $\int \psi^2(x,P_0) dP_0(x)$) among all regular asymptotically linear estimates at P_0 is called *efficient at* P_0. One way of framing the basic optimality theorem of estimation in semiparametric models in the i.i.d. case is as follows.

Theorem 1.

(a) *Suppose a regular asymptotically linear estimate* $\hat{\theta}_n^*$ *of* $\theta(P)$ *real exists, with influence function* $\psi^*(\cdot,P_0)$ *which is efficient at* P_0, *then* $\psi^* \in \dot{\mathcal{P}}(P_0)$. *Conversely, if a regular asymptotically linear estimate* $\hat{\theta}_n$ *has influence function* $\psi^*(\cdot,P_0)$ *belonging to* $\dot{\mathcal{P}}(P_0)$, *then* $\hat{\theta}_n$ *is efficient at* P_0.

(b) *If* ψ *is the influence function of any regular asymptotically linear estimate, then* $\psi^*(\cdot,P_0) = \Pi_0(\psi(\cdot,P_0) \mid \dot{\mathcal{P}}(P_0))$, *where* $\Pi_0(\cdot \mid \mathcal{L})$ *denotes projection in* $L_2^0(P_0)$ *on* $\mathcal{L} \subset L_2^0(P_0)$.

Note that, since $\dot{\mathcal{M}}(P_0) = L_2^0(P_0)$, it follows that if $\theta(P) = \int \omega dP$ is continuous linear then $\theta(\hat{P}_n)$ is regular asymptotically linear and hence efficient at all $P_0 \in \mathcal{M}$, the basic property of $\hat{P}_n$ described in (A).

These results are a subset of those obtained in Theorem 3.3.2 and Proposition 3.3.1 of BKRW, although we avoid the definition of pathwise differentiability of parameters and consideration of general regular estimates here.

More generally, $\hat{\theta}_n^*$ is more concentrated around $\theta(P_0)$ under P_0 than any other regular estimate (Hájek–LeCam Theorem), and has asymptotic minimax properties as well (see BKRW).

Result (a) enables us to quickly check if a candidate estimate is efficient by determining if its influence function belongs to $\dot{\mathcal{P}}(P_0)$, since characterizing "dense" subsets of $\dot{\mathcal{P}}(P_0)$ is usually simple.

Estimates of d-dimensional parameters $\boldsymbol{\theta}(\cdot)$, and even parameters $\theta(\cdot)$ taking values in a Banach space $\mathcal{T}$, can be viewed as the collection of estimates of all continuous linear functionals $b^*(\theta(\cdot))$, $b^* \in \mathcal{T}^*$, the dual space of $\mathcal{T}$. Now asymptotically linearity becomes a requirement for $b^*(\hat{\theta}_n)$, for all b^* in a large subset $\tilde{\mathcal{T}}^*$ of $\mathcal{T}^*$. Theorem 1 then applies directly. If $\{b^*(\hat{\theta}_n) : b^* \in \tilde{\mathcal{T}}^*\}$ possesses tightness properties as a stochastic process on $\tilde{\mathcal{T}}^*$, then efficiency of $\hat{\theta}_n^*$ also holds for plug-in estimates $q(\hat{\theta}_n^*)$ of certain real-valued $q(\theta(P))$, where q is non-linear. These issues are further discussed in BKRW, Chapters 5–7.

3. The Information Calculus

Our goal in this section is to develop a framework introduced in Bickel (1993) through which, once one has characterized the analogue of efficient influence functions in largest models (e.g., the nonparametric model for Markov chains), one can read off influence functions for submodels of interest geometrically as in the theorem. We illustrate these techniques in a variety of examples.

As in Section 2 suppose we have $X^{(n)} \sim P_b^{(n)}$, $b \in \mathcal{M} \subset \mathcal{B}$, a Banach space as our "nonparametric model". Let $\mathcal{B}^*$ be the dual space of $\mathcal{B}$, i.e., all continuous linear functionals b^* on $\mathcal{B}$ endowed with the operator norm. Let $r_n \downarrow 0$ at some rate ($n^{-1/2}$, in the examples we discuss).

We make the following assumptions, see also Levit (1978).

(A1) There is a sequence of estimates $\hat{b}^{(n)}$ of b in $\mathcal{M}$ such that, for all $b^* \in \mathcal{B}_0^* \subset \mathcal{B}^*$ where $\mathcal{B}_0^*$ is a linear space,

$$r_n^{-1}(b^*(\hat{b}^{(n)}) - b^*(b)) \Rightarrow \mathcal{N}(0, \sigma^2(b^*, b)) \tag{3.1}$$

under $P_b^{(n)}$.

(A2) For each $b \in \mathcal{M}$ there is a Hilbert space $\mathcal{H}_b$ with inner product $\langle \cdot, \cdot \rangle_b$, norm $\| \cdot \|_b$, and a mapping $T : \mathcal{B}_0^* \to \mathcal{H}_b$ with the following properties. If

$r_n^{-1}(b_n^*(\hat{b}^{(n)}) - b_n^*(b)) \Rightarrow \mathcal{N}(0, \tau^2)$ for a sequence $\{b_n^*\}$, then $\|Tb_n^* - h\|_b \to 0$ for some $h \in \mathcal{H}$ and $\tau^2 = \|h\|_b^2$. In particular, in (3.1), $\sigma^2(b^*, b) = \|h\|_b^2$ where $h = Tb^*$.

Definition 1. Let $\mathcal{Q} \equiv \{b_\eta \in \mathcal{M} : |\eta| < 1\}$ be a parametric submodel of $\mathcal{M}$ containing b_0. Then $\mathcal{Q}$ is *regular* at b_0 iff there exist $b_n^* \in \mathcal{B}_0^*$ and $\dot{l} \in \mathcal{H}_{b_0}$ such that $\|Tb_n^* - \dot{l}\|_{b_0} \to 0$, and if $t_n \to t$,

$$\log(dP^{(n)}_{b_{r_n t_n}}/dP^{(n)}_{b_0})(X^{(n)}) = tr_n^{-1}(b_n^*(\hat{b}^{(n)}) - b_n^*(b_0)) - \frac{t^2\|\dot{l}\|_{b_0}^2}{2} + o_p(1). \tag{3.2}$$

In view of (3.1), (3.2) implies that $\log(dP^{(n)}_{b_{r_n t_n}}/dP^{(n)}_{b_0})(X^{(n)}) \Rightarrow \mathcal{N}(-\frac{t^2}{2}\|\dot{l}\|_{b_0}^2, t^2\|\dot{l}\|_{b_0}^2)$ under $P_{b_0}^{(n)}$, which implies that the model $\{P_b^{(n)} : b \in \mathcal{Q}\}$ has the LAN property (LeCam and Yang (1990)).

Definition 2. For any submodel $\mathcal{P} \subset \mathcal{M}$, $b_0 \in \mathcal{M}$, let $\dot{\mathcal{P}}^0(b_0)(\subset \mathcal{H}_{b_0})$, *the tangent set of* $\mathcal{P}$ *at* b_0, be the set of all $\dot{l} \in \mathcal{H}_{b_0}$ corresponding to one-dimensional submodels of $\mathcal{P}$ regular at b_0. Let $\dot{\mathcal{P}}(b_0)$, *the tangent space of* $\mathcal{P}$ *at* b_0, be the linear closure in $\mathcal{H}_{b_0}$ of $\dot{\mathcal{P}}^0(b_0)$.

(A3) $\dot{\mathcal{M}}(b_0) = \mathcal{H}_{b_0}$ for all $b_0 \in \mathcal{M}$.

Definition 2, (A2) and (A3) characterize $\mathcal{M}$ as the "nonparametric model" corresponding to Hilbert space $\mathcal{H}_{b_0}$, as in Levit (1978).

Note that if $X^{(n)} = (X_1, \ldots, X_n)$, X_i i.i.d. P, if $\mathcal{X}$ is a complete separable metric space, and if we identify b with P viewed as an element of the Banach space of signed measures on $\mathcal{X}_1$, then we can make the following correspondences if $\mathcal{M}$ is the set of all probabilities on $\mathcal{X}_1$: $\hat{b}^{(n)} = \hat{P}_n$; $\mathcal{B}_0^* = \mathcal{B}^* = \{\text{Continuous bounded functions on } \mathcal{X}\}$; $b^*(b) = \int b^*(x)db(x)$; $\mathcal{H}_{b_0} = L_2^0(b_0)$. The mapping T is given by $Tb^* = b^* - E_{b_0}b^*$. Finally, the definition of tangent set and space given in Section 2 agree with the ones we have given more generally here provided only that we note that $\mathcal{H}_{b_0}$ is the closure of $T\mathcal{B}_0^*$.

We now can extend the notion of real parameter and regular asymptotically linear estimates directly. A real parameter θ on $\mathcal{P}$ just maps $\mathcal{P}$ to R. An estimate $\hat{\theta}_n(X^{(n)})$ of θ is *regular* at b_0 iff $\mathcal{L}_{b_{r_n t_n}}(r_n^{-1}(\hat{\theta}_n - \theta(b_{r_n t_n}))) \Rightarrow \mathcal{L}_0$ for some $\mathcal{L}_0$ independent of $\{t_n\}$, whenever $\{P_{b_\eta} : |\eta| < 1\}$ is a regular one-dimensional submodel of $\mathcal{P}$ and $|t_n| \le M < \infty$ for all n, some $M < \infty$. The estimate $\hat{\theta}_n$ is *asymptotically linear* at b_0 iff

$$r_n^{-1}(\hat{\theta}_n - \theta(b_0)) = r_n^{-1}b_n^*(\hat{b}^{(n)} - b_0) + o_p(1), \tag{3.3}$$

where $\|Tb_n^* - \psi(\cdot, b_0)\|_{b_0} \xrightarrow{P} 0$ for $\psi(\cdot, b_0) \in \mathcal{H}_{b_0}$ (under $P_{b_0}^{(n)}$). If $\hat{\theta}_n$ is asymptotically linear, it follows from (A1) and (A2) that $r_n^{-1}(\hat{\theta}_n - \theta(b_0)) \Rightarrow \mathcal{N}(0, \|\psi(\cdot, b_0)\|_{b_0}^2)$.

In complete agreement with Section 2 we call $\psi(\cdot, b_0)$ *the influence function or representer* of $\hat{\theta}_n$ at b_0.

Theorem 1 of Section 2 still applies since no special properties of the i.i.d. case are used. In particular, note that if $b^*(\hat{b})$ is a regular estimate of $b^*(b)$ on $\mathcal{M}$, then $b^*(\hat{b})$ is the asymptotically unique efficient estimate of $b^*(b)$. Even further, $\dot{l}$ of (3.2) can be called the "score function" since it is related to the efficient influence function $\psi(\cdot)$ for estimates of the parameter η of the model $\mathcal{Q}$ via the relation $\psi = \dot{l}/\|\dot{l}\|^2_{b_0}$. In fact, since the theory developed in Section 3.4 of BKRW that is based Section 3.3 there depends only on the geometry of $L_2^0(P_0)$, Theorem 3.4.1 of BKRW generalizes directly as well, as do the other definitions and propositions of that section. Extensions of Theorem 1 to vector and Banach-valued parameters that are proved in Chapters 3 and 5 of BKRW carry over as well, though formulation of regularity conditions undoubtedly requires work.

Finally, we note that the calculus also suggests what kind of representers we need to look for in estimates which are only *locally efficient* in the sense of Robins and Ritov (1995), i.e., $\sqrt{n}$-consistent at most P_b and efficient on a parametric subfamily (see also BKRW, especially Example 7.7.2). Specifically, given $\dot{l}_{10}(\theta)$, the score function of a parametric submodel $\mathcal{Q} = \{P_\theta : \theta \in \Theta \subset R^d\}$, we look for estimates with representer $l_0^*(\cdot, b) = \dot{l}_{10}(\theta) - \Pi_b(\dot{l}_{10}(\theta) \mid \dot{\mathcal{P}}_1(b))$ where $\dot{\mathcal{P}}_1(b)$ is the tangent space of the full model. We expect these estimates to be $\sqrt{n}$-consistent on $\mathcal{P}$ and efficient at all members of $\mathcal{Q}$ under suitable regularity conditions.

We now show that these notions apply to a number of interesting situations, and show that various estimates of important parameters in particular submodels are efficient.

Example 1. *The d sample problem.* Suppose $X^{(n)} = \{X_{ij} : 1 \le i \le n_j,\ j = 1, \ldots, d\}$ where $X_{ij} \in \mathcal{X}$ are i.i.d. $P^{(j)}$ for $j = 1, \ldots, d$, $\sum_{j=1}^d n_j = N$. Suppose $\hat{\pi}_j \equiv \frac{n_j}{N} \to \pi_j > 0$, $0 < \pi_j < 1$, $1 \le j \le d$. We can now take $b(P) = (P^{(1)}, \ldots, P^{(d)}) \in \mathcal{B}_1 \times \ldots \times \mathcal{B}_1$ where $\mathcal{B}_1 \equiv \{\text{Finite signed measures on } \mathcal{X}\}$. If $\mathcal{M} = \{b : P^{(j)} \ll \mu,\ 1 \le j \le d\}$, then $\hat{b} = (\hat{P}_n^{(1)}, \ldots, \hat{P}_n^{(d)})$ where $\hat{P}^{(j)}$ is the empirical measure of $\{X_{ij} : 1 \le i \le n_j\}$.

It is easy to see that if we take $r_N = N^{-1/2}$, $\mathcal{B}_0^* = \{\sum_{j=1}^d b_j^*(x_j) : b_j^* \text{ bounded continuous on } \mathcal{X},\ 1 \le j \le d\}$ and $\mathcal{H}_{b_0} = \{\sum_{j=1}^d h_j(x_j) : h_j \in L_2^0(P_0^{(j)})\}$ for $b_0 \leftrightarrow (P_0^{(1)}, \ldots, P_0^{(d)})$ and

$$\left\| \sum_{j=1}^d h_j \right\|^2 = \sum_{j=1}^d \pi_j \int h_j^2(x) dP_0^{(j)}(x), \tag{3.4}$$

then (A1)–(A3) hold.

The space $\mathcal{H}_{b_0}$ with norm (3.4) is the same as the tangent space at P_0 for the model (Z_i, Y_i), i.i.d. $P \in \mathcal{P}$, where $P[Z = z_j] = \pi_j$ known, $\mathcal{Z} = \{z_1, \ldots, z_d\}$, and $Y \mid Z = z_j$ has distribution $P^{(j)}$. Evidently tangent spaces of submodels also coincide given the correspondence between $(P^{(1)}, \ldots, P^{(d)})$ in the two models just given. We can, for instance, consider the regression submodel for this example:

$$X_{ij} = \theta_1 + \theta_2 z_j + \epsilon_{ij}, \tag{3.5}$$

where the $\epsilon_{ij} \in R$ are i.i.d. F with density f such that $I(F) = \int((f')^2/f)(x)dx < \infty$. As in Example 4.2.2 of BKRW, the tangent space at $b_0 \leftrightarrow (\theta_1, \theta_2, F)$ is

$$\dot{P}(b_0) = \Big\{ \sum_{j=1}^{d} h_j(x_j) : h_j(x_j) = (c_1 + c_2 z_j)\frac{f'}{f}(\epsilon_j), c_1, c_2 \in R \Big\}.$$

The MLE of θ_2 for f known is regular and has expansion

$$\hat{\theta}_2 = \theta_2 + \frac{1}{N}\sum_{j=1}^{d}\sum_{i=1}^{n_j} -\frac{(z_j - \bar{z}_N)}{\sigma_N^2}\frac{f'}{f}(\epsilon_{ij}) + o_p(N^{-1/2}),$$

where $\bar{z}_N = \sum_{j=1}^{d} \hat{\pi}_j z_j$, $\sigma_N^2 = \sum_{j=1}^{d}(z_j - \bar{z})^2$. It is efficient since its influence function belongs to the tangent space. Since Koul and Susarla's (1983) estimate is regular and has the same influence function, it is adaptive. There is nothing special about this argument and it is clear that treating the Z as fixed or random when their distribution is known, or is independent of the parameters governing the $P^{(j)}$, makes no difference in the construction of efficient estimates.

Next, we show how our framework relates to the counting process representations in the i.i.d. case which are of great importance in the formulation of models in survival analysis — see ABGK. Our approach can be viewed as, in part, putting results of Ritov and Wellner (1987) and Efron and Johnstone (1990) in our context. It is related to that of Greenwood and Wefelmeyer (1990) and Heyde (1988). The model we consider is less general than that of Greenwood and Wefelmeyer but permits us to link directly to Levit's very general Hilbert space representation. Heyde's approach corresponds to the treatment of Chapter 5 of BKRW generalized to the case of dependent variables, but deals with parameters defined by estimating equations rather than models and is tied to likelihood formulations. Our formulation enables us to read off efficiency properties which otherwise have to be shown by tediously going from elegant counting process representations to less elegant influence function (of the usual type) representations — See Example 3.4.2 of BKRW for instance.

Example 2. *Counting process representation for i.i.d. and d-sample models.* Let $(\mathbf{Z}_i, Y_i)$ be i.i.d. $P \in \mathcal{M}$, where we suppose $\mathbf{Z}$ has fixed distribution H_0

with finite support $\{\mathbf{z}_1, \ldots, \mathbf{z}_d\}$, and $Y \in R$ has conditional density $p(\cdot \mid \mathbf{z})$ with respect to Lebesgue measure and hazard rate $\lambda(\cdot \mid \mathbf{z}) \equiv p(\cdot \mid \mathbf{z})\bar{P}^{-1}(\cdot \mid \mathbf{z})$, where $\bar{P} \equiv 1 - P$, $P(y \mid \mathbf{z}) \equiv \int_{-\infty}^{y} p(t \mid \mathbf{z})dt$. We are interested in estimating parameters $(P^{(1)}, \ldots, P^{(d)})$ where $P^{(j)} \leftrightarrow p(\cdot|\mathbf{z}_j)$. To begin with, we suppose that $\pi_j = P[\mathbf{Z} = \mathbf{z}_j]$ are known and that $b := (\pi_1 P^{(1)}, \ldots, \pi_d P^{(d)})$ determines P. Not surprisingly, it will turn out that, as far as estimation relating to $(P^{(1)}, \ldots, P^{(d)})$ is concerned, knowing the π_j or not is irrelevant.

Let $\mathcal{B}$ be as in Example 1, and let $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_1$ be the space of d-tuples of finite signed measures on R with measures P of $\mathcal{M}$ corresponding to $(\pi_1 P^{(1)}, \ldots, \pi_d P^{(d)})$, rather than $(P^{(1)}, \ldots, P^{(d)})$ as in Example 1. Then, $\mathcal{B}^* = \{\sum_{j=1}^{d} b_j^* : b_j^* \in \mathcal{B}_1^*, b_j^*(b_j) = \int b_j^* db_j\}$, and b_j^* are bounded continuous. Moreover, $\hat{b} = (\pi_1 \hat{P}_n^{(1)}, \cdots, \pi_d \hat{P}_n^{(d)})$ where $\hat{P}_n^{(j)}$ is the conditional empirical distribution given $\mathbf{Z} = \mathbf{z}_j$, that is $\hat{P}_n^{(j)}(A) = \frac{1}{N_j} \sum_{i=1}^{n} 1(\mathbf{Z} = \mathbf{z}_j)1(Y_i \in A)$, $N_j = \sum_{i=1}^{n} 1(\mathbf{Z} = \mathbf{z}_j)$.

For $h \in L_2^0(P)$, define

$$R_j h(\cdot) = h(\cdot) - \bar{P}_j^{-1}(\cdot) \int_{\cdot}^{\infty} h(s) dP_j(s). \tag{3.6}$$

Represent

$$b^*(\hat{b}) - b^*(P) = \sum_{j=1}^{d} \pi_j \int \bar{b}_j^*(x) d\hat{P}_n^{(j)}(x), \tag{3.7}$$

where $\bar{b}_j^* = b_j^* - \int b_j^* dP_j$. Now

$$\int \bar{b}_j^* d\hat{P}_n^{(j)} = \int R_j(\bar{b}_j^*) d(\hat{P}_n^{(j)} - \hat{\Lambda}_n^{(j)}) \tag{3.8}$$

$$\hat{\Lambda}_n^{(j)}(u) = \sum_{i=1}^{n} 1(\mathbf{Z}_i = \mathbf{z}_j) \int_{-\infty}^{Y_i \wedge u} \lambda(t \mid \mathbf{z}_j) dt. \tag{3.9}$$

This is relation (3.15) of Ritov and Wellner (1988) and can readily be established by integration by parts. The process $\hat{\Lambda}_n^{(j)}$ is just the compensator of $n_j \hat{P}_n^{(j)}(\cdot)$. From (3.7), (3.8) and Martingale theory (Proposition 2.1(iv) of Ritov and Wellner (1988))

$$\sqrt{n}(b^*(\hat{b}) - b^*(P)) \Rightarrow \mathcal{N}\Big(0, \sum_{j=1}^{d} \pi_j \int R_j^2(\bar{b}_j^*) dP^{(j)}\Big). \tag{3.10}$$

Note that $R_j(b_j^*) = R_j(b_j^* + c)$.

We therefore can embed $\mathcal{B}_0^*$ into $L_2(P)$ via $Tb^* = \sum_{j=1}^{d} 1(\mathbf{Z} = \mathbf{z}_j)(R_j b_j^*)(Y)$. (A1) and (A2) follow from the i.i.d. case. Moreover, (A3) holds since $\sum_{j=1}^{d} 1(\mathbf{Z} = \mathbf{z}_j)(R_j b_j^*)(Y)$ is dense in $L_2^0(P)$. To see this let, for $h_j \in L_2(P^{(j)})$,

$$L_j h_j(\cdot) = h_j(\cdot) - \int_{-\infty}^{\cdot} h_j(y)\lambda(y \mid \mathbf{z}_j) dy. \tag{3.11}$$

By Ritov and Wellner (1988), L_j maps $L_2(P_j)$ onto $L_2^0(P_j)$. If $b_j^* = L_j h_j$, where h_j is continuous for $j = 1, \ldots, d$, then $b^* = \sum_{j=1}^d b_j^* \in \mathcal{B}_0^*$. Since the set of all such h_j is dense in $\mathcal{H}_b = L_2(P^{(j)})$, (A3) follows.

Lastly we note the following representation (Efron and Johnstone (1990)). If $\mathcal{Q}$ is a regular one-dimensional parametric submodel with conditional density $q(\cdot \mid \mathbf{z}, \theta)$ and conditional hazard rate $\lambda(\cdot \mid \mathbf{z}, \theta)$, then for $1 \le j \le d$, if $\lambda(\cdot \mid \mathbf{z}_j, \theta) \in L_2(P^{(j)})$,

$$L_j \frac{\partial}{\partial \theta} \log \lambda(\cdot \mid \mathbf{z}_j, \theta) = \frac{\partial}{\partial \theta} \log q(\cdot \mid \mathbf{z}_j, \theta). \tag{3.12}$$

Thus the score function in the classical sense, $\frac{\partial}{\partial \theta} \log q(\cdot \mid \mathbf{Z}, \theta) \in L_2^0(P)$, is now represented by $\frac{\partial}{\partial \theta} \log \lambda(\cdot \mid \mathbf{Z}, \theta) \in L_2(P)$. As we noted in Example 1, this analysis carries over directly to the d-sample model. Our approach from this point on evades intermediate calculations, such as in Ritov and Wellner ((1988), pp. 208-212). We note that the approach we have just given can carry over to the censored and truncated data case in many ways parallel to the presentation in Chapter 8 of ABGK. However, as in BKRW, our point of view makes calculations of efficiency more transparent.

We apply the representation to three examples.

(a) *Cox proportional hazard model.* Suppose $\mathcal{P} = \{P \in \mathcal{M} : \lambda(t \mid \mathbf{z}, \theta) = r(\mathbf{z}, \theta)\lambda(t), \lambda \text{ arbitrary}\}$. For simplicity take $\lambda = 1$. If λ is fixed, the "score function" for θ is

$$\dot{l}_1 = \frac{\partial}{\partial \theta} \log \lambda(t \mid \mathbf{z}, \theta) = \frac{\partial}{\partial \theta} \log r(\mathbf{z}, \theta). \tag{3.13}$$

On the other hand, the "tangent space" with respect to the nuisance parameter $\lambda(\cdot)$ is clearly $\{h(Y) : \int h^2 dP < \infty\}$. Therefore, by the generalization of Theorem 3.4.1 of BKRW, the "efficient score function" for θ is $l_1^* = \dot{l}_1 - E(\dot{l}_1 \mid Y)$. Specializing to the usual Cox case, $r(\mathbf{z}, \theta) = e^{\theta z}$, we obtain

$$l_1^* = \mathbf{Z} - E(\mathbf{Z} \mid Y). \tag{3.14}$$

This is in agreement with equation (4.15) in Ritov and Wellner (1988). Since $l_1^*/\|l_1^*\|^2$ is just the "influence function" of the Cox estimate we can conclude its efficiency without going through the tedious calculations needed to compute its influence function in the ordinary representation. (Of course, we still require $\mathbf{Z}$ to have finite support and $\partial\lambda/\partial\theta$ to be continuous, but we remove these contingencies below.)

(b) *The model of Nielsen, Linton and Bickel* (1998). Here $\lambda(t \mid \mathbf{z}, \theta) = \omega(\mathbf{z})\lambda_\theta(t)$ where ω is arbitrary. The argument given above yields the "efficient score function"

$$l_1^* = \frac{\partial}{\partial \theta} \log \lambda_\theta(Y) - E\left(\frac{\partial}{\partial \theta} \log \lambda_\theta(Y) \mid \mathbf{Z}\right), \tag{3.15}$$

and again efficiency of the type of estimate proposed by these authors follows. Admittedly Nielsen et al. assumed the distribution of $\mathbf{Z}$ absolutely continuous unknown, but again we dispense with this below.

(c) *A model of Chen and Wang* (2000). Consider the "accelerated hazard" model with $\lambda(y \mid \mathbf{z}, \theta) = \lambda_0(r(\mathbf{z}, \theta)y),\ r > 0$, θ and λ_0 arbitrary. The score function for λ_0 fixed is

$$\dot{l}_1 = [\log \lambda_0]'(r(\mathbf{Z}, \theta)Y)\frac{\partial}{\partial \theta} r(\mathbf{Z}, \theta)Y. \tag{3.16}$$

The tangent space for θ fixed is just $\{h(r(\mathbf{Z}, \theta)Y) : h(r(\mathbf{Z}, \theta)Y) \in L_2(P)\}$, and the "efficient score function" for θ is

$$l^* = [\log \lambda_0]'(r(\mathbf{Z}, \theta)Y)\left(\frac{\partial}{\partial \theta} r(\mathbf{Z}, \theta)Y - E\left(\frac{\partial}{\partial \theta} r(\mathbf{Z}, \theta)Y \mid r(\mathbf{Z}, \theta)Y\right)\right) \tag{3.17}$$

Chen and Wang construct a regular estimate on ad hoc grounds, but easily seen to be locally efficient at the Weibull family for $r(z, \theta) = e^{z\theta}$, $z = 0$ or 1, the two-sample model. In this case $\lambda_0(t) = (\alpha + 1)t^\alpha$, $t > 0, \alpha > -1$, and

$$\dot{l}_1(Z, Y, \theta) = \alpha Z, \tag{3.18}$$

$$l^*(Z, Y, \theta) = \alpha(Z - E(Z \mid Ye^{Z\theta})), \tag{3.19}$$

yielding asymptotic variance in the uncensored case of $\alpha^{-2}E\ \mathrm{Var}(Z - E(Z \mid Ye^{Z\theta}))$. The agreement with the expression in their Theorem 1 is not obvious, but follows from our discussion on estimating equations below, and their estimating equation (8).

Extensions

Admittedly, in all of these examples as originally dealt with, covariates were not limited to be finite-valued and often were time-varying. Time variation within a $(\mathbf{Z}, Y)$ observation clearly does not affect our general approach. Extension of the argument to arbitrary covariates is also possible. It is not necessary in the "nonparametric" case to efficiently estimate the function $\pi(\cdot)p(\cdot \mid \mathbf{z})$, but only linear functionals $\int b^*(\mathbf{z}, y)dP(\mathbf{z}, y)$ with $b^*(\cdot, \cdot) \in L_2(P)$ by $\int b^*(\mathbf{z}, y)dP_n(\mathbf{z}, y)$. These estimates have representers $L_{\mathbf{Z}}(b^*(\cdot, \mathbf{Z}))$ in $L_2(P)$, where $L_{\mathbf{Z}}$ is defined as in (3.12). In the same way, continuity of $\partial\lambda/\partial\theta$ is unnecessary. Only membership in $L_2(P)$ is needed. The rest of the development of tangent spaces and efficient score and influence functions is unchanged. Also note that the case of fixed covariates is covered by combining Examples 1 and 2.

Censoring, Truncation

We can more generally consider i.i.d. observations $X_i = (\mathbf{Z}_i, L_i, \Delta_i, Y_i),\ i = 1, \ldots, n$, where L_i corresponds to left truncation and Δ_i is the censoring indicator.

That is, C, T and L are independent given $\mathbf{Z}$ with $Y = \max\{L, T \wedge C\}$ and $\Delta = 1(T \leq C)$. Suppose the marginals of $\mathbf{Z}$ and L are fixed. If we now represent the data by the marked point processes

$$N_i(t) = 1(L_i \leq Y_i \leq \cdot) = \begin{cases} 1(L_i \leq T_i \leq \cdot), & \text{if } \Delta = 1 \\ 1(L_i \leq C_i \leq \cdot), & \text{if } \Delta = 0, \end{cases}$$

we are led via the usual argument to the representation of a one-dimensional regular model via

$$\dot{l}_1(\mathbf{Z}, L, \Delta, Y_1) = \Big[\Delta \frac{\partial}{\partial \theta} \log \lambda_1(Y \mid \mathbf{Z}, \theta) + (1 - \Delta) \frac{\partial}{\partial \theta} \log \lambda_0(Y \mid \mathbf{Z}, \theta)\Big] 1(Y \geq L), \tag{3.20}$$

where λ_0, λ_1 are the hazard rates of the censoring time C and lifetime T, respectively. Again, the nonparametric model corresponds to the possibility of arbitrary conditional hazard rates for T and C subject to truncation by L. Now $\dot{\mathcal{M}} = \{h(\mathbf{Z}, Y, \Delta) 1(Y \geq L) \in L_2(P)\}$ and the tangent space of the Cox model is just

$$\dot{\mathcal{P}} = \Big\{h(\mathbf{Z}, Y, \Delta) = \Big[\Delta\Big(\frac{\partial}{\partial \theta} \log r(\mathbf{Z}, \theta) + h_1(Y)\Big) + h_2(\mathbf{Z}, Y)\Big] 1(Y \geq L)\Big\},$$

where $h_1(Y)$ and $h_2(\mathbf{Z}, Y)$ range freely over members of $L_{\mathbf{Z}}(P)$ of the appropriate form. The efficient score function is just

$$l^* = \Delta\Big(\frac{\partial}{\partial \theta} r(\mathbf{Z}, Y, \theta) - E\Big(\frac{\partial}{\partial \theta} r(\mathbf{Z}, Y, \theta) \Big| Y\Big)\Big) 1(Y \geq L),$$

in agreement with Ritov and Wellner (1988), since the second term in (3.20) is projected out and it is easy to see that the first term has projection on the orthocomplement of $\dot{\mathcal{P}}$ as above.

Estimating Equations

As in the i.i.d. case the representer l^* suggests estimating equations but not linear ones. The compensator $\Lambda(\cdot \mid \mathbf{z})$ depends on P and can be estimated empirically if $\mathbf{Z}$ has finite support (and more generally — see ABGK Chapter 7). For instance, in the uncensored and untruncated finite valued $\mathbf{Z}$ case $d\hat{\Lambda}_n^{(j)}(t) = d\hat{P}_n^{(j)}(t)/\bar{\hat{P}}_n^{(j)}(t-)$ estimates $d\hat{\Lambda}_n^{(j)}(\cdot)$. We have, even ignoring the necessity of proving adequate behavior for the $\hat{\Lambda}_n^{(j)}$, the further problem of estimating the l^* which themselves in all our examples depend on conditional expectations. That is, say in the uncensored case, we look for estimates $\hat{l}^*(y, \mathbf{z}, \theta)$ of $l^*(y, \mathbf{z}, \theta, P)$ and then try to solve

$$\sum_{j=1} \pi_j \int \hat{l}^*(y, \mathbf{z}, \theta) d(\hat{P}_n^{(j)} - \hat{\Lambda}_n^{(j)})(y) = 0. \tag{3.21}$$

The dependence on π_j is illusory for parameters depending on $(P^{(1)}, \ldots, P^{(d)})$ only. In the Cox model, estimating $\hat{l}^*$ appears to require the estimation of $E(Z \mid Y)$. This is easier than expected once one notices

$$E(Z \mid Y = y) = \frac{E(r(Z,\theta)Z1(Y \geq y))}{E(r(Z,\theta)1(Y \geq y))}, \tag{3.22}$$

as in Ritov and Wellner ((1988), equation (4.15)). The natural (empirical) estimate of $E(Z \mid Y = y)$ is now just the Cox partial likelihood estimate. In the Nielsen, Linton, Bickel model, if $\mathbf{Z}$ is discrete, everything is readily estimable empirically. If $\mathbf{Z}$ is continuous, however, kernel estimation is required and we are led to the procedure suggested by Nielsen et al.

In the Chen–Wang model it appears that

(i) To adapt, obtain full efficiency always, one must estimate λ_0'/λ_0 and this requires density estimation.

(ii) Computing the conditional expectation given Y requires further density estimation.

This is unnecessary if (a) one is satisfied with local efficiency at λ_0, and (b) chooses $\lambda_0(t) = (\alpha+1)t^\alpha$ corresponding to a Weibull distribution, so that $\dot{l}_1(Z, Y, \theta)$ is a function of Z only.

This is the Chen–Wang approach which works because, as in the Cox model, for any function $\omega(\mathbf{Z}, \theta)$, $E(\omega(\mathbf{Z},\theta) \mid Yr(\mathbf{Z},\theta) = u) = \frac{\tilde{S}_1(u)}{\tilde{S}_0(u)}$ where, for $j = 0, 1$, $\tilde{S}_j(u) = E\left(\omega^j(\mathbf{Z},\theta)1(r(\mathbf{Z},\theta)Y \geq u)\right)$. Of course, proving that estimates based on such equations behave as we expect them to do requires extensive conditions as soon as one permits censoring and truncation—see Chen and Wang (2000), and more generally ABGK. However, the information calculus makes it clear which l^* or $\dot{l}_{10} - \Pi(\dot{l}_{10} \mid \dot{\mathcal{P}}_1)$ we need to estimate.

As is noted in BKRW and by Ritov and Wellner (1988), in connection with the Cox model, we can obtain locally efficient estimates in the following way. Start with the hazard rate representer of the score function of a "fixed shape" model, project on the orthocomplement of the tangent space in the representer space, and then solve an equation like (3.21). This principle, which we make explicit for a large class of examples, can be generalized. We do not pursue this here, but see the development in Chapter 6 of ABGK.

Example 3. *Real-valued Markov chains.* Suppose $X^{(n)} \equiv (X_1, \ldots, X_n) \sim P_b^{(n)}$, the n-dimensional marginal of $P_b^{(\infty)}$, the distribution of a real-valued stationary Markov chain with transition kernel corresponding to b, a probability measure on $R \times R$ with equal marginals, i.e., $(X_1, X_2) \sim b$. Take $\mathcal{B}$ to be the space of all signed measures on R^2 endowed with the total variation norm, and $\mathcal{M}$ to be the set of all b corresponding to a Markov chain obeying a Doeblin condition: for

some $r > 1$, the conditional distribution of X_r given X_1 is dominated by some σ-finite μ, and if $p_b^{(r)}(y \mid x)$ is the conditional density with respect to μ then

$$p_b^{(r)}(X_r \mid X_1) \geq \frac{d\nu}{d\mu}(X_r) \geq 0 \text{ a.s.} \tag{3.23}$$

for a measure ν such that $\nu(R) > 0$. The natural estimate of b is $\hat{b}^{(n)} \equiv \hat{P}_n^{(2)}$, the empirical distribution of (X_i, X_{i+1}), $1 \leq i \leq n-1$. Take $\mathcal{B}_0^* = \mathcal{B}^*$, represented as all continuous functions on R^2 via $b^*(b) = \int b^*(x_1, x_2) db(x_1, x_2)$. By (3.23) and the geometric φ mixing (see Doukhan (1994) for instance),

$$n^{1/2}(b^*(\hat{b}^{(n)}) - b^*(b)) = n^{-1/2} \sum_{i=1}^{n-1} (b^*(X_i, X_{i+1}) - b^*(b)) \Rightarrow \mathcal{N}(0, \sigma^2(b^*)),$$

where

$$\sigma^2(b^*) = \mathrm{Var}_b(b^*(X_1, X_2)) + 2 \sum_{k=1}^{\infty} \mathrm{Cov}_b(b^*(X_1, X_2), b^*(X_{k+1}, X_{k+2})). \tag{3.24}$$

Evidently, if

$$b_1^*(x_1, x_2) = b_2^*(x_1, x_2) + a(x_2) - a(x_1) + c \tag{3.25}$$

for some fixed c and a bounded a.s. (x_1, x_2), then $b_1^*(\hat{b}^{(n)})$ and $b_2^*(\hat{b}^{(n)})$ have the same limiting distribution. We thus, as in Bickel (1993) and Künsch (1984), define a mapping T on $L_2(P_0)$ by $Tb^*(x_1, x_2) = b^*(x_1, x_2) + a(x_2) - a(x_1)$ where

$$a(x) = \sum_{k=0}^{\infty} E_b(c^*(X_{k+1}) \mid X_1 = x), \tag{3.26}$$

$c^*(x) = E_b(\bar{b}^*(X_1, X_2) \mid X_1 = x)$ with $\bar{b}^* = b^* - Eb^*$.

Note that (3.23) implies that $a(x) = \sum_{k=0}^{\infty} (E_b(c^*(X_{k+1}) \mid X_1 = x) - E_b(c^*(X_1)))$ exists and has $|a(x)| \leq M \sum_{k=0}^{\infty} \varphi^k$, where M is the L_∞ norm of b^* and $|\varphi| < 1$ is the mixing coefficient. It is easy to verify that Tb^* satisfies

$$E_b(Tb^*(X_1, X_2) \mid X_1) = 0 \tag{3.27}$$

and, moreover, that

$$\sigma^2(Tb^*) = \mathrm{Var}_b(Tb^*(X_1, X_2)). \tag{3.28}$$

Take $\mathcal{H}_0 = \{h(X_1, X_2) : E_b(h^2(X_1, X_2)) < \infty,\ E_b(h(X_1, X_2) \mid X_1) = 0 \text{ a.s.}\}$ endowed with the $L_2(P_b^{(2)})$ norm. Clearly, T maps $L^2(P_0)$ into $\mathcal{H}_{b_0}$ and (A1) and (A2) hold for $\mathcal{B}_0^*$ as defined. Finally, note that if $h \in \mathcal{H}_0$, h continuous bounded and b is given, we can define a one-dimensional parameter submodel

$\mathcal{Q} \equiv \{Q_\eta : |\eta| < 1\}$ through P_b whose transition probability is dominated by that of $P_b^{(2)}$ and given by

$$\frac{dQ_\eta(x_2 \mid x_1)}{dP_b(x_2 \mid x_1)} = \exp\{\eta h(x_1, x_2) - A(\eta, x_1)\},$$

with $A(\eta, x_1)$ the appropriate normalizing constant. Evidently the Doeblin condition holds for each η and the μ corresponding to P_b. The model $\mathcal{Q}$ is also clearly regular with score function h. Since continuous bounded $h \in \mathcal{H}_0$ are dense in $\mathcal{H}_0$, (A3) follows.

The basic result of Greenwood and Wefelmeyer (1995) now follows: $b^*(\hat{P}_n^{(2)})$ is an efficient estimate of $b^*(P^{(2)})$ for $b^* \in \mathcal{B}_0^*$. More generally, plug-in works for parameters $\int h dP_b^{(2)}$, $b \in \mathcal{P} = \{b \in \mathcal{M} : \int h^2 dP_b^{(2)} < \infty\}$, and smooth parameters $v(P_b^{(2)})$ having linear derivatives $\int h dP_b^{(2)}$ in weak metrics on $\mathcal{M}$, such as Kolmogorov–Smirnov. On the other hand, it is clear that parameters such as $\int \psi(x_1, x_2, x_3) dP_b^{(3)}(\mathbf{x})$, ψ genuinely a nonlinear function of three variables, are not efficiently estimated by $\int \psi(x_1, x_2, x_3) d\hat{P}_n^{(3)}(\mathbf{x})$, since there is no function $h(X_1, X_2)$ with $h(X_1, X_2) \equiv \psi(X_1, X_2, X_3)$. Schick and Wefelmeyer (1999) used a sample splitting idea for construction of efficient estimators of such parameters, and Schick (2001) has recently shown how that idea may be extended to general semiparametric models. Kwon (2000) shows how to construct an estimator for the current situation which does not use an unrealistic sample splitting mechanism. The idea, suggested in Bickel (1993) and carried out by Kwon (2000), is to write

$$\theta(b) \equiv E_b \psi(X_1, X_2, X_3) = \int E_b\{\psi(x_1, X_2, X_3) \mid X_2 = x_2\} dP_b^{(2)}(x_1, x_2)$$

and use an empirical kernel or other estimate $\hat{p}(x_2 \mid x_1)$ of the transition density $p_b(x_2 \mid x_1)$ to estimate $E_b\{\psi(x_1, X_2, X_3) \mid X_2 = x_2\}$ by $\int \psi(x_1, x_2, x_3)\hat{p}(x_3 \mid x_2) dx_3$, ending with $\hat{\theta} = \frac{1}{n-1}\sum_{i=1}^{n-1} \int \psi(X_i, X_{i+1}, x_3)\hat{p}(x_3 \mid X_{i+1}) dx_3$. More generally (say) the lag k autocovariance $\theta_k \equiv \operatorname{cov}(X_1, X_{k+1})$, $k > 1$, when $EX_0 = 0$, can be estimated by $\hat{\theta}_k = \frac{1}{n}\sum_{i=1}^n X_i \int x\hat{p}^{(k)}(x|X_i) dx$, where $\hat{p}^{(k)}$ is defined recursively by

$$\hat{p}^{(j)}(z|x) = \int \hat{p}(z|y)\hat{p}^{(j-1)}(y|x) dy, \ j = 3, \ldots, k, \tag{3.29}$$

and $\hat{p}^{(2)}(\cdot|\cdot)$ is the kernel estimate of the transition density based on (X_i, X_{i+1}): $1 \le i \le n-1$. In practice, rather than computing $\hat{p}^{(k)}$, one would use a Markov bootstrap (Kwon (2000)).

Semiparametric submodels of $\mathcal{M}$

In this example we initially proceed more formally and permit chains which do not necessarily obey Doeblin conditions.

(a) The nonlinear autoregressive model

Let $\mathcal{P}$ be the submodel $X_{i+1} = g(X_i) + \epsilon_{i+1},\ 1 \le i \le n$, where g is unknown and ϵ_i are i.i.d. $E(\epsilon_1^2) < \infty$, $E(\epsilon_1) = 0$. Then $\{X_i, i = 1, 2, \ldots\}$ is a Markov chain. Suppose the ϵ_i are Gaussian. Then it is easy to see that

$$\dot{\mathcal{P}} = \{\epsilon_2 h(X_1) : h \in L_2(P^{(1)})\} \tag{3.30}$$

(see Section 4.3 of BKRW). Since these functions automatically satisfy (3.27), we only need estimates with "influence functions" in this space. In fact, it is easy to see that if we do not specify the distribution of the ϵ_i, but simply require that $E(\epsilon_1) = 0$, (3.30) still holds since then

$$\dot{\mathcal{P}} \supset \{h(X_1, X_2) : E(h(X_1, X_2) \mid X_1) = 0,\ E(h(X_1, X_2)\epsilon_2) = 0\} \oplus \left[\frac{f'}{f}(\epsilon_2)\right]$$

where [ ] denotes linear span.

It is evident that now the empirical distribution of (X_i, X_{i+1}), $1 \le i \le n-1$, does not yield influence functions in $\dot{\mathcal{P}}$. It is reasonable to conjecture, and not hard to verify (Kwon (2000)), that if g is sufficiently smooth then a "smoother" estimate $\hat{g}$ of $g(\cdot)$ based on the (X_i, X_{i+1}), $1 \le i \le n$, will yield efficient estimates of smooth parameters based on g only. For instance, if $\lambda(\cdot)$ is bounded and vanishes off compacts, $\int \hat{g}(t)\lambda(t)dt$ is an efficient estimate of $\int g(t)\lambda(t)dt$. If the density f of ϵ is known, if a stationary density p exists and there is appropriate mixing, one would expect that $\hat{p}$ solving $p(t) = \int f(t - \hat{g}(s))p(s)ds$, combined with $f(\cdot - \hat{g}(\cdot))$, can be expected to yield efficient estimates of smooth parameters of the joint distribution of (X_1, X_2). It is fairly straightforward, using the delta method, to argue that formally all of these estimates have influence functions belonging to $\dot{\mathcal{P}}$.

Note that, even though the Doeblin assumptions needed to justify the formulation does not hold in this instance, specifying a "least favorable" model is fairly routine once the form of the influence functions is identified. If the density of f is unknown and estimation of functionals such as $\int a(t)f(t)dt$ is of interest, $\dot{\mathcal{P}}$ now contains all $L_2(P^{(2)})$ functions $a(\epsilon_1)$. Then, if g were known, $n^{-1}\sum_{i=1}^{n-1} a(X_{i+1} - g(X_i))$ would have influence function $a(\epsilon_1)$ which belongs to the tangent space. Wefelmeyer (1994) shows that under suitable conditions, estimating ϵ_i by $\hat{\epsilon}_i = X_{i+1} - \hat{g}(X_i)$ yields an estimate that has the above influence function. Especially, one can use

$$\frac{1}{n}\sum_{i \in S_n} (\hat{g}(X_i) - g(X_i)) = o_P(n^{-1/2}) \tag{3.31}$$

for suitable $\hat{g}$, $S_n \subset \{1, \ldots, n\}$, e.g., using an initial stretch of length n^ϵ to estimate $\hat{g}$ and then summing only over X_i with $i \ge n^{2\epsilon}$.

Drost, Klaassen, and Werker (1994) show that in this example, and many others, if f is estimable then g can be estimated adaptively. Since this is a convex problem in the sense of Bickel (1982) it is clear that locally efficient estimates of g also exist.

(b) Fixing the stationary distribution

Kessler, Schick, and Wefelmeyer (2000) consider the possibly unrealistic model $\{P \in \mathcal{M} : P$ has a stationary distribution with density $p_0\}$ and give a "sample splitting" best efficient estimator. We derive the expected form of efficient influence functions which immediately suggests a natural estimate. Formally, $\dot{\mathcal{P}} = \{h(X_1, X_2) : E(h(X_1, X_2) \mid X_1) = 0, E(h(X_1, X_2)(b(X_2) - E(b(X_2) \mid X_1))) = 0$ for all $b \in L_2(P)\}$. This follows since the side condition on members $h \in \mathcal{H}_0$ required for membership in $\dot{\mathcal{P}}$ is $E(h(X_1, X_2)(a(X_1) + b(X_2))) = 0$ for all $a(X_1) + b(X_2) \in \mathcal{H}_0$. We claim that the projection of $h \in \mathcal{H}_0$ onto $\dot{\mathcal{P}}$ is $h - a_0(X_1) - b_0(X_2)$, where $a_0(X_1) + b_0(X_2)$ is the projection of h on the space S of *all* functions of the form $a(X_1) + b(X_2) \in L_2(P)$, called $\mathrm{ACE}(h \mid X_1, X_2)$ in BKRW, p. 440. This follows since ACE requires $0 = E(h(X_1, X_2) \mid X_1 = a_0(X_1) + E(b_0(X_2) \mid X_1))$. Now if $v(X_1, X_2)$ is bounded then, by our basic result, the influence function of the efficient estimate under $\mathcal{M}$ is $v(X_1, X_2) - q(X_1) + q(X_2)$, where q is of the form (3.26). But $v(X_1, X_2) - q(X_1) + q(X_2) - \mathrm{ACE}(v - q\pi_1 + q\pi_2 | X_1, X_2) = v - \mathrm{ACE}(v)$, where $\pi_j, j = 1, 2$ is the projection on the j'th coordinate. Therefore, we expect an efficient estimate of $\theta \equiv E(v(X_1, X_2))$ to be given by

$$\hat{\theta} \equiv \frac{1}{n-1} \sum_{i=1}^{n-1} (v(X_i, X_{i+1}) - \widehat{\mathrm{ACE}}(v \mid X_i, X_{i+1}) + \int \widehat{\mathrm{ACE}}(v | x_1, x_2) p_0(x_1) p_0(x_2) dx_1 dx_2), \tag{3.32}$$

where $\widehat{\mathrm{ACE}}$ is the projection of v on the sum space according to the (smoothed) empirical distribution of (X_i, X_{i+1}) (see Breiman and Friedman (1985)) and p_0 is the known stationary density. Note that this estimate is the natural analogue of the one for estimating a bivariate distribution subject to fixed marginals in BKRW.

Example 4. *The general stationary case.* Suppose $\mathcal{M}$ corresponds to the set of stationary geometrically φ-mixing probabilities on R^∞ and $b = (P^{(1)}, P^{(2)}, \ldots)$. Here the score and influence functions are at least formally approximable by functions $\psi(\ldots, x_{-1}, x_0)$ such that the linear approximations are

$$\hat{\theta} = \theta(P) + \frac{1}{n} \sum_{i=1}^{n} \psi(\mathbf{X}_{-\infty}^{i}) + o_p(n^{-1/2}), \tag{3.33}$$

where $\mathbf{X}^i_{-\infty} = (\ldots, X_0, X_1, \ldots, X_i)$, i.e., the i-shifted vector $\mathbf{X}^0_{-\infty}$, where

$$E_P(\psi(\mathbf{X}^i_{-\infty}) \mid \mathbf{X}^{i-1}_{-\infty}) = 0 \tag{3.34}$$

for all i. The natural space $\mathcal{H}_0$ is the subspace $\{\psi(\mathbf{X}^1_{-\infty}) : \psi \in L_2(P), E_P(\psi(\mathbf{X}^i_{-\infty}) \mid \mathbf{X}^{i-1}_{-\infty}) = 0,\ i \leq 1\}$. In this context the Markov model corresponds to the subspace where $\psi(\mathbf{X}^0_{-\infty}) = \psi(X_{-1}, X_0)$. It is possible to make this formalism more rigorous by, for instance, strengthening the definition of regularity to require uniform convergence on much larger sets. A conclusion is that, not very surprisingly, smooth functionals $v(P^{(k)})$ are efficiently estimated by $v(\hat{P}^{(k)}_n)$ or, going further such as in Ibragimov and Has'minskii (1981), that the usual empirical estimate of the spectral distribution function is efficient. However, studying what happens for, say, the semiparametric submodel obtained by adding Gaussian white noise to a nonparametric stationary sequence seems very difficult, and is left as a challenge.

4. Conclusion

An information calculus based on an appropriate definition of nonparametrics can be developed for semiparametric models in non-i.i.d. cases. In this paper we have discussed examples involving point processes representations in the i.i.d. case and stationary Markov processes. A full understanding of the utility of this calculus and extensions to the general stationary random field case, etc., remain open.

Acknowledgement

This research was partially supported by NSF Grant DMS9802960.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1995). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647-671.

Bickel, P. J. (1993). Estimation in semiparametric models. *Multivariate Analysis: Future Directions*, 55-73.

Bickel P. J. and Bühlmann, P. (1999). A new mixing notion and functional central limit theorems for a sieve bootstrap in time series. *Bernoulli* **5**, 413-446.

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

Bickel, P. J., Götze, F. and van Zwet, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statist. Sinica* **7**, 1-31.

Bickel, P. J. and Ritov, Y. (2000a). Nonparametric estimators which can be "plugged in". Technical Report, University of California, Berkeley.

Bickel, P. J. and Ritov, Y. (2000b). Non and semiparametric statistics: compared and contrasted. *J. Statist. Plann. Inference* **91**, 209-228.

Bickel, P. J., Ritov, Y. and Stoker, T. (1998). Testing and the method of sieves. Technical Report, University of California, Berkeley.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80**, 580-598.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14**, 1171-1179.

Chen Y. Q. and Wang M. C. (2000). Analysis of accelerated hazards models. *J. Amer. Statist. Assoc.* **95**, 608-618.

Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21**, 903-923.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-22.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B* **57**, 301-337.

Drost, F. C., Klaassen, C. A. J. and Werker, B. J. M. (1994). Adaptiveness in time series models. *Asymptotic Statistics : Proceedings of the Fifth Prague Symposium* **5**, 203-211.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Efron, B. and Johnstone, I. M. (1990). Fisher's information in terms of the hazard rate. *Ann. Statist.* **18**, 38-62.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81**, 310-320.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.* **91**, 674-688.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* **34**, 1386-1403.

Freedman, D. A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case, II. *Ann. Math. Statist.* **36**, 454-456.

Freedman, D. A. (1999). On the Bernstein-von Mises theorem with infinite dimensional parameters. *Ann. Statist.* **27**, 1119-1140.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, 500-531.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Greenwood, P. E. and Wefelmeyer, W. (1990). Efficiency of estimators for partially specified filtered models. *Stochastic Process. Appl.* **36**, 353-370.

Greenwood, P. E. and Wefelmeyer, W. (1995). Efficiency of empirical estimators for Markov chains. *Ann. Statist.* **23**, 132-143.

Hall, P. (1985). Resampling a coverage pattern. *Stochastic Process. Appl.* **20**, 231-246.

Heyde, C. C. (1988). Fixed sample and asymptotic optimality for classes of estimating functions. *Statist. Inference Stochastic Process.*, 241-247.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotical Theory*. Springer-Verlag, New York.

Jacobsen, M. and Keiding, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.* **23**, 774-786.

Kallenberg, W. C. M. and Ledwina, T. (1995). Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. *Ann. Statist.* **23**, 1594-1608.

Kessler, M., Schick, A. and Wefelmeyer, W. (2000). The information in the marginal law of a Markov chain. Submitted.

Koul, H. L. and Susarla, V. (1983). Adaptive estimation in linear regression. *Statist. Decisions* **1**, 379-400.

Kreiss, J. and Franke, J. (1992). Bootstrapping stationary autoregressive moving-average models, *J. Time Ser. Anal.* **13**, 297-317.

Künsch, H. R. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.* **12**, 843-863.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17**, 1217-1241.

Kwon, J. (2000). Efficiency calculus for general non and semiparametric models. Ph.D. dissertation, University of California, Berkeley.

Kwon, J. (2000). *Calculus of Statistical Efficiency in a General Setting; Kernel Plug-in Estimation for Markov Chains; Hidden Markov Modeling of Freeway Traffic.* Ph.D. Dissertation, Department of Statistics, University of California at Berkeley.
http://www.stat.berkeley.edu/users/kwon/index.html

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics. Some Basic Concepts.* Springer-Verlag, New York.

Levit, B. Y. (1978). Infinite-dimensional informational inequalities. *Theory Probab. Appl.* (Transl. of Teor. Verojatnost. i Primenen.) **23**, 371-377.

Nielsen, J. P., Linton, O. and Bickel, P. J. (1998). On a semiparametric survival model with flexible covariate effect. *Ann. Statist.* **26**, 215-241.

Paparoditis, E. and Politis, D. N. (1997). The local bootstrap for kernel estimators under general dependence conditions. Preprint.

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory.* Springer-Verlag, New York.

Politis, D. N. and Romano, J. P. (1994). Large-sample confidence-regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031-2050.

Politis D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling.* Springer, New York.

Rajarshi, M. B. (1990). Bootstrap in Markov-sequences based on estimates of transition density. *Ann. Inst. Statist. Math.* **42**, 253-268.

Rayner, J. C. W. and Best, D. J. (1989). *Smooth Tests of Goodness of Fit.* Oxford University Press.

Robins, J. M. and Ritov, Y. (1995). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statist. Medicine* **16**, 285-319.

Ritov, Y. and Wellner, J. A. (1988). Censoring, martingales and the Cox model. *Statist. Inference Stochastic Process.*, 191-219.

Schick, A. (2001). Sample splitting with Markov chains. *Bernoulli* **7**, 33-62.

Schick, A. and Wefelmeyer, W. (1999). Estimating joint distributions of Markov chains. Preprint.

Van der Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Statist.* **24**, 596-627.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer-Verlag, New York.

Van der Vaart, A. W. (2000). *Semiparametric Statistics.* In Lectures on Probability Theory, Ecole d'Ete de probabilites de St. Flour XXIX - 1999, P. Bernard, Ed. Springer, Berlin. To appear. (See http://www.cs.vu.nl/~aad/preprints/index-en.html.)

Wasserman, L. (1998). Asymptotic properties of semiparametric Bayesian procedures. In *Practical nonparametric and Semiparametric Bayesian Statistics* (Edited by D. Dey, P. Müller and D. Sinha), Springer-Verlag, New York.

Wefelmeyer, W. (1994). An efficient estimator for the expectation of a bounded function under the residual distribution of an autoregressive process. *Ann. Inst. Statist. Math.* **46**, 309 -315.

Department of Statistics, University of California, 367 Evans Hall, #3860, Berkeley, CA 94720-3860, U.S.A.

E-mail: bickel@stat.berkeley.edu

E-mail: kwon@stat.berkeley.edu

(Received October 2000; accepted July 2001)

COMMENTS

Jianqing Fan

Chinese University of Hong Kong and University of North Carolina

Bickel and Kwon are to be congratulated for this neat, insightful and stimulating paper on the general theory of semiparametric efficiency and for their successfully posing several important and challenge questions on semiparametric inferences. Semiparametric parametric models arise frequently in many applications. The interest in estimating certain principal parameters while imposing few assumptions on nuisance parameters gives rise to semiparametric models. The parameters of interest usually admit similar interpretations to those in parametric models. Most work focuses on efficient inferences on parameters of interest when semiparametric models are correctly specified. The question arises naturally how to validate whether a semiparametric model fits a given set of data, as asked by Bickel and Kwon. I welcome the opportunity to make a few comments and to provide additional insights.

1. Generalized Likelihood Ratio Test

One of the most celebrated methods in parametric inferences is the maximum likelihood ratio test. It is intuitive and easily applicable due to the Wilks type of results. An effort toward extending the scope of the likelihood ratio tests is empirical likelihood (Owen (1988)) and its various extensions. Yet, they cannot be directly applied to hypothesis testing problems in multivariate semiparametric and nonparametric models.

In an effort to derive a generally applicable testing procedure for multivariate nonparametric models, Fan, Zhang and Zhang (2001) propose a generalized

likelihood ratio test. The work is motivated by the fact that the nonparametric maximum likelihood ratio test may not exist. Further, even if it exists, it is not optimal even in the simplest nonparametric regression setting (see Fan et al. (2001)). Generalized likelihood ratio statistics, obtained by replacing unknown functions by reasonable nonparametric estimators, rather than MLE as in parametric models, enjoy several nice properties to be outlined below.

As an illustration, consider the varying-coefficient model

$$Y = a_1(U)X_1 + \cdots + a_p(U)X_p + \varepsilon, \tag{1}$$

where Y is the response variable, $(U, X_1, \cdots, X_p)$ is the covariate vector independent of the random noise ε. Consider the problem of testing homogeneity

$$H_0 : a_1(\cdot) = \theta_1, \cdots, a_p(\cdot) = \theta_p. \tag{2}$$

For simplicity, assume further $\varepsilon \sim N(0, \sigma^2)$ (As demonstrated in Fan et al. (2001), the normality assumption is only used to motivate the procedure). Given a random sample of size n, the likelihood under the null hypothesis can easily be obtained with parameters $\{\theta_j\}$ replaced by their MLE. Let $\ell_n(H_0)$ denote the log-likelihood under the null model. Under the more general model (1), the coefficient functions $a_1(\cdot), \cdots, a_p(\cdot)$ can easily be estimated by using, for example, a kernel method or local linear regression (Carroll, Ruppert and Welsh (1998), Hoover, Rice, Wu and Yang (1998), Fan and Zhang (1999)). Using these estimated functions, one can easily form the likelihood under the general model (1), though it does not maximize the nonparametric likelihood. Let $\ell_n(H_1, h)$ denote the log-likelihood, where h is the bandwidth used in the local linear regression estimate of functions $a_1(\cdot), \ldots, a_p(\cdot)$. Then, the generalized likelihood ratio statistic is simply

$$T_n(h) = \ell(H_1, h) - \ell(H_0). \tag{3}$$

This generalized likelihood ratio test admits the same intuitive interpretation as the classical likelihood ratio test.

Fan et al. (2001) unveil the following Wilks phenomenon: the asymptotic null distribution of $T_n(h)$ is independent of nuisance parameters in the model under the null hypothesis, and follows a χ^2-distribution (in a generalized sense) for testing homogeneity (2) versus (1). Thus, the P-values can easily be computed by either using the asymptotic distribution, or through simulations with parameter values taken to be the MLE under the null hypothesis. Further, they show that the resulting tests are asymptotically optimal in the sense of Ingster (1993).

The above Wilks phenomenon holds not only for testing parametric versus nonparametric hypotheses, but also for testing a nonparametric null hypothesis

versus a nonparametric alternative hypothesis. As an example, Fan et al. (2001) consider the problem of testing significance of variables

$$H_0 : a_1(\cdot) = a_2(\cdot) = \cdots = a_m(\cdot) = 0 (m \leq p).$$

The null hypothesis is still nonparametric because it involves nuisance functions $a_{m+1}(\cdot), \cdots, a_p(\cdot)$. Nevertheless, they show that the Wilks type of result continues to hold: the asymptotic null distribution is independent of these nuisance functions. Thus, the P-values can easily be computed by either using the asymptotic distributions or using simulations via fixing nuisance functions under the null hypothesis at their estimated values. These results are also extended to various other models.

The idea of the above generalized likelihood ratio method is widely applicable. It is easy to use because of the Wilks phenomenon, and is powerful as it achieves the optimal rates for hypothesis testing. This encourages me to propose the generalized likelihood ratio test as a possible tool to the open question (D) posed by Bickel and Kwon.

2. Validating Semiparametric Models

To fix the idea, consider the test against the partially linear model

$$H_0 : Y = g(U) + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon. \tag{4}$$

Again, for simplicity, we assume that $\varepsilon \sim N(0, \sigma^2)$. Let $\hat{g}$ and $\hat{\beta}_1, \cdots, \hat{\beta}_p$ be the estimates based on a sample of size n, using for example the profile likelihood approach (see e.g., Speckman (1988), Severini and Wong (1992), and Carroll, Fan, Gijbels and Wand (1997)). The profile likelihood gives semiparametric efficient estimators for parameters $\beta_1, \ldots, \beta_p$ and an optimal estimator for function g. With this, one can form the log-likelihood function under the null hypothesis, denoted by $\ell_n(H_0, h)$, where h is the bandwidth.

To test whether this model holds for a given data set, we need an alternative. Depending on the degree of prior belief on the model, one may consider the following possible alternative models.

1. **An additive model**: $H_{11} : Y = f_0(U) + f_1(X_1) + \cdots + f_p(X_p) + \varepsilon$.
2. **A varying-coefficient model**: $H_{12} : Y = f_0(U) + f_1(U)X_1 + \cdots + f_p(U)X_p + \varepsilon$.
3. **A full nonparametric model**: $H_{13} : Y = f(U, X_1, \cdots, X_p) + \varepsilon$.

The unknown nonparametric functions in the above models can easily be estimated, using for example kernel and local linear estimators with bandwidth h (for the additive model, one can use the backfitting algorithm as in Hastie and Tibshirani (1990)). Using the estimated nonparametric functions, one can form the nonparametric log-likelihood $\ell_n(H_{1j}, h)$ $(j = 1, 2, 3)$ as in Section 1, and the generalized likelihood ratio statistics $T_{n,j}(h) = \ell_n(H_{1,j}, h) - \ell_n(H_0, h), \quad j = 1, 2, 3.$ These form the generalized likelihood ratio test statistics for testing the semiparametric model (4) against the three nonparametric alternative models.

A few questions arise naturally. First of all, are the asymptotic null distributions for the test statistics independent of nuisance parameters in the null hypothesis? Secondly, do these test statistics achieve the optimal rates for hypothesis testing in the sense of Ingster (1993) and Spokoiny (1996)? Thirdly, what are the optimal rates for these three different alternatives?

In the additive model, Stone (1986) shows that one can estimate each additive component at the one-dimensional rate. Fan, Härdle and Mammen (1998) strengthen the result further in that one can estimate each additive component as well as if other components were known. The question then arises naturally if these kinds of results hold for hypothesis testing against the semiparametric model with the additive model as the alternative hypothesis.

3. Tests within Semiparametric Models

Suppose we have validated a semiparametric model. Various inference problems arise within the semiparametric model. For example, under the partially linear model, one may wish to test if certain variables are statistically significant, say

$$H_0 : \beta_1 = \cdots = \beta_m = 0.$$

More generally, one may consider the linear hypothesis:

$$H_0 : \mathbf{A}\boldsymbol{\beta} = 0, \tag{5}$$

where $\mathbf{A}$ is a given matrix and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$. This is a semiparametric null hypothesis versus a semiparametric alternative hypothesis. The testing problem is usually handled by using the Wald-type statistic, $W_n(h) = \hat{\boldsymbol{\beta}}^T \mathbf{A}^T (\mathbf{A}\hat{\Sigma}_h \mathbf{A}^T)^{-1} \mathbf{A}\hat{\boldsymbol{\beta}}$, where $\hat{\Sigma}_h$ is the estimated covariance matrix of $\boldsymbol{\beta}$, which involves estimated nonparametric function $\hat{g}$ and depends on a certain smoothing parameter h.

Note that under the null hypothesis (5), the problem is still a partially linear model. Hence, its parameters can be estimated by using the profile likelihood approach as in (4). The generalized likelihood ratio statistics can be computed by substituting the semiparametric estimators under both null and alternative hypotheses into the likelihood function, using the same bandwidth. Let the

resulting estimator be $T_n(h)$. The question then arises if the Wilks type of result holds. Between the two approaches $W_n(h)$ and $T_n(h)$, it remains to be seen which method is more powerful and which method gives a better approximation in terms of the size of the test.

For the partially linear model (1), one naive and simple approach is to use the partially linear structure to reduce the testing problem (5) to an approximate linear model. Let $(Y_i, U_i, X_{i1}, \cdots, X_{ip})$ be the random sample ordered according to the variable U. Then, by model (4),

$$\begin{aligned} Y_{2i+1} - Y_{2i} &= g(U_{2i+1}) - g(U_{2i}) + \beta_1(X_{2i+1,1} - X_{2i,1}) + \cdots \\ &\quad + \beta_p(X_{2i+1,p} - X_{2i,p}) + \varepsilon_{2i+1} - \varepsilon_{2i} \\ &\approx \theta_0 + \theta_1(U_{2i+1} - U_{2i}) + \beta_1(X_{2i+1,1} - X_{2i,1}) + \cdots \\ &\quad + \beta_p(X_{2i+1,p} - X_{2i,p}) + \varepsilon_{2i+1} - \varepsilon_{2i}. \end{aligned} \tag{6}$$

Note that the maximum distance between the spacing U_{2i+1} and U_{2i} is of order $O(n^{-1}\log n)$, when the density of U has a bounded support. Thus, the coefficients θ_0 and θ_1 in (6) can be taken to be zero 0. However, we keep these two parameters in the model to make the approximation more accurate. This kind of idea appears independently in Yatchew (1997) and Fan and Huang (2001). By using the approximate linear model (6), (5) becomes a linear hypothesis under the approximate linear model (6), and F-test statistics can be employed. One naturally asks how effective this simple and naive method is, compared with the more sophisticated Wald-test and the generalized likelihood ratio test. Note that we lose the information contained in the data $\{Y_{2i+1} + Y_{2i}\}$, which itself approximately follows (4). The data $\{Y_{2i+1} + Y_{2i}\}$ does not contain nearly as much information about $\boldsymbol{\beta}$ as $Y_{2i+1} - Y_{2i}$, since the former involves the nuisance function g. Thus, the efficiency based on (6) should, intuitively, be at least 50%.

Note that the above test can be regarded as a generalized likelihood ratio test with a very rough estimate of g. In fact, for given $\boldsymbol{\beta}$, one estimates g by taking the average of two neighboring points:

$$\begin{aligned} \hat{g}(u) = 2^{-1}\{&Y_{2i+1} + Y_{2i} - \beta_1(X_{2i+1,1} + X_{2i,1}) + \cdots \\ &+ \beta_p(X_{2i+1,p} + X_{2i,p})\}, \quad \text{for } u \in (\tfrac{U_{2i-1}+U_{2i}}{2}, \tfrac{U_{2i+1}+U_{2i+2}}{2}]. \end{aligned}$$

Substituting $\hat{g}$ into the models on Y_{2i+1}, we obtain

$$Y_{2i+1} - Y_{2i} = \beta_1(X_{2i+1,1} - X_{2i,1}) + \cdots + \beta_p(X_{2i+1,p} - X_{2i,p}) + 2\varepsilon_{2i+1}.$$

A similar equation is obtained by substituting $\hat{g}$ into the model on Y_{2i}:

$$Y_{2i+1} - Y_{2i} = \beta_1(X_{2i+1,1} - X_{2i,1}) + \cdots + \beta_p(X_{2i+1,p} - X_{2i,p}) - 2\varepsilon_{2i}.$$

The above two equations contain basically the same information as the model (6). Note that the estimator $\hat{g}$ here is significantly undersmoothed, but nonetheless gives reasonable inferences on the parametric component. It is consistent with a point hinted at in the paper by Bickel and Kwon.

After obtaining nonparametric estimate $\hat{g}$, researchers frequently ask if certain parametric model fits the nonparametric component. Namely, one wishes to test $H_0 : g(u) = g(u, \theta)$. Again, the generalized likelihood statistics can be constructed and its sampling properties need to be studied.

4. Choice of Bandwidth

Bickel and Kwon raise the question how to select bandwidths for semiparametric models. If the primary interest focuses on parametric components, the selected bandwidth should not create excessive biases in the estimation of nonparametric components. The reason is that the biases in the estimation of nonparametric components cannot be averaged out in the process of estimating parametric components, yet the variance in nonparametric estimates can be averaged out. This is evidenced in the approximate linear model (6), where g is estimated by the average of two neighboring points. If one wishes to choose a bandwidth that works well for parametric and nonparametric components simultaneously, a profile likelihood approach is needed, as demonstrated by Carroll et al. (1997). However, in semiparametric estimation problems, such as the partially linear model (4), one can also employ a two-step estimation scheme: choose a small bandwidth that efficiently estimates the parametric component, then treat the parametric component as if it were known and apply a nonparametric technique, with an optimally chosen bandwidth, to estimate the nonparametric component.

The problem of choosing an appropriate smoothing parameter arises also in the hypothesis testing problem. For each given bandwidth parameter h, one can regard the generalized likelihood test $T_n(h)$ (see e.g., (3)) as a proper test statistic. The question then becomes how to choose a good smoothing parameter that maximizes the power. The multi-scale test proposed in Fan (1996) appears to achieve good asymptotic power, as shown in Fan (1996) and Fan et al. (2001), though his formulation is in the frequency domain. The idea can simply be translated into the current setting. We refer to Zhang (2000) for some related work.

I have no intension to advocate using only the generalized likelihood ratio statistics for semiparametric and nonparametric inference. In fact, very few properties are known about the generalized likelihood ratio statistics. Even worse, the generalized likelihood statistics do not suggest any fixed procedure for estimating nonparametric components. Much more additional work is needed beyond the work in Fan et al. (2001). In light of no generally applicable guideline for

nonparametric and semiparametric testing problems, I outline some ideas, rather than some solutions, here in an attempt to address the model validation question raised by Bickel and Kwon, and to stimulate some further research in this area.

Acknowledgement

This research was partially supported by NSF grant DMS-0196041 and RGC grant CUHK 4262/01P of HKSAR.

Department of Statistics, Chinese University of Hong Kong, and Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, U.S.A.

E-mail: jfan@u30a.sta.cuhk.edu.hk

E-mail: jfan@stat.unc.edu

COMMENTS

Priscilla E. Greenwood, Anton Schick and Wolfgang Wefelmeyer

University of British Columbia, University of Binghamton and University of Siegen

In their thought-provoking essay, Bickel and Kwon (briefly, BK) touch on many important questions of semiparametric inference. We comment on only a few. Our Sections 1 to 4 concern BK's information calculus as applied to Markov chain models. In the first, we recall what BK call the traditional approach. The next two sections try to extract what we see as two essential points of the new information calculus in Markov chain models. The second of these points shows how to calculate efficient influence functions for Markov chains from corresponding bivariate i.i.d. models; this is particularly useful when the model and the parameter of interest are described in terms of the stationary law rather than the transition distribution. Section 4 is an aside on the converse: applying Markov chain results to bivariate i.i.d. models. Sections 5 to 7 discuss models more suited to the traditional approach: autoregression, conditional constraints, MCMC. Section 8 is on plugging kernel estimators into smooth functionals and into empirical estimators. Sections 9 and 10 briefly mention extensions of the traditional approach to continuous-time processes and to random fields.

1. The Traditional Approach.

In order to illustrate the power of BK's approach, we compare it with the traditional approach, which we recall first. For a review see Wefelmeyer (1999). Let $X^{(n)} = (X_1, \ldots, X_n)$ be observations from a stationary Markov chain with values in some state space E. (Here the state space may be arbitrary.) The natural parameter is the transition distribution, call it $q(x, dy)$. Let $\pi(dx)$, $b(dx, dy)$, and $P^{(n)}$ denote the laws of X_1, (X_1, X_2), and $X^{(n)}$, respectively. We have $b(dx, dy) = \pi(dx)q(x, dy)$. Consider (Hellinger differentiable) perturbations $q_{nh}(x, dy) \doteq q(x, dy)(1+n^{-1/2}h(x, y))$ of q. For q_{nh} to be a transition distribution, we must restrict h to $\mathcal{H}_0 = \{h \in L_2(b) : q_x h = 0\}$, where $q_x h = \int h(x, y)q(x, dy)$ denotes conditional expectation. The space $\mathcal{H}_0$ is the tangent space of the full nonparametric model. It is well known that we have local asymptotic normality at q,

$$\log \frac{dP^{(n)}_{nh}}{dP^{(n)}} = n^{-1/2} \sum_{i=1}^{n-1} h(X_i, X_{i+1}) - \frac{1}{2} \int h^2 \, db + o_p(1). \tag{1}$$

(We do not need the stronger form of local asymptotic normality used in BK, with perturbations involving factors t_n converging to some t.)

Consider now a submodel. It is given by a subset of transition distributions. Its *tangent space* at q is a subset of $\mathcal{H}_0$, say $\mathcal{H}_0^s$, which we take to be linear. Consider a real-valued functional $\vartheta(q)$ on the submodel. Assume it is *differentiable* at q with respect to the inner product induced by local asymptotic normality, with *gradient* $g \in \mathcal{H}_0$,

$$n^{1/2}(\vartheta(q_{nh}) - \vartheta(q)) \to \int hg \, db \quad \text{for all } h \in \mathcal{H}_0^s. \tag{2}$$

The *canonical* gradient is the projection g_s of g onto (the closure of) $\mathcal{H}_0^s$.

An estimator $\hat{\vartheta}$ of $\vartheta(q)$ is *regular* at q with *limit* L if L is a random variable such that

$$n^{1/2}(\hat{\vartheta} - \vartheta(q_{nh})) \Rightarrow L \quad \text{under } P^{(n)}_{nh} \quad \text{for all } h \in \mathcal{H}_0^s. \tag{3}$$

The convolution theorem says that $L = \left(\int g_s^2 \, db\right)^{1/2} \cdot N + M$ in distribution, with N standard Gaussian and M independent of N. This justifies calling a regular estimator $\hat{\vartheta}$ *efficient* for $\vartheta(q)$ if

$$n^{1/2}(\hat{\vartheta} - \vartheta(q)) \Rightarrow \left(\int g_s^2 \, db\right)^{1/2} \cdot N \quad \text{under } P^{(n)}.$$

An estimator $\hat{\vartheta}$ is *asymptotically linear* at q with *influence function* f if $f \in \mathcal{H}_0$ and

$$n^{1/2}(\hat{\vartheta} - \vartheta(q)) = n^{-1/2} \sum_{i=1}^{n-1} f(X_i, X_{i+1}) + o_p(1). \tag{4}$$

It is well known that an asymptotically linear estimator is regular if and only if its influence function is a gradient, and that a (regular) estimator is efficient if and only if it is asymptotically linear with influence function equal to the canonical gradient.

Example 1. Let us illustrate the traditional approach with a simple example, estimating a linear functional $\vartheta(q) = \int k\,db$, with $k \in L_2(b)$, in the full nonparametric model, with tangent space $\mathcal{H}_0$. For $h \in \mathcal{H}_0$ let $b_{nh}(dx, dy) = \pi_{nh}(dx)q_{nh}(x, dy)$. By a perturbation expansion (see e.g., Kartashov (1985), (1996)) we have

$$n^{1/2}\left(\int w\,db_{nh} - \int w\,db\right) \to \int h \cdot Tw\,db \quad \text{for all } w \in L_2(b), \tag{5}$$

where the operator $T : L_2(b) \to \mathcal{H}_0$ is $Tw(x, y) = w(x, y) - q_x w + \sum_{j=1}^{\infty}(q_y^j w - q_x^{j+1} w)$. This operator is a projection, $T = T^2$. It can also be written, as in BK, $Tw(x, y) = \overline{w}(x, y) - \sum_{j=1}^{\infty} q_x^j \overline{w} + \sum_{j=1}^{\infty} q_y^j \overline{w}$, where $\overline{w}(x, y) = w(x, y) - \int w\,db$ denotes centering. Relation (5), applied to $w = k$, says that the functional $\int k\,db$ has canonical gradient Tk in the sense of (2).

By a martingale approximation we have, for $w \in L_2(b)$,

$$n^{-1/2}\sum_{i=1}^{n-1}(\overline{w}(X_i, X_{i+1}) - Tw(X_i, X_{i+1})) = o_p(1). \tag{6}$$

(Relation (6) is called martingale approximation because $Tw(X_i, X_{i+1})$ are martingale increments. This approximation has been found independently by many authors, e.g., Gordin (1969), Maigret (1978), Dürr and Goldstein (1986) and Greenwood and Wefelmeyer (1995). See also Bradley (1988a,b) and Meyn and Tweedie (1993, Section 17.4). BK refer to Bickel (1993) and Künsch (1984)).

By the martingale approximation (6), applied to $w = k$, the empirical estimator $\hat{\vartheta} = \int k\,d\hat{b} = \frac{1}{n-1}\sum_{i=1}^{n-1} k(X_i, X_{i+1})$ satisfies

$$n^{1/2}\left(\int k\,d\hat{b} - \int k\,db\right) = n^{-1/2}\sum_{i=1}^{n-1} Tk(X_i, X_{i+1}) + o_p(1).$$

Hence the influence function, in the sense of (4), of the empirical estimator is Tk, the canonical gradient, and the estimator is regular and efficient by the two characterizations above.

2. An Equivalence Relation

The first point of BK on information calculus for Markov chains can be phrased as follows. Call $w, z \in L_2(b)$ *equivalent* if $Tw = Tz$. Then by the

martingale approximation (6), $n^{-1/2}\sum_{i=1}^{n-1}(\overline{w}(X_i, X_{i+1}) - \overline{z}(X_i, X_{i+1})) = o_p(1)$. Now parametrize locally with equivalence classes in $L_2(b)$ rather than their representatives in $\mathcal{H}_0$. Then for $h = Tw$, local asymptotic normality (1) can be written

$$\log \frac{dP_{nh}^{(n)}}{dP^{(n)}} = n^{-1/2}\sum_{i=1}^{n-1} \overline{w}(X_i, X_{i+1}) - \frac{1}{2}\int (Tw)^2\, db + o_p(1).$$

This is local asymptotic normality in the sense of Definition 1 of BK. Extend differentiability (2) of $\vartheta(q)$ correspondingly, calling m *gradient* of $\vartheta(q)$ if $m \in L_2(b)$ and

$$n^{1/2}(\vartheta(q_{nh}) - \vartheta(q)) \to \int h \cdot Tm\, db \quad \text{for all } h \in \mathcal{H}_0^s. \tag{7}$$

Any gradient m with Tm in (the closure of) $\mathcal{H}_0^s$ may then be called *canonical*. Extend asymptotic linearity (4) of $\hat{\vartheta}$, calling m *influence function* of $\hat{\vartheta}$ if $m \in L_2(b)$ and

$$n^{1/2}(\hat{\vartheta} - \vartheta(q)) = n^{-1/2}\sum_{i=1}^{n-1} \overline{m}(X_i, X_{i+1}) + o_p(1). \tag{8}$$

Then appropriate versions of the characterizations of regular and efficient estimators continue to hold.

Example 2. BK apply their approach in particular to the simple example above, estimating $\vartheta(q) = \int k\, db$ in the full nonparametric model, with tangent space $\mathcal{H}_0$. Write

$$n^{1/2}\left(\int k\, d\hat{b} - \int k\, db\right) = n^{-1/2}\sum_{i=1}^{n-1} \overline{k}(X_i, X_{i+1}) + o_p(1).$$

We have $Tk \in \mathcal{H}_0$. Hence k is a canonical gradient in the extended sense (7), and efficiency of the empirical estimator follows.

This proof is much shorter than the traditional one. Note, however, that the martingale approximation is also used here, namely for extending influence functions to equivalence classes. Efficiency of the empirical estimator was shown first by Penev (1990, 1991); he uses the perturbation expansion but circumvents the martingale approximation. Greenwood and Wefelmeyer (1995) show that the perturbation expansion follows from the martingale approximation.

3. From Bivariate Models to Markov Chains

The second point of BK in their information calculus applied to Markov chains is that canonical gradients can be obtained as in bivariate models, with i.i.d. observations (X_i, Y_i). This is extremely useful, especially for models and functionals which are more easily described in terms of the joint law b than of the transition distribution q.

Parametrize the Markov chain by the law b of (X_1, X_2) rather than by q. Then b must have equal marginals π: $\int v(x)\, b(dx, dy) = \int v(y)\, b(dx, dy)$ for all $v \in L_2(\pi)$. Consider (Hellinger differentiable) perturbations $b_{nw}(dx, dy) \doteq b(dx, dy)(1 + n^{-1/2} w(x, y))$. For b_{nw} to be a probability measure, we must have $\int w\, db = 0$. Since b_{nw} must also have equal marginals, we get

$$\int v(x) w(x, y)\, b(dx, dy) = \int v(y) w(x, y)\, b(dx, dy) \quad \text{for all } v \in L_2(\pi).$$

Hence the tangent space at b, say $\mathcal{H}$, is defined by having the following orthogonal complement in $L_2(b)$: $\mathcal{H}^\perp = \{v(x) - v(y) : v \in L_2(\pi)\}$.

Locally, the parameters b and q are related as follows. To go from b to q, factor b_{nw} as $b_{nw}(dx, dy) = \pi_{nw}(dx) q_{nw}(x, dy)$. Then π_{nw} is perturbed as

$$\pi_{nw}(dx) = b_{nw}(dx, E) \doteq \pi(dx)(1 + n^{-1/2} q_x w). \tag{9}$$

Hence $q_{nw}(x, dy) \doteq q(x, dy)(1 + n^{-1/2} w_0(x, y))$, where $w_0(x, y) = w(x, y) - q_x w$ denotes conditional centering. In particular, $\mathcal{H}_0 = \{w_0 : w \in \mathcal{H}\}$. To go from q to b, start from a perturbation $q_{nh}(x, dy) \doteq q(x, dy)(1 + n^{-1/2} h(x, y))$ with $h \in \mathcal{H}_0$. Write $b_{nh}(dx, dy) = \pi_{nh}(dx) q_{nh}(x, dy)$. For $w \in L_2(b)$ and $h \in \mathcal{H}_0$ write

$$\int h \cdot Tw\, db = \int Sh \cdot w\, db, \tag{10}$$

with an operator $S : \mathcal{H}_0 \to \mathcal{H}$ which we may call the adjoint of T. We do not need the explicit form of S; see Greenwood and Wefelmeyer (1999) for it. From (10) and the perturbation expansion (5) we obtain the perturbation

$$b_{nh}(dx, dy) \doteq b(dx, dy)(1 + n^{-1/2} Sh(x, y)). \tag{11}$$

In particular, $\mathcal{H} = \{Sh : h \in \mathcal{H}_0\}$. An analogous local parameter change, between densities and hazard functions, is described in Ritov and Wellner (1988).

Now consider a submodel described by some set of joint laws of (X_1, X_2). Its tangent space at b is a subset of $\mathcal{H}$, say $\mathcal{H}^s$, which we take to be linear. Consider a real-valued functional $\vartheta(b)$ on the submodel. Call it *differentiable* with *gradient* m if $m \in L_2(b)$ and

$$n^{1/2}(\vartheta(b_{nw}) - \vartheta(b)) \to \int wm\, db \quad \text{for all } w \in \mathcal{H}^s. \tag{12}$$

The *canonical* gradient is the projection m_s of m onto (the closure of) $\mathcal{H}^s$. Writing $w = Sh$ and using (10), we can characterize m_s as the function in (the closure of) $\mathcal{H}^s$ which fulfills

$$0 = \int Sh \cdot (m - m_s)\, db = \int h \cdot (Tm - Tm_s)\, db \quad \text{for all } h \in \mathcal{H}_0.$$

This means that Tm_s is the canonical gradient, in the traditional sense (2), in the Markov chain model. This is essentially Theorem 1 in Greenwood and Wefelmeyer (1999a). BK's second point is the following interpretation of this result. Suppose we have i.i.d. observations (X_i, Y_i). Consider a bivariate model of distributions $b(dx, dy)$, with equal marginals, and a real-valued differentiable functional $\vartheta(b)$ on this model. Calculate its canonical gradient m_s in the sense of (12). The canonical gradient of the corresponding Markov chain model is then Tm_s. Hence, using BK's first point, any function $z \in L_2(b)$ with $Tz = Tm_s$, in particular m_s itself, is a canonical gradient and efficient influence function in their extended sense.

Example 3. BK illustrate their second point with their Example 3b, a Markov chain model with *known* marginal distribution π. Then we must have $\pi_{nw}(dx) = \pi(dx)$ and hence $q_x w = 0$ by (9), and similarly $\pi(dy) = \pi_{nw}(dy) \doteq \pi(dy)(1 + n^{-1/2} q_y^- w)$. Here q^- is the transition distribution of the *reversed* chain, defined by $\pi(dx)q(x, dy) = \pi(dy)q^-(y, dx)$, and $q_y^- w = \int q^-(y, dx)w(x, y)$ is the conditional expectation under q^-, acting on the *first* argument of w. Hence the tangent space is $\mathcal{H}^s = \{w \in L_2^0(b) : qw = q^- w = 0\}$. Following BK, for $w \in L_2^0(b)$ we can write $qw = q^- w = 0$ as $\int (u(x) + v(y))w(x, y)\, b(dx, dy) = 0$ for all $u, v \in L_2(\pi)$. In words: w is orthogonal to functions of the form $u(x) + v(y)$. Now let $\vartheta(b)$ be differentiable, in the sense (12) of the bivariate model, with gradient $m \in L_2(b)$, say. As BK note, the *canonical* gradient in this sense can be obtained from Bickel, Klaassen, Ritov and Wellner ((1998), p. 440) as $m_s = m - \mathrm{ACE}(m)$, where $\mathrm{ACE}(m)$ is the projection of m onto the space of functions $u(x) + v(y)$. For $w \in \mathcal{H}^s$ we have $qw = 0$, i.e., $w(x, y) = w(x, y) - q_x w = w_0(x, y)$. The model is therefore degenerate: $\mathcal{H}^s = \mathcal{H}_0^s$. Hence m_s is also the traditional canonical gradient and efficient influence function in the sense (2).

Example 4. We agree that the Markov chain model with known marginals is possibly unrealistic. It is, however, not, as BK suggest, the model considered by Kessler, Schick and Wefelmeyer (2001). The latter assume not that the marginal is fixed but that it belongs to some parametric family π_ϑ, with ϑ one-dimensional, and they construct an efficient estimator for ϑ. The justification for such models comes from financial time series in which the marginal can be modeled more convincingly than the dynamics, especially when one has discrete observations from a continuous-time process. The efficient estimator is a complicated one-step improvement. We will consider elsewhere the possibility of finding a conceptually simpler estimator using BK's approach.

Example 5. Here is another application of BK's approach. Greenwood and Wefelmeyer (1999a) consider the model of all *reversible* Markov chains. This

means that b is symmetric, $b(dx, dy) = b(dy, dx)$, or equivalently, $q = q^-$. They prove that the symmetrized empirical estimator

$$\hat{\vartheta}_s = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (k(X_i, X_{i+1}) + k(X_{i+1}, X_i))$$

is efficient for $\int k\, db$. The proof is also based on their Theorem 1 used above. For a more elegant version of this proof we follow BK and parametrize with b. The tangent space of the bivariate model is $\mathcal{H}^s = \{w \in L_2^0(b) : w(x,y) = w(y,x)\}$. Consider a real-valued functional $\vartheta(b)$ which is differentiable, in the sense (12), with gradient $m \in L_2^0(b)$. The *canonical* gradient in the bivariate model is the symmetrized $m_s(x,y) = \frac{1}{2}(m(x,y) + m(y,x))$. Hence the canonical gradient in the Markov chain model is Tm_s. Hence, by BK's first point, m_s is also an efficient influence function in the Markov chain model. This proves that if $\hat{\vartheta}$ is asymptotically linear with influence function m in the Markov chain model, then its symmetrization $\hat{\vartheta}_s = \frac{1}{2}(\hat{\vartheta}(X_1, \ldots, X_n) + \hat{\vartheta}(X_n, \ldots, X_1))$ is regular and efficient in the model of all reversible Markov chains. In particular, the symmetrized *empirical* estimator is efficient.

Example 6. Müller, Schick and Wefelmeyer (2001b) consider the nonparametric Markov chain model with linear constraint $\int z\, db = 0$ for some d-dimensional vector $z \in L_2(b)^d$. They construct efficient estimators for linear functionals $\int k\, db$, following the traditional approach and Levit (1975), who considers the i.i.d. case. The canonical gradient is $Tk - c_*^\top Tz$ with $c_* = \left(\int Tz \cdot Tz^\top\, db\right)^{-1} \int Tz \cdot Tk\, db$. Let us derive this result using BK's approach. Parametrize by b. We have $n^{1/2}\,(\int z\, db_{nw} - \int z\, db) \to \int zw\, db$. Because of the constraints $\int z\, db = \int z\, db_{nw} = 0$ we must have $\int zw\, db = 0$. Hence the tangent space of the corresponding bivariate model is $\mathcal{H}^s = \{w \in \mathcal{H} : \int zw\, db = 0\}$. By Example 2, k is a gradient of $\int k\, db$ in the extended sense (7). Write $w_\mathcal{H}$ for the projection of a function $w \in L_2(b)$ onto $\mathcal{H}$. In the bivariate model, because of the constraint $\int z\, db = 0$, all functions $k - a^\top z$ with $a \in \mathbf{R}$ are gradients, and hence all functions $k_\mathcal{H} - a^\top z_\mathcal{H}$ are gradients in $\mathcal{H}$. The canonical gradient is the projection of any of these gradients onto $\mathcal{H}^s$. It must minimize $\int (k_\mathcal{H} - a^\top z_\mathcal{H})^2\, db$ in a, call it a_*. By BK's second point, $k_\mathcal{H} - a_*^\top z_\mathcal{H}$ is also a canonical gradient in the constrained Markov chain model, in their extended sense (7).

Of course, $k_\mathcal{H} - a_*^\top z_\mathcal{H}$ must be equivalent to the traditional canonical gradient $Tk - c_*^\top Tz$ in the bivariate model. This follows from two observations.

1. If $w \in L_2(b)$ and $w_\mathcal{H}$ is its projection onto $\mathcal{H}$, then $w - w_\mathcal{H}$ is in $\mathcal{H}^\perp$, i.e., of the form $v(x) - v(y)$. Such functions are annihilated by T. Hence w and $w_\mathcal{H}$ are equivalent: $Tw = Tw_\mathcal{H}$. In particular, $k_\mathcal{H} - a_*^\top z_\mathcal{H}$ is equivalent to $Tk - a_*^\top Tz$.

2. The operator ST is a projection onto $\mathcal{H}$. Hence we obtain, using (10),

$$\int Tw \cdot Tm \, db = \int w \cdot STm \, db = \int wm \, db \quad \text{for all } w, m \in \mathcal{H}.$$

In particular, $a_* = c_*$.

An efficient estimator for $\int k \, db$ under the constraint $\int z \, db = 0$ is the improved empirical estimator

$$\frac{1}{n-1} \sum_{i=1}^{n-1} (k(X_i, X_{i+1}) - \hat{a}_*^\top z(X_i, X_{i+1})).$$

It requires a consistent estimator $\hat{a}_*$ for a_*. Such an estimator is constructed in Müller, Schick and Wefelmeyer (2001b). It is based on an explicit representation of a_*. Calculating a_* requires calculating the projections of z and k onto $\mathcal{H}$. Example 7 shows how projections $w_\mathcal{H}$ of functions $w \in L_2(b)$ onto $\mathcal{H}$ are obtained, via the traditional approach, as $w_\mathcal{H} = STw$. One checks that by (10) this gives again $a_* = c_*$.

This example shows that even if the model and functional of interest are in terms of the joint law b rather than the transition distribution q, the traditional approach is not necessarily more awkward than the approach via the bivariate model. One reason is the following. The traditional approach parametrizes by q and uses an unpleasant local parameter space $\mathcal{H}_0$, equipped however with the natural norm $\int w^2 \, db$. If we introduce equivalence classes as suggested in BK's first point, then we end up with a simple local parameter space $L_2(b)$, but now equipped with the unpleasant semi-norm $\int (Tw)^2 \, db$. On the other hand, if we parametrize by b as suggested in BK's second point, then we end up with the natural norm but with an unpleasant local parameter space $\mathcal{H}$.

4. From Markov Chains to Bivariate Models

We have seen in Section 3 how canonical gradients in Markov chain models can be obtained from canonical gradients in bivariate models. The converse is of course also possible and, more surprisingly, sometimes useful.

Consider a Markov chain model described by some set of transition distributions. Its tangent space at q is a subset $\mathcal{H}_0^s$ of $\mathcal{H}_0$, taken to be linear. Let $\vartheta(q)$ be a real-valued functional which is differentiable, in the (traditional) sense (2), with canonical gradient $g_s \in \mathcal{H}_0^s$. Set $h = Tw$ and use (10) to rewrite differentiability (2) as

$$n^{1/2}(\vartheta(q_{nh}) - \vartheta(q)) \to \int Tw \cdot g_s \, db = \int w \cdot Sg_s \, db \quad \text{for all } w \in \mathcal{H}^s.$$

This is differentiability in the sense (12) of the bivariate model. Hence Sg_s is the canonical gradient of $\vartheta(q)$, viewed as functional on the bivariate model.

Example 7. This is already useful in the simplest example, estimating the linear functional $\vartheta(q) = \int k\,db$, with $k \in L_2(b)$, in the full nonparametric Markov chain model. Its (canonical) gradient is $g = Tk$. The corresponding bivariate model is the model with equal marginals. It follows that $Sg = STk$ is the canonical gradient in this model. The explicit form of ST can be obtained from results for Markov chain models, see Greenwood and Wefelmeyer (1999a). An efficient estimator in the bivariate i.i.d. model with equal marginals is constructed in Peng and Schick (2001). It does not use the explicit form of the canonical gradient.

5. Regression and Autoregression

An important class of Markov chain models are autoregressive models $X_{i+1} = r(X_i)+\varepsilon_{i+1}$, where the innovations ε_i are i.i.d. with mean zero and finite variance σ^2 and have an absolutely continuous and positive density f with finite Fisher information $J = \int \ell^2\,dF$ for location, where $\ell = -f'/f$ and F is the distribution function of f. For convenience we consider only first-order autoregression. For the model to be ergodic, the autoregression function r must satisfy some growth conditions; see e.g., Bhattacharya and Lee (1995). BK consider the nonparametric model, with r unknown. Submodels are the linear model, with $r(x) = \vartheta x$, and nonlinear models with parametric families $r_\vartheta(x)$ of autoregression functions. Here it suggests itself to follow the traditional approach and describe the model by its transition distribution $q(x, dy) = f(y - r(x))\,dy$.

The information calculus of Section 3 would suggest looking at the bivariate i.i.d. model described by the joint law $b(dx, dy) = \pi(dx)q(x, dy)$ of (X_1, X_2). Perturbation of q would, however, result in a complicated perturbation of π, see (11), and in a complicated tangent space of the bivariate model.

Nevertheless, it pays to look at an i.i.d. model analogous to the Markov chain model, namely regression $Y_i = r(X_i) + \varepsilon_i$, with ε_i as before, and i.i.d. covariates X_i, independent of the ε_i, with known law $c(dx)$, say. The joint law of (X_1, Y_1) is $c(dx)f(y - r(x))\,dy$. Tangent spaces and gradients for autoregression are therefore the same as for regression. Schick (1993) considers functionals of (c, r); for extensions to heteroscedastic regression see Schick (1994).

Following the traditional approach to autoregression, see Koul and Schick (1997), consider (Hellinger differentiable) perturbations $f_{nv} \doteq f(1 + n^{-1/2}v)$. Since the innovations are assumed to have mean zero, the local parameters v must be in the orthogonal complement V in $L_2(F)$ of the polynomials of degree at most one,

$$V = \{v \in L_2(F) : \int v(\varepsilon)\,dF(\varepsilon) = \int \varepsilon v(\varepsilon)\,dF(\varepsilon) = 0\}.$$

The model also specifies a family of autoregression functions. Consider (π-square-differentiable) perturbations $r_{nu} \doteq r + n^{-1/2}u$. The model restricts u

to some subset of $L_2(\pi)$, say U, which we take to be (closed and) linear. The transition density determined by f_{nv} and r_{nu} is $f_{nv}(y - r_{nu}(x)) \doteq f(\varepsilon)(1 + n^{-1/2}(u(x)\ell(\varepsilon) + v(\varepsilon)))$. Hence the tangent space of the autoregressive model is $\mathcal{H}_0(U) = \{u(x)\ell(\varepsilon) + v(\varepsilon) : u \in U, v \in V\}$. The tangent space is the sum of the tangent spaces $\{u(x)\ell(\varepsilon) : u \in U\}$ for known f, and $\{v(\varepsilon) : v \in V\}$ for known r. It is well known that one can estimate (all smooth functionals of) f and r adaptively with respect to each other if and only if these two spaces are orthogonal.

Example 8. Schick and Wefelmeyer (2001b) obtain efficient estimators for $\int a\,dF$ when the autoregression functions are restricted to a parametric family r_ϑ. For simplicity, we take ϑ one-dimensional here. Then U is the linear span $[\dot r_\vartheta]$ of the derivative of r_ϑ with respect to ϑ, and the tangent space is $\mathcal{H}_0([\dot r_\vartheta]) = \{t\dot r_\vartheta(x)\ell(\varepsilon) + v(\varepsilon) : t \in \mathbf{R}, v \in V\}$. Unless $\int \dot r_\vartheta\,d\pi = 0$, the tangent space is not an orthogonal sum, and f and r cannot be estimated adaptively with respect to each other. A natural estimator of $\int a\,dF$ is the empirical estimator $\frac{1}{n-1}\sum_{i=1}^{n-1} a(\hat\varepsilon_{i+1})$ based on estimated innovations $\hat\varepsilon_{i+1} = X_{i+1} - r_{\hat\vartheta}(X_i)$. It can be improved using that the innovations have mean zero,

$$\hat A = \frac{1}{n-1}\sum_{i=1}^{n-1}(a(\hat\varepsilon_{i+1}) - \hat c\hat\varepsilon_{i+1}), \tag{13}$$

with $\hat c$ a consistent estimator for the optimal constant

$$c = \sigma^{-2}\int \varepsilon a(\varepsilon)\,dF(\varepsilon). \tag{14}$$

An obvious choice is $\hat c = \sum_{i=1}^{n-1}\hat\varepsilon_{i+1}a(\hat\varepsilon_{i+1})/\sum_{i=1}^{n-1}\hat\varepsilon_{i+1}^2$. The influence function of $\hat A$ requires some notation, and we do not give it here. In the non-adaptive situation, with $\int \dot r_\vartheta\,d\pi$ not zero, for $\hat A$ to be efficient we must estimate $\varepsilon_{i+1} = X_{i+1} - r_\vartheta(X_i)$ using an *efficient* estimator for ϑ.

Plug-in of finite-dimensional estimators in not necessarily adaptive situations is studied in Klaassen and Putter (1997, 2000) for i.i.d. models, and more generally in Müller, Schick and Wefelmeyer (2001a).

Example 9. In their Example 3a, BK consider estimating $\int a\,dF$ in the nonparametric autoregressive model, with r unknown except for mean zero. Then $U = L_2(\pi)$, and the tangent space is $\mathcal{H}_0(L_2(\pi)) = \{u(x)\ell(\varepsilon)+v(\varepsilon) : u \in L_2(\pi), v \in V\}$. This is not an orthogonal sum. Hence f and r cannot be estimated adaptively with respect to each other. (BK state that the tangent space equals that with Gaussian innovation distribution with known variance, their (3.30), and later that it contains all functions $v(\varepsilon)$ with $v \in L_2(\pi)$. These statements are not consistent with each other and with the tangent space obtained here.) The canonical gradient for $\int a\,dF$ is the same as in the corresponding regression model,

Müller, Schick and Wefelmeyer (2001c), namely $\overline{a}(\varepsilon) - \int a\ell\,dF \cdot \varepsilon$. One can show that the empirical estimator $\frac{1}{n-1}\sum_{i=1}^{n-1} a(\hat{\varepsilon}_{i+1})$ based on estimated innovations $\hat{\varepsilon}_{i+1} = X_{i+1} - \hat{r}(X_i)$ has this influence function. To check that this function is indeed in the tangent space $\mathcal{H}_0(L_2(\pi))$, rewrite it as

$$\overline{a}(\varepsilon) - \int a\ell\,dF \cdot \varepsilon = -\sigma^2 \int a\ell_V\,dF \cdot \ell(\varepsilon) + a_V(\varepsilon) + \sigma^2 \int a\ell_V\,dF \cdot \ell_V(\varepsilon),$$

where a_V and ℓ_V are the projections of a and ℓ onto V, $a_V(\varepsilon) = \overline{a}(\varepsilon) - c\varepsilon$, $\ell_V(\varepsilon) = \ell(\varepsilon) - \sigma^{-2}\varepsilon$. We note that in this non-adaptive model, the canonical gradient for *known* regression function r is indeed different: it is just the projection a_V of a onto V, and an efficient estimator is the *improved* empirical estimator $\frac{1}{n-1}\sum_{i=1}^{n-1}(a(\varepsilon_{i+1}) - \hat{c}\varepsilon_{i+1})$ based on *true* innovations. Compare also Example 8 on parametric autoregression functions r_ϑ.

These results are not consistent with the statements of BK that the empirical estimators with true and estimated innovations are asymptotically equivalent, that their influence function is $a(\varepsilon)$, and that this function is in the tangent space, which would imply that $\frac{1}{n-1}\sum_{i=1}^{n-1} a(X_{i+1} - \hat{r}(X_i))$ is adaptive with respect to r.

Example 10. BK ascribe their statements about $\frac{1}{n-1}\sum_{i=1}^{n-1} a(X_{i+1} - \hat{r}(X_i))$ in nonparametric autoregression to Wefelmeyer (1994). But the latter treats only *linear* autoregression $X_{i+1} = \vartheta X_i + \varepsilon_{i+1}$, and proves that the *improved* empirical estimator $\hat{A}$, now with innovations estimated by $\hat{\varepsilon}_{i+1} = X_{i+1} - \hat{\vartheta}X_i$, is efficient. Linear autoregression is a special case of the nonlinear model above, with $r_\vartheta(x) = \vartheta x$ and $\dot{r}_\vartheta(x) = x$. The tangent space is therefore $\mathcal{H}_0^s = \{tx\ell(\varepsilon) + v(\varepsilon) : t \in \mathbf{R}, v \in V\}$. Since the innovations have mean zero, so has the stationary law π. This implies that the tangent space is an orthogonal sum, and ϑ and f can be estimated adaptively with respect to each other. In particular, $\hat{A}$ is efficient for $\int a\,dF$ even when an inefficient estimator $\hat{\vartheta}$ is used in the estimated innovations $\hat{\varepsilon}_{i+1} = X_{i+1} - \hat{\vartheta}X_i$.

Example 11. Another adaptive example is nonparametric autoregression with innovations that are *symmetric about zero*. The tangent space is $\mathcal{H}_0^s = \{u(x)\ell(\varepsilon) + v(\varepsilon) : u \in L_2(\pi), v \in L_2(F)$ symmetric about zero$\}$. Here $\ell(\varepsilon) = -\ell(-\varepsilon)$. Hence $\int v\ell\,dF = 0$ for all v that are symmetric about zero, and the tangent space is an orthogonal sum. Koshevnik (1996) shows that the symmetrized empirical distribution function based on estimated innovations is efficient.

Example 12. Kwon (2000) and BK also consider estimating $\int r(x)\lambda(x)\,dx$ in the nonparametric autoregression model with mean zero innovations. Here λ is known and has compact support. They suggest that an efficient estimator is obtained by plugging in a suitable (kernel) estimator $\hat{r}$ for r. Kwon (2000) shows

that the estimator $\int \hat{r}(x)\lambda(x)\,dx$ has influence function $\varepsilon\lambda(x)/f(x)$. From Schick ((1993), (3.5)), the canonical gradient of $\int r(x)\lambda(x)\,dx$ is obtained as

$$\Big(\frac{\lambda(x)}{f(x)} - \int \lambda(y)\,dy\Big)\frac{\ell(\varepsilon)}{J} + \int \lambda(y)\,dy \cdot \varepsilon,$$

with J the Fisher information for location of the innovation distribution. This is the influence function of BK's estimator only if the innovation distribution is Gaussian, so their estimator is efficient only if the true innovation distribution happens to be Gaussian.

The traditional approach has also been used in more complicated autoregressive models. For example, Schick (1999a) treats the semiparametric model $X_{i+1} = \vartheta X_i + r(X_{i-1}) + \varepsilon_{i+1}$. Maercker (1997) and Schick (2001) treat the heteroscedastic autoregressive model $X_{i+1} = \vartheta X_i + s(X_i)\varepsilon_{i+1}$ with symmetric errors, while Schick (1999b) considers it with arbitrary errors. Efficient estimation in invertible linear processes is treated in Schick and Wefelmeyer (2001c).

6. Conditional Constraints

Another class of submodels described through transition distributions rather than joint laws are models with constraints $E(v_\vartheta(X_1, X_2)|X_1) = 0$ for some d-dimensional vector $v_\vartheta \in L_2(b)$. These comprise quasi-likelihood models, with parametric models for conditional mean and variance of the Markov chain:

$$E(X_2|X_1) = r_\vartheta(X_1),$$
$$E((X_2 - r_\vartheta(X_1))^2|X_1) = s^2_\vartheta(X_1).$$

Here $v_\vartheta(x, y)$ has components $y - r_\vartheta(x)$ and $(y - r_\vartheta(x))^2 - s^2_\vartheta(x)$. The quasi-maximum-likelihood estimator solves an estimating equation of the form

$$\sum_{i=1}^{n-1} w_\vartheta(X_i, X_{i+1})(X_{i+1} - r_\vartheta(X_i)) = 0,$$

with weights w_ϑ chosen to minimize the asymptotic variance. It does not use the information in the specification of the conditional variance and is not efficient. Efficient estimating equations are constructed in Wefelmeyer (1996). For similar regression models with i.i.d. observations, quite different efficient estimators are introduced in Li (2000) and (2001). Efficient estimation of invariant laws in such models is discussed in Schick and Wefelmeyer (1999).

7. MCMC

A third class of submodels described by transition distributions are Monte Carlo Markov chains. Here one starts with a distribution $\pi(dx)$ which is in

principle known, and constructs a transition distribution $q(x, dy)$ with π as invariant law. Then one runs the corresponding Markov chain and approximates, e.g., $\int a(x)\pi(dx)$ by the empirical estimator $\sum_{i=1}^{n} a(X_i)$. Greenwood, McKeague and Wefelmeyer (1998) calculate the information in the knowledge that a Gibbs sampler was used. A review is Greenwood and Wefelmeyer (2001).

8. Plug-in Estimators

As BK point out, $n^{1/2}$-consistent and even efficient estimators can often be obtained by plugging density estimators or regression function estimators into smooth functionals or into "empirical estimators" involving such functions. BK's estimators for $\int r(x)\lambda(x)\,dx$ and $\int a\,dF$ in nonparametric autoregression are examples of plug-in into a smooth functional and into an empirical estimator.

For i.i.d. observations with density f, smooth functionals of f can be estimated efficiently by plugging in (undersmoothed) kernel estimators; see Abramson and Goldstein (1991), Goldstein and Messer (1992) and Goldstein and Khas'minskii (1995).

For expectations of functions of more than two arguments, e.g., $E\psi(X_1, X_2, X_3)$, the empirical estimator based on Markov chain observations is not efficient in the nonparametric Markov chain model. Writing $E\psi(X_1, X_2, X_3) = \int \psi(x, y, z)\, b(dx, dy)q(y, dz)$, one sees that for discrete state space a better estimator is obtained by replacing b and q by their empirical estimators. For general state space, Schick and Wefelmeyer (2001a) construct a complicated efficient estimator as one-step improvement of the empirical estimator. Bickel (1993) has suggested a conceptually simpler estimator, using the empirical estimator for b as before, and plugging in a nonparametric estimator $\hat{q}$ for the transition density. Kwon (2000) treats a modification of this idea, writing the density of the joint law of (X_1, X_2, X_3) as $p(x, y)p(y, z)/g(y)$ with g and p the densities of X_1 and (X_1, X_2), respectively, and replacing these densities by kernel estimators.

Example 13. Here is another application of plug-in. For moving average processes $X_{i+1} = \varepsilon_{i+1} - \vartheta\varepsilon_i$, the density $g(x)$ of X_{i+1} can be written as convolution of the density f of ε_{i+1} and the density of $\vartheta\varepsilon_i$, $g(x) = \int f(x + \vartheta y)f(y)dy$. Saavedra and Cao (1999) and (2000) propose plugging in (undersmoothed) kernel estimators $\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n} K_c(z - \hat{\varepsilon}_i)$, where $K_c(u) = K(u/c)/c$ and $\hat{\varepsilon}_i$ are estimated innovations. They obtain $n^{1/2}$-consistency of their estimator $\int \hat{f}(x + \hat{\vartheta}y)\hat{f}(y)dy$. Schick and Wefelmeyer (2001e) propose the asymptotically equivalent, but simpler, U-statistic

$$\hat{g}(x) = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} K_c(x - \hat{\varepsilon}_i + \hat{\vartheta}\hat{\varepsilon}_j)$$

and prove that it is efficient. The estimator can be written (approximately) as the plug-in estimator $\frac{1}{n}\sum_{i=1}^{n}\hat{f}(x+\hat{\vartheta}\hat{\varepsilon}_i)$.

We note that estimators based on U-statistics have many applications in semiparametric inference. For example, U-statistics with *fixed* kernel are used in Schick and Wefelmeyer (2001d) to estimate expectations under the stationary law of invertible linear processes.

9. Continuous-time Processes

The traditional approach generalizes immediately to continuous-time processes X_t, $t \geq 0$, observed on an increasing time interval $[0, n]$, say. For counting processes, the intensity plays the role of the transition distribution as natural parameter; diffusion processes $X_t = r(X_t)dt + s(X_t)dB_t$ are parametrized by drift r and diffusion coefficient s. More generally, semimartingales are parametrized by their predictable characteristics, Jacod and Shiryaev (1987) is the standard reference for structure theory and limit theorems. Other types of asymptotics are also possible. For counting processes we may let the intensity increase. For diffusion processes, we may let the diffusion coefficient decrease, see Kutoyants (1994). In survival analysis one usually considers an increasing number of paths, a comprehensive reference including non- and semiparametric efficiency results is Andersen, Borgan, Gill and Keiding (1993).

Efficient plug-in estimators for the stationary density of diffusion processes are obtained in Kutoyants (1997), (1998) and (1999). Empirical estimators are shown to be efficient in nonparametric Markov step process and semi-Markov process models by Greenwood and Wefelmeyer (1994a) and (1996), and in nonparametric multivariate point process models by Greenwood and Wefelmeyer (1994b). It seems possible to use versions of BK's approach in such models.

10. Random Fields

The traditional approach also generalizes to homogeneous random fields on lattices, where the transition distribution is replaced by the local characteristic, the conditional distribution at a site given the rest of the configuration. For random fields with local interactions, Greenwood and Wefelmeyer (1999b) determine which empirical estimators are efficient.

Acknowledgement

This research was supported in part by NSERC, Canada, and was supported in part by NSF Grant DMS 0072174.

Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada.

E-mail: pgreenw@math.ubc.ca

Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, U.S.A.

E-mail: anton@marge.math.binghamton.edu

FB6 Mathematics, University of Siegen, Walter Flex St. 3, 57068 Siegen, Germany.

E-mail: wefelmeyer@mathematik.uni-siegen.de

COMMENTS

Chris A. J. Klaassen

University of Amsterdam

This paper is very challenging. Peter Bickel and Jaimyoung Kwon put semiparametric inference in perspective by describing the historical development from parametric to semiparametric statistics. They also sketch the important features of this development for the i.i.d. case. This leads them to formulating a list of five research questions. In fact, each of these questions depicts a whole area of interesting and indeed challenging problems, which undoubtedly will stimulate research in semiparametrics. The authors are to be congratulated for their well-organized, clearly written, and useful presentation, in which they elaborate on one of the research problems on their list, namely the generalization of semiparametric theory to the non-i.i.d. world.

To me, the crucial idea in their approach to this problem is their generalization of the concept of asymptotic linearity of estimators. The authors view the average of the influence function at the i.i.d. observations as the expectation of this influence function under the empirical and replace the empirical by an appropriate estimator of the 'core' distribution in the non-parametric version of the non-i.i.d. model under study; cf. (A1), (3.1), and (3.3). This is a simple but attractive idea, which works well in the many special cases studied in the paper.

Asymptotics is by far the most frequently used approach to mathematical statistical problems, and so the authors focus in their future research questions on this approach. Taking limits typically simplifies the mathematical problems and hopefully yields good approximations to the finite sample size truth. In this vein asymptotic normality is quite relevant. However, asymptotic consistency, and to a lesser extent asymptotic $\sqrt{n}$-consistency and regularity are more difficult to interpret at the level of finite sample sizes; most of all because uniformity issues are not considered, typically. Therefore, we should have an eye to finite

sample sizes, and I would like to suggest to add the ultimate question (Z) to the list (A) – (E) of Bickel and Kwon about the main goal—in my opinion—of all semiparametric research, namely the problem of finite sample size (optimal) behavior of inference procedures in semiparametric models.

Within the context of the semiparametric symmetric location problem an early attempt to construct bounds to the finite sample size performance of estimators may be found in Section 3.2 of Klaassen (1980). These bounds imply that the convergence to normality of adaptive location estimators cannot be uniform at all (see also Klaassen (1979)).

In question (A) Bickel and Kwon note that in many semiparametric models estimation of irregular parameters, such as densities and their derivatives, seems necessary in order to estimate the parameter of interest efficiently. They suggest to study the selection of bandwidth and other regularization parameters for these problems, in particular for those models in which these irregular parameters can be estimated at specified rates. In fact, it has been proved that in general, efficient estimation of the parameter of interest is possible if and only if its efficient influence function can be estimated consistently by a $\sqrt{n}$-unbiased estimator; see Klaassen (1987). To me this suggests that rates of convergence for estimators of the influence function, an irregular parameter, are related to the amount of uniformity that can be attained in the convergence of an efficient estimator of the parameter of interest.

Another point I would like to raise is about quite a curious phenomenon. Typically, in semiparametric models the Euclidean parameter of interest is identifiable and has positive semiparametric Fisher information. Some of these models can be extended in such a way that the Euclidean parameter of interest is still identifiable, but has vanishing semiparametric Fisher information at all distributions in the model. Lenstra (1998) has shown this for the regression parameter in the mixed proportional hazards model, which extends the Cox proportional hazards model via a nonparametric unobservable random frailty factor; see also Klaassen and Lenstra (2000). This vanishing Fisher information is bound to have serious consequences both asymptotically and for finite sample sizes. It might well be that this interesting phenomenon could occur more easily in the non-i.i.d. cases that Bickel and Kwon stimulate us to study.

Finally, I would like to comment on a minor but cute point. As the authors mention in Example 3, quite often constructions of efficient semiparametric procedures use the technique of splitting up the sample in independent parts, thus simplifying the proof of efficiency. As far as I know, the first paper constructing an efficient, adaptive semiparametric procedure is by Hájek (1962) for the linear regression problem. He applies this sample splitting technique in that he uses a vanishingly small part of the sample to estimate the score function for location

and subsequently applies this estimate in a testing procedure for the hypothesis of a vanishing regression function based on the remaining part of the sample. For nonadaptive cases the sample has to be split up in at least two substantial parts, in the sense that they may not be vanishingly small; see Schick (1986) and Klaassen (1987) for the first papers using this type of sample splitting. At first sight, sample splitting seems unnatural or unrealistic, as the authors say in Example 3. However, for nonadaptive situations the crux of sample splitting is that an average is taken of estimators essentially based on the independent parts in which the sample is split up. Since averaging is intimately related to normality and since efficient estimators are to be asymptotically normal, sample splitting might well be reasonable and even quite natural. To support this claim we present the following result on Edgeworth expansions.

Proposition 1. *Let $X_1, \dots, X_n$ be i.i.d. random variables and let $T_n = t_n(X_1, \dots, X_n)$ be a one-dimensional statistic such that $\sqrt{n}T_n$ has distribution function $F_n(\cdot)$ and Edgeworth expansion*

$$\tilde{F}_n(x) = \Phi(x) - \varphi(x)\left\{ \frac{c_1}{\sqrt{n}}(x^2 - 1) + \frac{c_2}{n}(x^3 - 3x) + \frac{1}{2}\frac{c_1^2}{n}(x^5 - 10x^3 + 15x)\right\}, \quad x \in \mathbb{R}, \tag{1}$$

for some constants c_1 and c_2 with

$$\sup_x n|F_n(x) - \tilde{F}_n(x)| \to 0, \quad \text{as } n \to \infty. \tag{2}$$

Split the sample into two parts, $X_1, \dots, X_m$ and $X_{m+1}, \dots, X_n$ respectively. Let $T_{1,m} = t_m(X_1, \dots, X_m)$ be the statistic based on the first part and $T_{2,n-m} = t_{n-m}(X_{m+1}, \dots, X_n)$ the statistic based on the second part. Define the sample splitting statistic as

$$\tilde{T}_n = \frac{m}{n}T_{1,m} + \frac{n-m}{n}T_{2,n-m}. \tag{3}$$

If

$$0 < \liminf_{n\to\infty} \frac{m}{n} \le \limsup_{n\to\infty} \frac{m}{n} < 1 \tag{4}$$

holds, then $\sqrt{n}\tilde{T}_n$ has the same Edgeworth expansion (1) *as $\sqrt{n}T_n$.*

Proof. Uniformly in $z \in \mathbb{R}$ we have

$$\begin{aligned} P(\sqrt{n}\tilde{T}_n \le z) &= \int F_m\left(\sqrt{\frac{n}{m}}(z-y)\right) dF_{n-m}\left(\sqrt{\frac{n}{n-m}}\, y\right) \\ &= \int \tilde{F}_m\left(\sqrt{\frac{n}{m}}(z-y)\right) dF_{n-m}\left(\sqrt{\frac{n}{n-m}}\, y\right) + o\left(\frac{1}{n}\right) \end{aligned}$$

$$= \int F_{n-m}\left(\sqrt{\frac{n}{n-m}}(z-y)\right) d\tilde{F}_m\left(\sqrt{\frac{n}{m}}\, y\right) + o\left(\frac{1}{n}\right)$$
$$= \int \tilde{F}_{n-m}\left(\sqrt{\frac{n}{n-m}}(z-y)\right) d\tilde{F}_m\left(\sqrt{\frac{n}{m}}\, y\right) + o\left(\frac{1}{n}\right). \quad (5)$$

If $T_n = n^{-1}\sum_{i=1}^n X_i$ is the sample mean with $\mathrm{E}X_i = 0$, then $\tilde{T}_n = T_n$ and (5) shows

$$\int \tilde{F}_{n-m}\left(\sqrt{\frac{n}{n-m}}(z-y)\right) d\tilde{F}_m\left(\sqrt{\frac{n}{m}}\, y\right) = \tilde{F}_n(z) + o\left(\frac{1}{n}\right), \quad (6)$$

uniformly in $z \in \mathbb{R}$. In our general case $\tilde{F}_n$ has the same structure (cf. Theorem VI.3.1, p.159, of Petrov (1975)) and hence (6) holds for $\tilde{F}_n$ from (1). Equations (5) and (6) imply, uniformly in $z \in \mathbb{R}$,

$$P(\sqrt{n}\tilde{T}_n \le z) = \tilde{F}_n(z) + o\left(\frac{1}{n}\right). \quad (7)$$

Straightforward but lengthy computations also prove (7) in the general case.

Remark 1. Pearson's inequality on the skewness κ_3 and the kurtosis κ_4 of a random variable states (cf. Klaassen, Mokveld and Van Es (2000))

$$\kappa_3^2 - \kappa_4 \le 2. \quad (8)$$

For the coefficients c_1 and c_2 in the Edgeworth expansion (1) of the sample mean, this means

$$c_2 \ge \frac{3}{2}c_1^2 - \frac{1}{12}. \quad (9)$$

However, this restriction is not essential and (6) holds for all values of c_1 and c_2.

Most asymptotically normal statistics have an Edgeworth expansion of type (1). This holds for L-statistics (Helmers (1980)), R-statistics (Does (1983)), and U-statistics (Bickel, Götze, and Van Zwet (1986)). Consequently, for many estimators, sample splitting has no effect asymptotically to third order. At least this suggests that in semiparametrics sample splitting is a natural option.

To conclude I would like to stress that I think this is a very nice paper, which gives a clear perspective also for future research and which presents an appropriate generalization of asymptotic linearity of estimators from i.i.d. non- and semiparametric models to more general non-i.i.d. models. Therefore, my comments have focused on rather minor issues. Nevertheless, I am very interested in the opinion of the authors on them.

Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands.

E-mail: chrisk@science.uva.nl

COMMENTS

Mark J. van der Laan and Zhuo Yu

University of California, Berkeley

Let us first compliment the authors on their nice paper on calculation of the efficient influence function in semiparametric models. The authors provide a framework for LAN-semiparametric models embedded in a nonparametric model, which allows one to calculate the efficient influence function as the projection of the nonparametric efficient influence function onto the tangent space. This generalizes the i.i.d.-result which says that the projection of any influence function onto the tangent space equals the efficient influence function and thus that an influence function which is an element of the tangent space must be equal to the efficient influence function.

The current frontiers "locally efficient estimation" and its relation to "estimation of non-smooth parameters" in the i.i.d. case, which the authors discussed, drew our particular attention.

Why locally efficient estimation?

If an estimator of a regular parameter is a reasonably smooth functional of the empirical distribution P_n, then its first order linear approximation (i.e., the empirical mean of its influence function) represents its finite sample behavior. It is clear that efficiency of an estimator and smoothness of the estimator as a functional of the empirical distribution P_n are typically tradeoffs, so that it is no surprise that in many models the maximum likelihood estimator suffers from the curse of dimensionality (i.e., lack of smoothness) while many practical good estimators are available. In fact, for most current data sets and their parameters of interest (e.g., causal inference, censored data), maximum likelihood estimation is a too restrictive methodology.

For example, suppose one observes right-censored data on a survival time T and a 25-dimensional covariate W, where each of the 25 components of W is discrete with 20 possible outcomes. Let F_T be the parameter of interest. Then the outcome space of W has 20^{25} values w_j. In this case, the maximum likelihood estimator of $F_{T|W}(\cdot \mid w_j)$ is the Kaplan-Meier estimate based on the subsample of subjects with $W_i = w_j$. Therefore one needs a sample size of the order of 20^{25} observations in order to have that the MLE of $F_T(t)$ has a reasonable practical performance. In this case the curse of dimensionality causes a miserable practical performance for any practical sample size. On the other

hand, Robins and Rotnitzky (1992) develop practical estimators which are locally efficient at a user supplied submodel for the right-censored data structure ($\tilde{T} \equiv \min(T, C), I(\tilde{T} \leq T), \bar{W}(\tilde{T})$) with a time-dependent covariate process $\bar{W}(\tilde{T}) = (W(t) : t \leq \tilde{T})$. Below, we provide a low-dimensional two sample semiparametric model with a regression parameter β and unspecified marginal distribution, where the maximum likelihood estimator of β requires so much smoothing that repairing it will still only result in a non-practical estimator. In other words, the curse of dimensionality occurs with low-dimensional and high-dimensional data structures though it affects many more parameters in the latter case.

Data sets encountered in practice are nowadays typically high dimensional (e.g gene expression profiles, genetic profiles, etc.) and involve time-dependent covariate processes so that modeling the *complete* data generating distribution with a parametric model is an extremely challenging problem and, even if one succeeds, then maximum likelihood estimation requires maximizing a high dimensional surface with lots of local maxima and can therefore be a computational nightmare. Therefore semiparametric models are generally preferable to parametric models in these situations so that the estimating function methodology is essential.

Estimating functions

The theory of estimating functions provides a general methodology for construction of estimators, ranging from globally efficient (such as a maximum likelihood estimator) to locally efficient at a small parametric submodel, or just inefficient everywhere. We here provide a short overview. For an extensive description of the methods applied to censored data and causal inference data structures, we refer to the upcoming book of van der Laan and Robins (2001) "Unified methods for censored longitudinal data and causality".

Let $\mu = \mu(F_X) \in \mathbb{R}^k$ be a euclidean pathwise differentiable parameter of interest of the data generating distribution $F_X \in \mathcal{M}$, where $\mathcal{M}$ denotes the model for F_X. In general, the nuisance scores are given by the scores of parametric submodels F_ϵ for which μ does not locally vary, i.e., $\frac{d}{d\epsilon}\mu(F_\epsilon)|_{\epsilon=0} = 0$. The nuisance tangent space $T_{nuis}(F_X) \subset L_0^2(F_X)$ is now the closure of the linear space generated by these nuisance scores.

Consider a class of k-dimensional estimating functions $\{D_h(X \mid \mu, \rho) : h \in \mathcal{H}\}$ indexed by an index h ranging over a set $\mathcal{H}$. An estimating function is unbiased if

$$E_{F_X} D_h(X \mid \mu(F_X), \rho(F_X)) = 0 \text{ for all } F_X \in \mathcal{M}.$$

Suppose now that the estimating functions are an element of the orthogonal complement $T_{nuis}^{\perp}$ of the nuisance tangent space in the sense that, for all $h \in \mathcal{H}$,

$$D_h(\cdot \mid \mu(F_X), \rho(F_X)) \in T_{nuis}^{\perp}(F_X)^k \text{ at all } F_X \in \mathcal{M}. \tag{1}$$

Such a rich class of estimating functions can be derived by finding appropriate representations of $T_{nuis}^{\perp}(F_X)$, or equivalently of the space generated by all influence functions. One wants to choose $\mathcal{H}$ so that there exists a $h_{opt} = h_{opt}(F_X) \in \mathcal{H}$ with $D_{h_{opt}}(\cdot \mid \mu(F_X), \rho(F_X))$ equal to the efficient influence function of μ at F_X. For example, if β is the parameter of interest in a generalized linear regression model $E(Y \mid Z) = m(Z \mid \beta)$, then $T_{nuis}^{\perp} = \{h(Z)\epsilon(\beta) : h\} \cap L_0^2(F_X)$ which naturally implies a rich class of estimating functions $\{h(Z)\epsilon(\beta) : \sup_z \mid h(z) \mid < \infty\}$ which have no nuisance parameter ρ. Note that the unbiasedness of the optimal estimating function $D_{h_{opt}}(X \mid \mu, \rho)$ is protected against misspecification of the index h_{opt} so that one can construct locally efficient estimators by estimating h_{opt} according to a guessed low dimensional submodel.

A two sample semiparametric estimation problem

Suppose we observe n_0 i.i.d. observations of $X_0 \sim f_0$ and n_1 i.i.d. observations of $X_1 \sim f_1$, where f_0, f_1 are Lebesgue densities. Consider this as a sample of $n = n_0 + n_1$ i.i.d. observations (X_i, ξ_i), $i = 1, \ldots, n$, where $\xi_i \in \{0, 1\}$ indexes the 2 populations. Let $F_j(x) = P(X \leq x \mid \xi = j)$, $j = 0, 1$, be the corresponding distribution functions. For example, X_0 and X_1 might represent a measurement on a randomly drawn subject from a population of lung-cancer patients and a healthy population, respectively. We are concerned with estimation of parameters comparing F_0 and F_1 such as $\mu_1 - \mu_0$, where $\mu_j = EX_j$, $j = 0, 1$. In many applications n_0 is very small relative to n_1. In these situations it is beneficial to have a statistical framework which allows one to borrow information from the X_1-sample when estimating F_0. Dominici and Zeger (2001) consider a parametric model for the function $p \to F_1 F_0^{-1}(p)$, $p \in [0, 1]$, and develop a least squares estimation method. In van der Laan, Dominici and Zeger (2001), we propose to model the quantile-quantile function (as in the structural nested causal inference models of Robins) that maps the quantiles of F_0 into the quantiles of F_1:

$$F_1^{-1} F_0(q) = m(q \mid \beta), \tag{2}$$

and yields a simpler parametrization of the likelihood. Here $m(\cdot \mid \beta)$: $[a_0, b_0] \to [a_1, b_1]$ is a known increasing absolutely continuous function in q with range $[a_1, b_1] \equiv [F_1^{-1}(0), F_1^{-1}(1)]$, parametrized by a k-dimensional parameter β. Notice that $F_0(x) = P(X_0 \leq x) = P(X_1 \leq F_1^{-1} F_0(x)) = F_1(m(x \mid \beta))$ and that $m(X_0 \mid \beta) \sim F_1$. Assumption (2) defines a semiparametric model for the data generating distribution with infinite dimensional parameter the cumulative distribution function F_1 and finite dimensional parameter β.

Calculation of the orthogonal complement of the nuisane tangent space of β yields the following class of estimating functions for β (van der Laan, Dominici

and Zeger (2001)):

$$\{D_h(X, \xi \mid \beta) \equiv (\xi = 0)(1 - p)h(m(X \mid \beta)) - (\xi = 1)ph(X) : h\}, \tag{3}$$

where the index $h = (h_1, \ldots, h_k)$ can be any k-dimensional vector of real valued functions and $p = P(\xi = 1)$. (3) represents a class of unbiased estimating functions indexed by a user-supplied h which does not involve a nuisance parameter f_1. The efficient score for β is given by $D_{h_{opt}}(X, \xi \mid \beta)$, where

$$h_{opt}(X \mid f_1, \beta) = \frac{f_1'(x)}{f_1(x)} m^{(1)}(m^{-1}(x)) + \frac{m'^{,(1)}(m^{-1}(x))}{m'(m^{-1}(x))}.$$

Here m' denotes the derivative w.r.t. x, $m^{(1)}(x)$ denotes the k-dimensional vector of first derivatives w.r.t. β_j of $m(x \mid \beta)$, $j = 1, \ldots, k$, and $m'^{,(1)}(x)$ denotes the k-dimensional vector of first derivatives w.r.t. β_j, $j = 1, \ldots, k$, of $m'(x \mid \beta)$. In other words, the class of unbiased estimating functions (3) includes, in particular, the efficient score of β which is the optimal estimating function. One can estimate β with the solution

$$0 = \sum_{i=1}^{n} D_{h_{opt}(f_{1n}, \beta_n^0)}(Y_i \mid \beta),$$

where we assume that (f_{1n}, β_n^0) is the parametric maximum likelihood estimator of (f_1, β) assuming a certain parametric model for f_1. If the parametric model is correct, then the resulting estimator β_n will be efficient, while it remains consistent and asymptotically normally (CAN) distributed if (f_{1n}, β_n^0) converges to a wrong (f^*, β^*). In other words, β_n will be CAN and efficient at the guessed parametric submodel. Note that globally efficient estimation will require estimation of a derivative of the density f_1, which explains why the maximum likelihood estimator is inconsistent, and that a regularized maximum likelihood estimator is not a preferred route of estimation.

Protection against misspecification of the nuisance parameter

If the estimating functions have a nuisance parameter ρ (note that we treat h_{opt} as an index and not as a nuisance parameter) which is high-dimensional, then globally consistent estimation of ρ might still be too much to ask. The orthogonality (1) of the estimating functions implies that derivatives w.r.t to ρ along one-dimensional directions (as allowed by the model $\mathcal{M}$) are equal to zero, so that ad hoc consistent estimation of ρ does not affect the asymptotic performance of the solution μ_n of the estimating equation $0 = \sum_i D_h(X_i \mid \mu, \rho_n)$. Fortunately, the orthogonality can even provide protection of the consistency of μ_n against inconsistent estimation of ρ. For example, the following lemma shows

that if the nuisance parameter space for ρ is convex in the sense that ρ can be varied along lines $\epsilon F_1 + (1-\epsilon)F \in \mathcal{M}$, then it remains unbiased even when the nuisance parameter is misspecified.

Lemma 0.1. *Consider an estimating function $D(X \mid \mu, \rho)$ which satisfies (1). Assume that μ is pathwise differentiable at each $F \in \mathcal{M}$ along a class of one-dimensional models including nuisance score lines $F_\epsilon = \epsilon F_1 + (1-\epsilon)F \in \mathcal{M}$ indexed by a set of F_1's with 1) dF_1/dF and dF/dF_1 being uniformly bounded, 2) $d/d\epsilon \mu(F_\epsilon)|_{\epsilon=0} = 0$, and 3) $\{\rho(F_1) : F_1\}$ covers the whole parameter space $\{\rho(F) : F \in \mathcal{M}\}$. Then*

$$E_{F_X} D(X \mid \mu(F_X), \rho_1) = 0 \text{ for all } \rho_1 \in \{\rho(F) : F \in \mathcal{M}\}.$$

Proof. Let $F_1, F, \rho_1 = \rho(F_1), \rho = \rho(F)$ be as in the lemma. Then $F_{\epsilon,h} = \epsilon F_1 + (1-\epsilon)F$ is a one dimensional submodel of $\mathcal{M}$ with score $h = d(F_1 - F)/dF$. Since it is a nuisance score model we have $0 = \frac{d}{d\epsilon}\mu(F_{\epsilon,h})\Big|_{\epsilon=0}$. By the fact that μ is pathwise differentiable along $F_{\epsilon,h}$ at F, we have for any gradient $\ell(X \mid F)$:

$$0 = \int \ell(x \mid F) \frac{d(F_1 - F)}{dF} dF = \int \ell(x \mid F) dF_1(x).$$

Since a standardized version of $D(\cdot \mid \mu(F_X), \rho(F_X))$ is a gradient, this implies also that for any such pair F_1, F, $0 = \int D(x \mid \mu(F), \rho(F)) dF_1(x) = \int D(x \mid \mu(F_1), \rho(F)) dF_1(x)$, which proves the lemma by just interchanging the role of F_1 and F.

It is the protection of the unbiasedness of estimating functions against misspecification of nuisance parameters which allows locally efficient estimation, since one can estimate $\rho(F_X)$ according to a guessed submodel of $\mathcal{M}$ without losing the consistency of the corresponding estimator μ_n.

Double protection of estimating functions in censored data and causal inference models

Suppose now that the parameter of interest is still $\mu(F_X)$, but we only observe n i.i.d. observations of censored data $Y = \Phi(C, X) \sim P_{F_X,G}$, where $X \in F_X \in \mathcal{M}^{Full}$ and the conditional distribution $G(\cdot \mid X)$ of C, given X, is assumed to satisfy coarsening at random (Heitjan and Rubin (1991), Jacobsen and Keiding (1995), Gill, van der Laan and Robins (1997)). Since causal inference data structures are missing data problems where the full data is the collection of all potential outcomes, this also covers causal inference models (for a unified treatment we refer to van der Laan and Robins (2001)). In this model for the observed data Y, the orthogonal complement of the nuisance tangent space $T_{nuis}^{\perp}$

of μ is given by the range of $A_{F_X}\left\{A_G^\top A_{F_X}\right\}^- : T_{nuis}^{Full,\perp} \to L_0^2(P_{F_X,G})$ of the orthogonal complement $T_{nuis}^{F,\perp}$ of the nuisance tangent space of μ in the full-data model $\mathcal{M}^F$, where $A_{F_X} : L_0^2(F_X) \to L_0^2(P_{F_X,G})$ is the nonparametric score operator $A_{F_X}(h)(Y) = E(h(X) \mid Y)$ and $A_G^\top : L_0^2(P_{F_X,G}) \to L_0^2(F_X)$ is its adjoint $A_G^\top(V)(X) = E(V(Y) \mid X)$. Here we implicitly assume that $T_{nuis}^{Full,\perp}$ is in the range of the nonparametric information operator $I_{F,G} = A_G^\top A_F$, but this can be weakened by replacing $A_F I_{F,G}^{-1}(D)$ by the projection of $D(X)$ onto the closure of the range of the score operator. Therefore, given a class of full-data estimating functions $\{D_h(X \mid \mu, \rho) : h \in \mathcal{H}^F\}$ in the full-data model, the class of estimating functions for μ in the observed data model is given by

$$\left\{IC(Y \mid F, G, D_h(\cdot \mid \mu, \rho)) \equiv A_F I_{F,G}^- D_h(\cdot \mid \mu, \rho) : h \in \mathcal{H}^F\right\}.$$

These are estimating functions for μ with nuisance parameter (ρ, F, G). By applying Lemma 0.1. (for G it follows directly and for F one needs to note that the lemma can be applied for each choice of μ) we obtain protection against misspecification of G, given F_X, ρ, and misspecification of F_X, given G, ρ: if $E_{F_X} D(X) = 0$, then $E_{F_X,G} IC(Y \mid F_1, G_1, D_h) = 0$, if either $F_1 = F_X$ or $G_1 = G$. This double protection property (which has been noted by various authors) of the estimating function implies that the estimator μ_n corresponding with the estimating equation $0 = \sum_i IC(Y_i \mid F_n, G_n, D_{h_n}(\cdot \mid \mu, \rho_n))$, where F_n, G_n are estimated according to guessed submodels for F_X and G, will be consistent if at least one of the guessed submodels is correct, assuming that the full-data estimating function is asymptotically unbiased. In Gill, van der Laan and Robins (2000) and van der Laan and Robins (2001) it is shown how one can also construct locally efficient estimators based on the least squares representation of the efficient influence function, as provided in Bickel, Klaassen, Ritov and Wellner (1993).

Double protection w.r.t. non-convex parameters

Consider the semiparametric regression model:

$$T = m(A|\alpha) + g(Z) + \epsilon, \tag{4}$$

where m is a known function, say, $m(A|\alpha) = \alpha A$, $g(Z)$ is unspecified, the conditional distribution $H(A \mid Z)$ of A, given Z is unspecified, and $E(\epsilon|A, Z) = 0$. Note that in a study where a treatment A is randomly assigned to a subject based on covariates, this conditional distribution H would be known by design. This model was studied in Newey (1990) and in Robins, Mark and Newey (1993), who

proposed a variety of estimators. In Zhuo and van der Laan (2001) it is shown that the orthogonal complement of the nuisance tangent space is given by

$$T_{nuis}^{\perp} = \{D_h(X|\alpha, g, H) \equiv \{T - m(A|\alpha) - g(Z))\}\{h(A,Z) - E_H(h(A,Z)|Z)\} : h\},$$

and the optimal estimating function $D_{h_{opt}}(X \mid \alpha, g, H)$ (i.e., the efficient influence function) is presented in closed form. Consider this class of estimating functions $\{D_h(X \mid \alpha, g, H) : h\}$ for α with nuisance parameters g and $H(A \mid Z)$. Lemma 0.1. predicts protection of the unbiasedness property of these estimating functions against misspecification of H, but not against misspecification of g_1. However, the double protection property can be directly verified: $E_{F_X} D_h(X \mid \alpha, g_1, H_1) = 0$, if either $g_1 = g(F_X)$ or $H_1 = H(F_X)$. Where $g(F_X)$, $H(F_X)$ denote the true regression curve g and true conditional distribution of A, given Z. For example, if $H_{A|Z}$ is known by design, or it is known that A is independent of Z, then one can estimate $g(Z)$ in the estimating equation $0 = \sum_i D_{h_n}(X_i \mid \alpha, g_n, H_n)$ as if it is linear in Z (thus no smoothing required) without any risk of getting an inconsistent estimator of α. We refer to Zhuo and van der Laan (2001) for simulations addressing the performance of the corresponding locally efficient estimators relative to estimators proposed in Newey (1990) and in Robins, Mark and Newey (1993).

Estimation of non-smooth parameters

Consider n i.i.d. observations on the right-censored data structure $(\tilde{T} = (\min(T, C), \Delta = I(T \leq C), \bar{W}(\tilde{T}))$ on a survival time T. Suppose that the density f_T is the parameter of interest. As suggested in the paper under discussion, given a kernel k and bandwidth b, one could estimate $\mu_b = \int f_T(s) k((s-t)/b)/b ds$ with a locally efficient estimator $\mu_{b,n}$ (Robins and Rotnitzky (1992)) and estimate $f_T(t)$ with $\mu_{b_n,n}$, where b_n is a cross-validated bandwidth estimator. Let us compare this estimator with a smoothed Kaplan-Meier estimate and assume independent censoring, so that the latter estimator is consistent. It can be shown that the locally efficient estimators (under both correct and incorrect specification of a guessed submodel) are *asymptotically* equivalent with the smoothed Kaplan-Meier estimator, though these estimators are known to gain in efficiency relative to Kaplan-Meier for smooth parameters. In other words, it is not possible to asymptotically improve estimation of the density by using covariate information. Consider now n i.i.d. observations on the current status data structure $(C, I(T \leq C), \bar{W}(C) = (W(s) : s \leq C))$ on a failure (e.g., onset of tumor) time T. Suppose that the parameter of interest is $F_T(t) = P(T \leq t)$, which can be approximated by the smooth parameter $\mu_b = \int F_T(s) k((s-t)/b)/b ds$. In this case the locally efficient estimators $\mu_{b,n}$ of μ_b of van der Laan and Robins (1998) are

now also asymptotically exploiting the covariate information to gain in asymptotic efficiency relative to the smoothed marginal NPMLE, as shown in van der Vaart and van der Laan (2001). This difference in asymptotic use of covariate information for the two data structures raises a general interesting question.

Division of Biostatistics, University of California, Berkeley, School of Public Health Warren Hall #7360, Berkeley, California 94720-7360, U.S.A.
E-mail: laan@stat.berkeley.edu

COMMENTS

Brad McNeney and Jon A. Wellner

University of Simon Fraser and University of Washington

Bickel and Kwon present an interesting survey of recent work on semiparametric models and statistical methods for such models. They correctly note that the primary focus has been on the i.i.d. data case or in models formulated in terms of counting processes. After reviewing developments for the i.i.d. case and presenting five Questions (A-E), they focus on Question C, and in answer to this question they present an approach to extending current results from the i.i.d. theory to more general models.

Our discussion will first focus on the general themes and issues raised by Questions A-E, and then return to the specific Answer developed to Question C. First some comments on Questions A,B,D, and E.

Question A. How should bandwidth estimation be accomplished when "smoothing" is necessary to attain efficiency?

This question is difficult and deserves considerable further study. The study of "adaptive estimation" in nonparametric models has been underway over the past 5-8 years: see e.g., Birgé and Massart (1999), Barron, Birgé, and Massart (1997), Efromovich (1998), Spokoiny (1996), and Lepski and Spokoiny (1997). The approaches developed for nonparametric models need to be brought to bear on semiparametric models. Perhaps another way to put the question is as follows.

Question A′. In a semiparametric model, can we estimate the infinite-dimensional parameter "adaptively" (here we use the term in the sense of the nonparametric function estimation literature, and not in the sense of Bickel (1982) or

Bickel and Kwon Question A) and still efficiently estimate and test hypotheses about and form confidence intervals for the finite-dimensional parameter?

Bunea (2000) has started to address this question in the special case of partially linear regression models.

Question B. What are appropriate model selection criteria for estimation of the variances of estimates of Euclidean parameters in i.i.d. and non-i.i.d. settings.

We view this question as being closely connected with Question A: can we eat our efficiency/adaptivity cake and make valid inferences too? Of course additional complications will ensue in non-i.i.d. problems.

Question D. How can we test goodness-of-fit for a given semiparametric model? What are the appropriate (best?) model diagnostics?

Another side to this coin is the establishment of properties of efficient semiparametric estimators beyond the models for which they are derived. As an example, consider the study of the Cox partial likelihood estimators beyond the Cox model given by Lin and Wei (1989) and BKRW pages 330 - 335. More studies of this type are needed. In situations in which a whole class of consistent estimators is available for estimation of the Euclidean parameters of a model, trade-offs between robustness and efficiency should be considered.

Question E. What are the asymptotic behaviors of semiparametric (and nonparametric) Bayes estimators?

This is a very important question since many (most?!) current estimators for semiparametric models are being computed via Markov Chain Monte Carlo methods with virtually no understanding of their (frequentist) properties. As Bickel and Kwon note, there has been some recent progress by Wasserman (1998), Ghosal and van der Vaart (2000), (2001), and Shen and Wasserman (2001), but much remains to be done in this area. Moreover, questions concerning frequentist properties (such as consistency) and inference remain largely unexplored.

Question E raises the interesting issue of development of algorithms for frequentist methods (Non- and Semi-Parametric Maximum Likelihood estimators, Generalized Estimating Equations,...). While some work has been done in this direction (see e.g., Böhning (1986), (1995) and Jongbloed (1998)), there is a large scope for further development of fast, high-quality, scalable algorithms.

A further set of problems involves the examination of semiparametric models under "functional model" (or incidental parameter) hypotheses as well as the more usual "structural model" (or i.i.d.) hypotheses. See Murphy and van der Vaart (1996) for an interesting study of the particular case of (linear) errors-in-variables regression models, and see Pfanzagl (1993) and Strasser (1996) for discussion of some of the general issues.

Now we turn to Bickel and Kwon's Answer to Question C.

Their approach is based on finding a suitable replacement $\hat{b}^{(n)}$ for the empirical measure $\mathbb{P}_n$. This replacement has several key properties, outlined in their conditions (A1) – (A3), in common with the empirical measure. The authors discuss how these three simple properties allow a generalization of many i.i.d. results on asymptotic efficiency and, in particular, how one might characterize efficient estimators. They present several interesting applications of their formulation, such as d-sample models and real-valued Markov chains. Their discussion of the key conceptual ideas needed to extend the i.i.d. theory is very insightful.

Hájek-LeCam style convolution theorems, which form the basis of the familiar notion of asymptotic efficiency in Bickel, Klaassen, Ritov and Wellner (1993) (BKRW), are of course not restricted to i.i.d. data. Rather it is the study of sufficient conditions to satisfy the convolution theorem hypotheses of local asymptotic normality (LAN) and regularity of the parameter of interest that has met with the most success to date in the i.i.d. setting. Hence an alternate way to extend i.i.d. results to non-i.i.d. models is to generalize these sufficient conditions. This is the approach taken in McNeney and Wellner (2000). In what follows we will comment on some of the connections between the two approaches.

In Definition 1 Bickel and Kwon define regular 1-dimensional submodels and describe the form of the local approximations to the log-likelihood ratio. The sequence of functionals $\{b_n^*\}$ that appears in the leading term in the expansion is connected to a tangent $\dot{l}$ via an operator T. This leads to a definition of the tangent set $\dot{\mathcal{P}}^0$ and the tangent space $\dot{\mathcal{P}}$ as the closed linear span of $\dot{\mathcal{P}}^0$. In general parameter spaces, linearity of $\dot{\mathcal{P}}^0$ is required (cf. Assumption (iii) of Theorem 5.2.1 in BKRW). Is this linearity of the tangent set implicitly assumed here?

The form of the approximation for the log-likelihood ratios is the basic building block for a study of efficiency. In McNeney and Wellner (2000), this is also the starting point. We chose a formulation with a more traditional triangular array of observations $\{X_{nk};\, k = 1, \ldots, m_n;\, n = 1, 2, \ldots\}$ and our approximation is in terms of a Martingale difference array $\{h_{nk} - E[h_{nk} \mid X_{n1}, \ldots, X_{n,k-1}];\, k = 1, \ldots, m_n;\, n = 1, 2, \ldots\}$. The Martingale difference array conveniently provides the asymptotic normality needed in the local approximation, and we assume a connection between the h_{nk}'s and a corresponding element h of the tangent space. We do not actually require the tangent space to be closed although in our discussion of asymptotic linearity and regularity of estimators, following Bickel (1993), we do define a "largest model of interest" and assume the tangent space for the largest model of interest is closed. As in Definition 2 of the present paper, asymptotically linear estimators are those which satisfy an expansion of the same form as the expansion of the log-likelihood ratios in a submodel of the largest model of interest.

The authors technical description is for estimation of regular 1-dimensional parameters, but they note that extensions to Banach-valued parameters follow as well. However, in considering Banach-valued parameters, there are several possible notions of regularity and asymptotic linearity of estimators; see for instance Definition 5.2.5 of BKRW regarding asymptotic linearity. By working exclusively with estimates of general parameters as collections of estimates of 1-dimensional parameters, it appears that the authors will adopt the weak versions of these definitions – weak regularity of parameters, weakly asymptotically linear estimators, and a resulting weak form of efficiency wherein an estimator T_n of a general parameter $b \in \mathcal{B}$ is efficient if the limiting distribution of $b^*(\sqrt{n}(T_n - b))$ is as concentrated as possible. This form of optimality does not, however, translate into the stronger conclusion that the asymptotic distribution of $\sqrt{n}(T_n - b)$ is optimal. It does not even imply $\{T_n\}$ is consistent. See the discussion of Example 1 in Section 4 of McNeney and Wellner (2000, p.464), for an example.

In their Example 1 discussing $d-$sample problems, the authors derive the same tangent space for the case of fixed covariates as in a randomized version that produces i.i.d. data, so that information bounds are the same in the two problems. Note that this is more generally true in that information bounds will be the same in two problems whenever the two tangent spaces are isomorphic as Hilbert spaces. See Corollary 4.4 of McNeney and Wellner (2000).

Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, B.C. V5A 1S6, Canada.

E-mail: mcneney@stat.sfu.ca

Department of Statistics, University of Washington Box 354322, Seattle, WA 98195, U.S.A.

E-mail: jaw@stat.washington.edu

COMMENTS

James M. Robins[1] and Andrea Rotnitzky[1,2]

[1]*Harvard School of Public Health and* [2]*Di Tella University*

We thank the editors for giving us this opportunity to discuss Bickel and Kwon's stimulating article and to give our perspective on the future of semiparametric inference. We found Bickel and Kwon's extension of methods for calculating information bounds to non i.i.d. models both enlightening and novel, and their list of five open questions relevant and challenging. In this discussion

we would like to pose a sixth question and to describe some initial attempts at an answer. Specifically we will discuss the question of how to approach the estimation of a finite dimensional parameter θ in very large semi parametric models, like those studied in Ritov and Bickel (1992), in which the semiparametric variance bound for $n^{1/2}$-consistent estimators of θ is finite (i.e., θ is a regular parameter) and yet θ is not estimable at rate n^{α} for any $\alpha > 0$. Robins and Ritov (1997) have argued that the study of these models is of major importance because the asymptotic behavior of an estimator in these very large models accurately mimics the finite sample behavior of estimators in the high dimensional models typically used in biomedical applications. Here we argue, following Scharfstein, Rotnitzky and Robins (1999) and Robins Rotnitzky and Van der Laan (2000), that in such large models one promising partial answer to our question is to employ so called doubly robust (DR), equivalently doubly protected, estimators when such estimators exist. Section 1 of our discussion will serve as motivation for and an introduction to DR estimation. Section 2-4 will summarize the current state of knowledge. Section 3 also outlines an approach to DR estimation of non-regular parameters. A discussion and bibliographic history of DR estimation concludes.

1. Motivation

Consider a follow-up study with data on outcome Y, a binary treatment R, and a high-dimensional vector of potential confounding factors V, many of which are continuous, such as age, red blood count, white blood count, liver function tests and weight. In realistic epidemiologic studies it would not be unusual for the sample size n to be between 500 and 2000 and yet for V to be 50-100 dimensional. Because V is high-dimensional and continuous, neither nonparametric smoothing nor stratification can be used for confounder control. As a consequence, statistical models are required for dimension reduction. Typically this involves regressing the outcome on the treatment and the confounders using linear, logistic, or log linear models.

For example if Y were continuous, we might choose to fit by ordinary least squares (OLS) the linear outcome regression (OR) model

$$E(Y \mid R, V) = \beta_0 + \beta^{'} V + \theta R$$

owing to the infeasibility of fitting the semiparametric regression (SR) model $E(Y \mid R, V) = \omega(V) + \theta R$ by multivariate non-parametric (e.g. kernel) smoothing, where $\omega(V)$ is an unknown arbitrary function. In the absence of measurement error or confounding by unmeasured factors, the parameter θ of the SR model represents the treatment effect. Even if, as we assume, the SR model assumption of no treatment-covariate interaction is correct, the OLS estimate $\widehat{\theta}_{OR}$ from the OR model may be badly biased if $\omega(V)$ cannot be well approximated

by $\beta_0 + \beta' V$. In particular, if the nonlinear part of $\omega(V)$ is highly correlated with R and highly predictive of Y, then $\widehat{\theta}_{OR}$ will be badly biased, even though the estimated regression function $\widehat{\beta}'_{OLS,0} + \widehat{\beta}'_{OLS} V + \widehat{\theta}_{OR} R$ may be highly predictive of the response Y and the power of standard global lack of fit tests may be small. Partially non/semiparametric dimension reducing techniques such as generalized additive models may improve somewhat upon a linear regression model but cannot solve the dimensionality problem. For example, GAM models ignore interactions among components of V.

Recently alternative methods of confounder control based on an estimated propensity scores have been introduced. The propensity score $P \equiv pr(R = 1 \mid V)$ is the conditional probability of exposure given the covariates (Rosenbaum and Rubin (1983)). Because P is unknown, and the fitting of the nonparametric logistic regression model $\texttt{logit}\ pr(R = 1 \mid V) = \gamma(V)$, with $\gamma(V)$ an unknown unrestricted function is infeasible, we might choose to estimate P by the predicted value $\widehat{P} = \texttt{expit}\left(\widehat{\alpha}_0 + \widehat{\alpha}' V\right)$ from the maximum likelihood fit of a linear logistic model

$$\texttt{logit}\ pr(R = 1 \mid V) = \alpha_0 + \alpha' V,$$

Here $\texttt{logit}\ x = \ln\{x/(1-x)\}$ and $\texttt{expit}(x) = \{1 + \exp(-x)\}^{-1}$. A suitable propensity score estimator $\widehat{\theta}_P$ of θ turns out to be the estimator of θ in the OLS fit of the model $E(Y \mid R, V) = \beta_0 + \theta R + \varsigma \widehat{P}$ (Robins (2000)).

There has been considerable debate as to which approach to confounder control is to be preferred, as the first is biased if the outcome regression model is misspecified while the second approach is biased if the treatment regression, i.e., propensity, model is misspecified. This controversy could be resolved if an estimator were available that was guaranteed to be consistent for θ whenever at least one of the two models was correct under an asymptotic sequence in which the outcome and treatment regression models remain fixed as the sample size n increases to infinity. We refer to such combined methods as doubly-robust or doubly-protected as they can protect against misspecification of either the outcome or treatment model, although not against simultaneous misspecification of both.

A natural first guess that turns out to be correct is that the OLS estimator $\widehat{\theta}_{DR}$ based on an expanded model $E(Y \mid R, V) = \beta_0 + \beta' V + \theta R + \varsigma \widehat{P}$ that adds $\widehat{P}$ as a regressor is doubly robust. One could wonder about the actual advantage of using DR estimators as, in practice, all models including the outcome and treatment regression models are misspecified and thus even the DR estimator of θ may be considerably biased. In our opinion, a DR estimator has the following advantage that argues for its routine use: if either the model for the outcome or the model for the propensity score is nearly correct, then the bias of a DR estimator of θ will be small. Thus, the DR estimator $\widehat{\theta}_{DR}$, in contrast with

both the usual outcome regression estimator $\widehat{\theta}_{OR}$ or the propensity estimator $\widehat{\theta}_P$, gives the analyst two chances, instead of only one, to get nearly correct inference about the treatment effect. Of course, there can be an efficiency cost to using a DR estimator rather than the outcome regression estimator of θ: if the outcome regression model is correct both the $\widehat{\theta}_{DR}$ and $\widehat{\theta}_{OR}$ will be consistent but the DR estimator will be less efficient. However, in our opinion, we have already paid homage to the need for efficiency by using parametric, albeit high dimensional, models in the outcome and treatment regressions; at this juncture the hope to control bias due to model misspecification with DR estimators trumps further efficiency concerns.

A further advantage of DR estimation is that comparison of the three estimators $\widehat{\theta}_{DR}$, $\widehat{\theta}_P$, and $\widehat{\theta}_{OR}$ with one another serves as a useful informal goodness of fit test. Specifically if the DR estimator differs from both the propensity and outcome regression estimator by much more than can be explained by sampling variation (say, as evaluated using the bootstrap) then we can conclude that both the propensity and outcome regression model must have been badly misspecified and that all three estimators probably suffer from substantial bias. In that event the specification of both the propensity and outcome regression model should be modified, say by adding additional nonlinear and interaction terms to the model. If $\widehat{\theta}_{DR}$ and $\widehat{\theta}_P$ are close but differ greatly from $\widehat{\theta}_{OR}$, one can take that as some evidence that the propensity model may be nearly correct, that the outcome regression model is probably badly misspecified, and that $\widehat{\theta}_{DR}$ and $\widehat{\theta}_P$ may suffer from only a small amount of bias. Similar remarks apply with the roles of $\widehat{\theta}_{OR}$ and $\widehat{\theta}_P$ reversed. This informal goodness of fit test is based directly on estimators of the parameter θ of interest and thus will presumably be both more sensitive and inferentially relevant than global goodness of fit tests of the outcome and propensity regression models themselves.

Doubly robust estimators do not always exist, and even when they do, their construction may not always be obvious. As an example, suppose that in our motivating problem, either the outcome Y is Bernoulli and we fit a linear logistic outcome model $\texttt{logit}E(Y \mid R, V) = \beta_0 + \beta^{'}V + \theta R$ or the outcome Y is a count variable and we fit the log linear outcome regression model $\log E(Y \mid R, V) = \beta_0 + \beta^{'}V + \theta R$. In both cases the iteratively reweighted least squares (IRLS) estimator of θ (i.e., the ML estimator under the Bernoulli and Poisson likelihoods respectively) obtained by adding the term $\varsigma\widehat{P}$ to the model is, in contrast with the linear regression model, inconsistent whenever the outcome regression model is misspecified and the true value of θ is non-zero, even if the propensity model is correct. Indeed, as we discuss in Sections 2 and 3, (i) no DR estimator exists for the linear logistic model and (ii) a DR estimator exists in the log linear model but it is not constructed by adding functions of $\widehat{P}$ to a log linear regression model.

In Sections 2-4 we use the semiparametric theory developed by Bickel and others to provide some preliminary answers to the question of existence and construction of DR estimators.

2. The Formal Problem and Doubly Robust Estimating Functions

To formalize our problem we consider inference about a possibly vector valued functional $\theta \equiv \theta(\rho)$ under a model $M(\mathcal{R})$ indexed by an infinite dimensional parameters $\rho \in \mathcal{R}$ for n i.i.d. copies of a random vector X. We are interested in settings in which the parameter space $\mathcal{R}$ is very large and inference about θ is practically unfeasible due to the curse of dimensionality. Specifically, following Ritov and Bickel (1992), Robins and Ritov (1997) and Robins, Rotnitzky and van der Laan (2000), we consider models $M(\mathcal{R})$ which have the following properties: (i) the semiparametric variance bound for $n^{1/2}$-consistent estimators of θ is finite at all $\rho \in \mathcal{R}$, and yet, no estimator is consistent for θ uniformly over $\rho \in \mathcal{R}$, much less uniformly asymptotically normal (UAN); (ii) no estimator of θ attains a pointwise (i.e., non-uniform) rate of convergence of n^{α} at all $\rho \in \mathcal{R}$ for any $\alpha > 0$; (iii) there does not exist a regular asymptotically linear estimator (RAL) of θ at any $\rho \in \mathcal{R}$. In this setting in both theory and practice some method of dimension reduction is necessary by imposing additional modelling restrictions. One dimension reduction strategy often used in practice is to introduce a parametrization (κ, γ), $\kappa \in \mathcal{K}$ and $\gamma \in \Gamma$, of ρ with κ and γ variation independent, i.e., $\mathcal{R} = \mathcal{K} \times \Gamma$ and replace model $M(\mathcal{R})$ by either a working submodel $M(\mathcal{K}_{sub} \times \Gamma)$ or a working submodel $M(\mathcal{K} \times \Gamma_{sub})$ where $\mathcal{K}_{sub} \subset \mathcal{K}$ and $\Gamma_{sub} \subset \Gamma$, and hope that RAL estimators can be found in one or both of the working submodels. However, because $\mathcal{K}_{sub}$ and Γ_{sub} are only working submodels, it is unknown whether the true value of γ is in Γ_{sub} or the true value of κ is in $\mathcal{K}_{sub}$. Thus, the best that can be hoped for is an estimator that is RAL in the union model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$ that assumes that the true value of ρ either lies in $\mathcal{K} \times \Gamma_{sub}$ or in $\mathcal{K}_{sub} \times \Gamma$. We refer to such an estimator as doubly robust or doubly protected under the parametrization $\rho = (\kappa, \gamma)$ and submodels Γ_{sub} and $\mathcal{K}_{sub}$.

The ultimate goal would be to characterize necessary and sufficient conditions for the existence of a DR estimators and, where they exist, provide constructive methods for finding them. In this discussion, we summarize current progress. Before studying the union model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$ of interest, it will be advantageous to study unbiased estimating functions in the special case $M(\kappa \times \Gamma) \cup M(\mathcal{K} \times \gamma)$ in which $\mathcal{K}_{sub}$ and Γ_{sub} are singletons.

Definition 1. We say that a function $U(\theta, \kappa, \gamma) \equiv u(X, \theta, \kappa, \gamma)$ is a DR estimating function for a, possibly vector valued, functional $\theta(\kappa, \gamma)$ in model

$M(\mathcal{K}\times\Gamma)$ under parametrization (κ,γ) if it is an unbiased estimating function in the union model $M(\kappa\times\Gamma)\cup M(\mathcal{K}\times\gamma)$. That is for all (κ,γ) and $(\kappa^*,\gamma^*)\in\mathcal{K}\times\Gamma$, $E_{\kappa^*,\gamma^*}\left[U\left(\theta\left(\kappa^*,\gamma^*\right),\kappa^*,\gamma\right)\right]=E_{\kappa^*,\gamma^*}\left[U\left(\theta\left(\kappa^*,\gamma^*\right),\kappa,\gamma^*\right)\right]=0$ and $\partial E_{\kappa^*,\gamma^*}\left[U\left(\theta,\kappa,\gamma^*\right)\right]/\partial\theta\Big|_{\theta=\theta(\kappa^*,\gamma^*)}$ and $\partial E_{\kappa^*,\gamma^*}\left[U\left(\theta,\kappa^*,\gamma\right)\right]/\partial\theta\Big|_{\theta=\theta(\kappa^*,\gamma^*)}$ are invertible.

Henceforth, we always assume invertibility of $\partial E_{\kappa^*,\gamma^*}[U(\theta,\kappa,\gamma^*)]/\partial\theta\Big|_{\theta=\theta(\kappa^*,\gamma^*)}$ and $\partial E_{\kappa^*,\gamma^*}[U(\theta,\kappa^*,\gamma)]/\partial\theta\Big|_{\theta=\theta(\kappa^*,\gamma^*)}$. Clearly, a necessary (but not sufficient) condition for the existence of a doubly robust estimating function is that there is an unbiased estimating function $U_1(\theta,\gamma)$ for $\theta(\kappa,\gamma)$ in model $M(\mathcal{K}\times\gamma)$ and an unbiased estimating function $U_2(\theta,\kappa)$ for $\theta(\kappa,\gamma)$ in model $M(\kappa\times\Gamma)$. Further progress requires semiparametric theory definitions.

Given an arbitrary semiparametric model $M(\Psi_1\times\Psi_2)$ indexed by variation independent, possibly infinite dimensional parameters, ψ_1 and ψ_2, and a, possibly p-dimensional vector valued, functional $\theta(\psi)$ where $\psi=(\psi_1,\psi_2)$, let $\mathcal{L}_2^0(\psi)$ be the Hilbert space of random vectors of the dimension of θ with mean zero and covariance inner product under ψ. Let $\Lambda_{\Psi_j}(\psi)\subset\mathcal{L}_2^0(\psi)$ and $\Lambda_{\Psi}(\psi)\subset\mathcal{L}_2^0(\psi)$ be the tangent spaces (i.e., closed linear span of scores) for ψ_j, $j=1,2$, and for ψ, respectively, when the data is generated under ψ and let $\Lambda_{\Psi_j}^{\perp}(\psi),j=1,2$, and $\Lambda_{\Psi}^{\perp}(\psi)$ denote their orthogonal complements in $\mathcal{L}_2^0(\psi)$. Finally, in any semiparametric model $M(\Psi)$ indexed by ψ, let $IF(\psi)$ denote the influence function space for θ at ψ. That is, $IF(\psi)$ is the direct sum of $\Lambda_{\Psi}^{\perp}(\psi)$ and the linear space spanned by the efficient influence function EIF (ψ) for θ. In many models $IF(\psi)$ is called the orthogonal complement to the nuisance tangent space for θ. To make our discussion concrete we use two models $M(\mathcal{K}\times\Gamma)$ to illustrate our results. In the first, $\theta(\kappa,\gamma)$ may be a function of both κ and γ. In the second, $\theta(\kappa,\gamma)=\theta(\kappa)$ only depends on κ. When $\theta(\kappa,\gamma)=\theta(\kappa)$, we define $IF(\kappa,\gamma)$ in model $M(\kappa\times\Gamma)$ to be $\Lambda_{\Gamma}^{\perp}(\kappa,\gamma)$.

Example 1. A Partially Missing Response Model

Suppose we have a model with underlying full data (R,Y,V), Y and R Bernoulli and V highly multivariate and continuous. The parameter of interest θ is the mean of Y. However, if $R=1$ then Y is not observed. Thus $X=(R,V,RY)$ is observed. Scharfstein, Rotnitzky and Robins (1999) consider the model $M(\mathcal{K}\times\Gamma)$ for X that imposes the sole assumption that

$$pr(R=1|Y,V;\gamma)=\phi\{\gamma(V)+\alpha Y\}\equiv\Phi(\gamma) \tag{1}$$

where α is a known selection bias parameter, $\phi(\cdot)$ is a known differentiable, strictly increasing, cumulative distribution function with support on $(-\infty,\infty)$, and $\Gamma=\{\gamma=\gamma(\cdot)\}$ is the set of all functions of V. When $\alpha=0$, the

data are said to be coarsened at random (CAR) and the missingness is said to be ignorable. When $\alpha \neq 0$, missingness is said to be nonignorable. Write $pr(Y = 1|V, R = 1; \omega) = \omega(V)$ and let $\Omega = \{\omega = \omega(\cdot); 0 < \omega(V) < 1\}$ be the set of all integrable functions taking values in $(0, 1)$, so ω is the conditional mean function of Y given V in the subpopulation of units with Y observed. At $\alpha = 0$, ω is also the conditional mean function for the entire population. Let $\mathcal{N} = \{\eta = \eta(\cdot)\}$ be the set of all densities for V and let $\mathcal{K} = \mathcal{N} \times \Omega$. Robins and Rotnitzky (RR) (2001a) show that κ and γ are variation independent, i.e., their joint parameter space is the product space $\mathcal{K} \times \Gamma$. The individual likelihood contribution is $\mathcal{L}(\kappa, \gamma) = \mathcal{L}_1(\kappa)\mathcal{L}_2(\kappa, \gamma)$ where $\mathcal{L}_1(\kappa) = \eta(V)[\omega(V)^Y\{1 - \omega(V)\}^{1-Y}]^R$ and $\mathcal{L}_2(\kappa, \gamma) = E_{\omega,\gamma}(R|V)^R\{1 - E_{\omega,\gamma}(R|V)\}^{1-R}$. RR (2001a) show that $E_{\omega,\gamma}(R|V) = E_\omega\{\Phi(\gamma)^{-1}|V, R = 1\}^{-1}$. Under CAR, i.e., under $\alpha = 0$, $\theta(\kappa, \gamma) = \theta(\kappa), E_{\omega,\gamma}(R|V) = E_\gamma(R|V)$ does not depend on ω and $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$ factors into a function of κ only and a function of γ only. Rotnitzky, Robins and Scharfstein (1998) showed that model $M(\mathcal{K} \times \Gamma)$ is a non-parametric model for the law F_X of the observed data and that the joint law $F_{Y,R,V}$ is identified. In particular, the marginal mean of Y, $E_{\kappa,\gamma}(Y)$ is $\theta(\kappa, \gamma) = E_\eta[E_\omega\{Y/\Phi(\gamma)|V, R = 1\}/E_\omega\{1/\Phi(\gamma)|V, R = 1\}]$.

RR (2001a) show that the efficient influence function for $\theta(\kappa, \gamma)$ is $S_{eff}(\kappa, \gamma, \theta(\kappa, \gamma))$ where

$$S_{eff}(\kappa, \gamma, \theta) = \frac{R}{\Phi(\gamma)}\left(Y - \frac{E_\omega\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}Y|R = 1, V\right\}}{E_\omega\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}|R = 1, V\right\}}\right) + \frac{E_\omega\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}Y|R = 1, V\right\}}{E_\omega\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}|R = 1, V\right\}} - \theta,$$

and the influence function space for θ in model $M(\mathcal{K} \times \Gamma)$ is thus $\{cS_{eff}(\kappa, \gamma, \theta(\kappa, \gamma)); c \in \mathbf{R}\}$.

We later use Lemma 1 below to show that a DR estimating function exists for the parametrization (κ, γ) if and only if the ratio $B(\omega, \gamma) = E_\omega\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}Y|R = 1, V\right\}$ $/E_\omega\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}|R = 1, V\right\}$ is free of γ. RR (2001a) proved that this ratio is free of γ if and only either (i) $\alpha = 0$ and thus there is CAR or, (ii) the known CDF $\phi(\cdot)$ satisfies

$$\phi(x) = \frac{\exp\{k\alpha[x/\alpha] + q(x - \alpha[x/\alpha])\}}{1 + \exp\{k\alpha[x/\alpha] + q(x - \alpha[x/\alpha])\}}, \tag{2}$$

where k is any positive constant, $[x]$ is the largest integer less than or equal to x, and $q(u)$ is any increasing and differentiable function on $[0, \alpha)$ such that $k\alpha[x/\alpha] + q(x - \alpha[x/\alpha])$ is differentiable. The choice $q(u) = u$ and $k = 1$ gives the logistic cumulative distribution function $\phi(x) = \exp(x)/\{1 + \exp(x)\}$. However the logistic CDF is not the only possible choice for $q(u)$. For example, $q(u) = I_{[0,1/2)}(u)\, 2u^2 + I_{(1/2,1]}(u)\{1 - 2(u - 1)^2\}$ is a valid choice for $\alpha = 1$ and $k = 1$.

When (i) or (ii) hold, it is easy to check that $B(\omega,\gamma) = E_\omega\{e^{-k\alpha Y}Y|R=1,V\}/E_\omega\{e^{-k\alpha Y}|R=1,V\}$ and $\{cS_{eff}(\kappa,\gamma,\theta); c\in\mathbf{R}\}$ is a set of DR estimating functions. Robins and Rotnitzky (2001a) use Lemma 1 below to prove that this set contains all DR estimating functions.

A necessary condition for the existence of a DR estimating function is the existence of both an unbiased estimating function for θ with κ known and another with γ known. A natural question is whether there exist simple primitive sufficient conditions for the existence of these estimating functions. Example 1 suggests not as the sufficient condition (ii) is very complex and nonintuitive.

Example 2. Semiparametric Regression

Model $M(\mathcal{K}\times\Gamma)$ for $X=(R,V,Y)$ imposes the sole assumption that $\phi^{-1}\{E(Y|R,V)\} = \theta R+\omega(V)$, with Y and R being continuous, count, or dichotomous outcome and treatment variables, V a highly multivariate continuous random vector with support in $\mathcal{V}$, $\phi^{-1}(x)$ a known 1-1 link function with range $(-\infty,\infty)$, $\Omega = \{\omega=\omega(\cdot)\}$ the set of all functions of V. To avoid distracting technicalities, we will additionally impose the assumption that $f(V)$ is known. This we do without loss of generality because the influence function space for θ is the same whether $f(V)$ is known or unknown. Let $\epsilon(\theta,\omega)$ denote $Y-\phi\{\theta R+\omega(V)\} \equiv Y-\Phi\{\theta,\omega\}$. Then $\mathcal{K}=\{\kappa=(\theta,\omega,\eta);\theta\in R^1,\ \omega\in\Omega,\ \eta\in\mathcal{N}(\theta,\omega)\}$, where $\mathcal{N}(\theta,\omega)$ is the set of all mean zero conditional densities for $\epsilon(\theta,\omega)$ given V (except when Y is binary in which case $\kappa=(\theta,\omega)$). We define $\Gamma=\{\gamma\equiv\gamma(R|V)\}$ to be the set of all conditional densities for R given V. The individual likelihood contribution factors as $\mathcal{L}(\kappa,\gamma)=\mathcal{L}_1(\kappa)\mathcal{L}_2(\gamma)$, where $\mathcal{L}_1(\kappa)=\eta(\epsilon(\theta,\omega)|V)$ and $\mathcal{L}_2(\gamma)=\gamma(R|V)$.

Bickel, Klaassen, Ritov and Wellner (1993) and RR (2001b) show that the influence function space for θ is

$$IF(\kappa,\gamma)=\{U(\kappa,\gamma,g,\phi)=\epsilon(\theta(\kappa),\omega(\kappa))\{g(R,V)-M(\gamma,\kappa,g,\phi)\}; g\in\mathcal{G}\},$$

with $\mathcal{G}$ the set of all functions of (R,V), and $M(\gamma,\kappa,g,\phi) = E_\gamma\{g(R,V)|V\}$ if ϕ^{-1} is the identity link, $M(\gamma,\kappa,g,\phi) = E_\gamma\{g(R,V)e^{\theta(\kappa)R}|V\}/E_\gamma\{e^{\theta(\kappa)R}|V\}$ if ϕ^{-1} is the log link and $M(\gamma,\kappa,g,\phi) = \frac{E_\gamma[g(R,V)\Phi\{\theta(\kappa),\omega(\kappa)\}[1-\Phi\{\theta(\kappa),\omega(\kappa)\}]|V]}{E_\gamma[\Phi\{\theta(\kappa),\omega(\kappa)\}[1-\Phi\{\theta(\kappa),\omega(\kappa)\}]|V]}$ if ϕ^{-1} is the logit link. Throughout, we write $\theta(\kappa)$ and $\omega(\kappa)$ when we wish to emphasize that θ and ω are formally functions of κ. Note that $IF(\kappa,\gamma)$ does not depend on η, so without loss of generality we write it as $IF(\theta,\omega,\gamma)$. RR (2001b) have proved that the only links for which $M(\gamma,\kappa,g,\phi)$ depends on κ only through $\theta(\kappa)$ for all $g\in\mathcal{G}$ are the identity and exponential. It is straightforward to check that, for ϕ^{-1} the identity or the log-link, the set $\mathcal{D}=\{U(\theta,\kappa,\gamma,g,\phi)=\epsilon(\theta,\omega(\kappa))\{g(R,V)-M_{est}(\gamma,\theta,g,\phi)\}; g\in\mathcal{G}\}$ is comprised of DR estimating functions, where $M_{est}(\gamma,\theta,g,\phi) = E_\gamma\{g(R,V)|V\}$ if ϕ^{-1} is

the identity link and $M_{est}(\gamma, \theta, g, \phi) = E_\gamma\{g(R,V)e^{\theta R}|V\}/E_\gamma\{e^{\theta R}|V\}$ if ϕ^{-1} is the log link. Note $U(\theta, \kappa, \gamma, g, \phi)$ is $U(\kappa, \gamma, g, \phi)$ with $\theta(\kappa)$ replaced by the free parameter θ.

We next summarize results in RR (2001a) and apply them to prove that the set $\mathcal{D}$ contains all DR estimating functions that depend on κ only through $\omega(\kappa)$ when ϕ^{-1} is either the identity or the log link, and that no DR estimating function exists when Y is binary and ϕ^{-1} is the logit link.

- A necessary condition for $U(\theta, \kappa, \gamma)$ to be DR is that $U(\theta(\kappa, \gamma), \kappa, \gamma)$ must be an element of the influence function space $IF(\kappa, \gamma)$ in the unrestricted model $M(\mathcal{K} \times \Gamma)$ at each (κ, γ).
- Suppose that $\theta(\kappa, \gamma) = \theta(\kappa)$ is a function of κ alone. Then *(i)* our model $M(\mathcal{R})$ will admit parameterizations $\mathcal{R} = \mathcal{K} \times \Gamma$ with $\mathcal{K} = \{\kappa = (\theta, \delta) : \theta \in \Theta,\ \theta(\kappa) = \theta \text{ and } \delta \in \Delta(\theta)\}$, where $\Delta(\theta)$ is a set that can possibly depend on θ, (ii) $IF(\kappa, \gamma)$ can be expressed as the set $IF(\theta, \delta, \gamma) = \left\{\widetilde{V}(\theta, \delta, \gamma) \equiv \widetilde{U}((\theta, \delta), \gamma); \widetilde{U}(\kappa, \gamma) \in IF(\kappa, \gamma)\right\}$ of functions of (θ, δ, γ), and (iii), by the previous remark, a necessary condition for an estimating function $U(\theta, \kappa, \gamma) = U(\theta, \delta, \gamma)$ that depends on κ only through δ to be DR is that it be an element of $IF(\theta, \delta, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$.

Example 2. (continuation). In Example 2, take $\delta = (\omega, \eta)$ and $\Delta(\theta) = \{(\omega, \eta) : \omega \in \Omega, \eta \in \mathcal{N}(\theta, \omega)\}$. Then since $\kappa = (\theta, \delta)$ with $\theta(\kappa, \gamma) = \theta(\kappa) = \theta$, any DR estimating function $U(\theta, \omega, \gamma)$ must be in the set $IF(\theta, \delta, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$. But as pointed out above, $IF(\theta, \delta, \gamma) = IF(\theta, \omega, \gamma)$ does not depend on η. Now, when ϕ^{-1} is the logit link, RR (2001a) showed that $E_{\theta,\omega,\gamma^*}[U(\theta, \omega, \gamma)] \neq 0$ for all $U(\theta, \omega, \gamma) \in IF(\theta, \omega, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$. Thus, no DR estimating function can exist when ϕ^{-1} is the logit link. When ϕ^{-1} is the identity or log link, we noted above that all elements $U(\theta, \omega, \gamma)$ of $IF(\theta, \omega, \gamma)$ are doubly robust. This proves that $IF(\theta, \omega, \gamma)$ is the set of all DR estimating functions $U(\theta, \kappa, \gamma) = U(\theta, \omega, \gamma)$ that depends on κ only through ω for ϕ^{-1} the identity or the log link.

When $\theta(\kappa, \gamma)$ depends on κ and γ the above strategy is not available. The following Lemma provides a sometimes useful way to prove the absence of DR estimating functions in this case.

Lemma 1. *A necessary condition for the existence of a doubly robust estimating function is that, for each θ in $\Theta(\gamma) \equiv \{\theta(\kappa, \gamma) : \kappa \in \mathcal{K}\}$, we have $\cap_{\kappa^*:\theta(\kappa^*,\gamma)=\theta} IF(\kappa^*, \gamma) \neq \varnothing$ in model $M(\mathcal{K} \times \gamma)$, and for each θ in $\Theta(\kappa) \equiv \{\theta(\kappa, \gamma) : \gamma \in \Gamma\}$, we have $\cap_{\gamma^*:\theta(\kappa,\gamma^*)=\theta} IF(\kappa, \gamma^*) \neq \varnothing$ in model $M(\kappa \times \Gamma)$.*

Informally, we interpret Lemma 1 as saying that in the model $M(\mathcal{K} \times \gamma)$ there exists an element of $IF(\kappa^*, \gamma)$ that depends on κ^* only through $\theta(\kappa^*, \gamma)$,

and in model $M(\kappa \times \Gamma)$ there is an element of $IF(\kappa, \gamma^*)$ that depends on γ^* only through $\theta(\kappa, \gamma^*)$.

Example 1. (continuation). In model $M(\kappa \times \Gamma)$, RR (2001a) prove that $IF(\kappa, \gamma) = \{S_{eff}(\kappa, \gamma, \theta(\kappa, \gamma)) + a(X); a \in \mathcal{A}(\kappa)\}$, where

$$\mathcal{A}(\kappa) = \Big\{ a(X) = Rs(Y,V) + (1-R)E_\omega[s(Y,V)|R=1,V] - \varphi(s;\kappa);\, s \text{ unrestricted} \Big\},$$

and $\varphi(s;\kappa) = E_\eta\{E_\omega[s(Y,V)|R=1,V]\}$. RR (2001a) show that in order for $\cap_{\gamma:\theta(\kappa,\gamma)=\theta} IF(\kappa,\gamma) \neq \emptyset$ it must be that $B(\omega,\gamma)$ does not depend on γ.

We have described ways to rule out DR estimating functions for $\theta(\kappa, \gamma)$ by checking necessary conditions for their existence. We now explore ways to rule in DR estimating function by finding further sufficient conditions for their existence. Consider the following.

Condition 1. $IF(\kappa^*, \gamma)$ in model $M(\mathcal{K} \times \gamma)$ depends on κ^* only through $\theta(\kappa^*, \gamma)$, and $IF(\kappa, \gamma^*)$ in model $M(\kappa \times \Gamma)$ depends on γ^* only through $\theta(\kappa, \gamma^*)$.

Condition 2. $\theta(\kappa, \gamma)$ is function of κ alone and
$\mathcal{K} = \{\kappa = (\theta, \delta) : \theta \in \Theta,\ \theta(\kappa, \gamma) = \theta,\ \text{and } \delta \in \Delta(\theta)\}$.

Condition 3. There exists a function $q(\cdot)$ such that $q(\theta)$ is a linear functional of both the law indexed by κ (with γ fixed) and the law indexed by γ (with κ fixed), in the sense that for some $U_1(\gamma)$ and $U_2(\kappa)$, $E_{\kappa\gamma}\{U_1(\gamma)\} = E_{\kappa\gamma}\{U_2(\kappa)\} = q\{\theta(\kappa, \gamma)\}$. The following Theorem, proved in RR (2001a), states that when Condition 1 and one of Conditions 2 or 3 hold, then there exist DR estimating equations. Indeed, the theorem shows how to construct them.

Theorem 1. (a) *If Conditions* 1 *and* 2 *hold, then all elements* $U(\theta, \delta, \gamma)$ *of* $IF(\theta, \delta, \gamma) = IF(\kappa, \gamma)$ *in model* $M(\mathcal{K} \times \Gamma)$ *are doubly robust;* (b) *if* Conditions 1 *and* 3 *hold,* $IF(\kappa, \gamma)$ *in model* $M(\mathcal{K} \times \Gamma)$ *has the form* $IF(\kappa, \gamma) = \{\widetilde{U}(\gamma) - q[\theta(\kappa, \gamma)]; \widetilde{U}(\gamma) \in \widetilde{\mathcal{U}}(\gamma)\}$, *where* $\widetilde{\mathcal{U}}(\gamma)$ *is the set of all random variables* $\widetilde{U}(\gamma)$ *for which* $\widetilde{U}(\gamma) - q[\theta(\kappa, \gamma)] \in IF(\kappa, \gamma)$ *in both* $M(\mathcal{K} \times \gamma)$ *and* $M(\kappa \times \Gamma)$. *Further the set* $\{\widetilde{U}(\gamma) - q[\theta]; \widetilde{U}(\gamma) \in \widetilde{\mathcal{U}}(\gamma)\}$ *consists of DR estimating functions.*

A necessary and sufficient condition for Condition 1 to hold is that in model $M(\kappa \times \Gamma)$ there exists an unbiased estimating function $U_2(\theta, \kappa)$ for θ and the orthogonal complement $\Lambda_\Gamma^\perp(\kappa, \gamma) = \Lambda_\Gamma^\perp(\kappa)$ to the tangent space for γ is the same for all (i.e., does not depend on) $\gamma \in \Gamma$, and in model $M(\mathcal{K} \times \gamma)$ there exists an unbiased estimating function $U_1(\theta, \gamma)$ for θ and $\Lambda_\mathcal{K}^\perp(\kappa, \gamma) = \Lambda_\mathcal{K}^\perp(\gamma)$ is the same for all $\kappa \in \mathcal{K}$. This raises the question of when one might expect $\Lambda_\mathcal{K}^\perp(\kappa, \gamma) = \Lambda_\mathcal{K}^\perp(\gamma)$ and/or $\Lambda_\Gamma^\perp(\kappa, \gamma) = \Lambda_\Gamma^\perp(\kappa)$. RR (2001a) used ideas from Bickel (1982) on convex models to show that these identities hold if model $M(\kappa \times \Gamma)$ is convex in its parameter γ and $M(\mathcal{K} \times \gamma)$ is convex in κ. A model $M(\Psi)$ is convex if for all ψ^*,

$\psi \in \Psi$, any mixture of the laws governed by ψ^* and ψ lies in the model. We say that κ and γ are mutually convex in model $M(\mathcal{K} \times \Gamma)$ if both model $M(\mathcal{K} \times \gamma)$ and model $M(\kappa \times \Gamma)$ are convex. The models of both Example 1 and Example 2 have κ and γ mutually convex.

Example 1.(continuation). Model $M(\mathcal{K} \times \gamma)$ is convex in κ. Thus $\Lambda^{\perp}_{\mathcal{K}}(\kappa^*, \gamma) = \{(\frac{R}{\Phi(\gamma)} - 1)g(V); g \in \mathcal{G}\}$ does not depend on κ^*. Model $M(\kappa \times \Gamma)$ is convex in γ. Thus $\Lambda^{\perp}_{\Gamma}(\kappa, \gamma^*) = \mathcal{A}(\kappa)$ does not depend on γ^*.

Robins and Rotnitzky (2001a) provide an example that shows that condition 1 alone is not sufficient for the existence of doubly robust estimating functions.

Theorem 1 provides sufficient conditions for the existence of DR estimating functions only in models under the quite strong Condition 1. The following two theorems provide sufficient conditions for the existence of DR estimating functions in models that need not satisfy this condition. The first theorem considers the case where $\theta(\kappa, \gamma) = \theta(\kappa)$. It strengthens the suppositions of Theorem 1a by assuming that the likelihood factors into a $\kappa - part$ and a $\gamma - part$. It relaxes the suppositions of Theorem 1a by no longer assuming either that model $M(\kappa \times \Gamma)$ admits an unbiased estimating function for θ or that $\Lambda^{\perp}_{\mathcal{K}}(\kappa, \gamma)$ in model $M(\mathcal{K} \times \gamma)$ depends only on γ.

Theorem 2. (Robins, Rotnitzky and Van der Laan, (2000)). *Suppose in model $M(\mathcal{K} \times \Gamma)$, the parameter $\theta(\kappa, \gamma) = \theta(\kappa)$ depends only on κ, the likelihood $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$ factors, $\Lambda^{\perp}_{\Gamma}(\kappa, \gamma) = \Lambda^{\perp}_{\Gamma}(\kappa)$ in model $M(\kappa \times \Gamma)$ depends only on κ, and there exist an unbiased estimating function $\widetilde{U}(\theta, \gamma)$ in model $M(\mathcal{K} \times \gamma)$, i.e., $E_{\kappa^*, \gamma}\{\widetilde{U}(\theta(\kappa^*), \gamma)\} = 0$ for all (κ^*, γ). Then $U(\theta, \kappa, \gamma) = \widetilde{U}(\theta, \gamma) - \Pi_{\kappa, \gamma}[\widetilde{U}(\theta, \gamma) | \Lambda_{\Gamma}(\gamma)]$ is a DR estimating equation where $\Pi_{\kappa, \gamma}[A|\mathcal{B}]$ is the projection of the random variable A on the closed linear space $\mathcal{B}$, and $\Lambda_{\Gamma}(\kappa, \gamma) = \Lambda_{\Gamma}(\gamma)$ by the factorization of the likelihood.*

Example 2. (continuation). Robins and Rotnitzky (2001b) proved that for any choice of ϕ^{-1} other than the identity or the log link and for R continuous or discrete, model $M(\mathcal{K} \times \gamma)$ was not convex in κ and $\Lambda^{\perp}_{\mathcal{K}}(\kappa, \gamma)$ varies with κ. Thus Theorem 1 cannot be used to guarantee the existence of a DR estimating function. However it is clear that the suppositions of Theorem 2 hold for any choice of ϕ, except perhaps the existence of an unbiased estimating function $\widetilde{U}(\theta, \gamma)$ in model $M(\mathcal{K} \times \gamma)$. Further for R continuous and $\phi(x) = x^3$, Robins and Rotnitzky (2001b) proved that $\widetilde{U}(\theta, \gamma) \equiv \epsilon(\theta)h(R, V, \gamma)$ is an unbiased estimating function, where $\epsilon(\theta) = Y - \theta^3 R^3$ and $h(R, V, \gamma) = R^3 - E_{\gamma}(R^3 B|V)\{E_{\gamma}(BB^T|V)\}^{-1}B$ is the residual from the population conditional least squares regression of R^3 on $B = (1, R, R^2)^T$ provided $h(R, V, \gamma) = 0$ wp1 is false.

The following is an alternative to Theorem 1b that does not require mutual convexity in κ and γ.

Theorem 3. *Suppose there exists a function $q(\cdot)$ such that $q(\theta)$ is a linear functional of both the law indexed by κ (with γ fixed) and the law indexed by γ (with κ fixed) in the sense that for some $U_1(\gamma)$ and $U_2(\kappa), E_{\kappa\gamma}\{U_1(\gamma)\} = E_{\kappa\gamma}\{U_2(\kappa)\} = q(\theta(\kappa,\gamma))$ and the model $M(\mathcal{K}\times\Gamma) = M(\mathcal{R})$ is convex in $\rho = (\kappa,\gamma)$. Then there exists a DR estimating function. Specifically for any $S(\kappa,\gamma) \in IF(\kappa,\gamma)$ in model $M(\mathcal{K}\times\Gamma), U(\kappa,\gamma) - q(\theta) = S(\kappa,\gamma) + q(\theta(\kappa,\gamma)) - q(\theta)$ is doubly robust.*

Example 4. Suppose we have a nonparametric i.i.d. model for densities of X absolutely continuous variable w.r.t. Lebesgue measure parameterized by $\kappa = Var(X)$, $\kappa \in K = \{\kappa; \kappa > 0\}$, and $\gamma = (\mu, \eta(\cdot))$, where $\mu = E(X)$ and $\eta(\cdot)$ is the density of $(X-\mu)/\kappa^{1/2}$, with $\gamma \in \Gamma = \{\mu, \eta(\cdot); \mu \in R^1, \eta(\cdot)$ a density with mean 0 and variance 1$\}$. Let $\theta(\kappa,\gamma) = \mu\kappa$. The space $\Lambda_{\mathcal{K}}^{\perp}(\kappa,\gamma)$ varies with κ and thus $M(\mathcal{K}\times\gamma)$ is not convex in κ so the suppositions of Theorem 1 do not hold. Nonetheless, by the convexity of the entire nonparametric model $M(\mathcal{K}\times\Gamma)$ and $E_{\kappa,\gamma^*}[X\kappa] = \theta(\kappa,\gamma^*)$ and $E_{\kappa^*,\gamma}[\mu(X^2-\mu^2)] = \theta(\kappa^*,\gamma)$, a DR estimating function exists by Theorem 3. Indeed, since $IF(\kappa,\gamma)$ in model $M(\mathcal{K}\times\Gamma)$ is the span of $S(\kappa,\mu) = (X-\mu)(\kappa-2\mu^2) - \mu(X^2-\mu^2-\kappa)$, the function $S(\kappa,\mu) + \mu\kappa - \theta$ is a DR estimating function. Of course, this example does not suffer from the curse of dimensionality, but it does confirm that Theorem 3 can be applied in settings in which Theorem 1b does not apply.

The suppositions of Theorem 1b do not imply those of Theorem 3 because, as can be demonstrated with the semiparametric regression model of Example 2 for the identity link, mutual convexity in κ and γ does not imply convexity in $\rho = (\kappa,\gamma)$. Conversely Example 4 demonstrates that convexity in $\rho = (\kappa,\gamma)$ does not imply mutual convexity in κ and γ.

3. Estimation When Γ_{sub} or $\mathcal{K}_{sub}$ Are not Singletons

We now consider inference when Γ_{sub} or K_{sub} are not singletons, this being the problem of practical interest.

Definition. An estimator $\widehat{\theta}$ is a doubly robust * estimator for $\theta(\kappa,\gamma)$ in model $M(\mathcal{K}\times\Gamma)$ with respect to the parametrization (κ,γ) and submodels $\Gamma_{sub} \subset \Gamma$ and $\mathcal{K}_{sub} \subset \mathcal{K}$ if $\widehat{\theta}$ is a * estimator for $\theta(\kappa,\gamma)$ in the union model $M(\mathcal{K}\times\Gamma_{sub}) \cup M(\mathcal{K}_{sub}\times\Gamma)$, where * can represent any property of interest such as consistent, RAL, etc.

RR (2001a) show that when the true value of (κ,γ) lies in the intersection submodel $M(\mathcal{K}_{sub}\times\Gamma_{sub})$, the influence function of any RAL DR estimator must be an element of the influence function space $IF(\kappa,\gamma)$ of the full model $M(\mathcal{K}\times\Gamma)$. Following the introduction of some notation we discuss a number of settings where DR estimators exist.

Throughout, for a given random function $U(\theta, \kappa, \gamma)$, we write $U(\theta, \kappa, \gamma) = \widetilde{U}(\theta, k, j)$ where $k = k(\kappa)$ and $j = j(\gamma)$ are maximal coarsening functions of κ and γ with respect to the function $U(\theta, \kappa, \gamma)$, in the sense that if $U(\theta, \kappa_1, \gamma_1) = U(\theta, \kappa_2, \gamma_2)$ then $k(\kappa_1) = k(\kappa_2)$ and $j(\gamma_2) = j(\gamma_2)$.

Example 1. (continuation). $S_{eff}(\kappa, \gamma, \theta) \equiv \widetilde{S}_{eff}(\theta, k(\kappa), j(\gamma))$ with $k(\kappa) = \omega$ and $j(\gamma) = \gamma$.

One setting where DR RAL estimators exist is when (i) $U(\theta, \kappa, \gamma)$ is a DR estimating function for $\theta(\kappa, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$ and (ii) given $U(\theta, \kappa, \gamma) = \widetilde{U}(\theta, k(\kappa), j(\gamma))$, one can construct a consistent estimator $\widehat{j}$ of $j(\gamma)$ in model $M(\mathcal{K} \times \Gamma_{sub})$ and a consistent estimator $\widehat{k}$ of $k(\kappa)$ in model $M(\mathcal{K}_{sub} \times \Gamma)$. Under (i) and (ii), the estimator $\widehat{\theta}(\widehat{k}, \widehat{j})$ solving $P_n[\widetilde{U}(\theta, \widehat{k}, \widehat{j})] = 0$ will be a DR RAL estimator under regularity conditions provided, as we assume, that the size of Γ_{sub} and $\mathcal{K}_{sub}$ are chosen small enough so that $\widehat{j}$ converges to $j(\gamma), \gamma \in \Gamma_{sub}$ under (κ, γ) and $\widehat{k}$ converges to $k(\kappa)$, $\kappa \in \mathcal{K}_{sub}$ under (κ, γ), at sufficiently fast rates. P_n is the empirical distribution expectation operator.

Supposition (ii) holds when the likelihood factors as $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$ in model $M(\mathcal{K}_{sub} \times \Gamma_{sub})$, since then the scores $S_\gamma(\gamma)$ and $S_\kappa(\kappa)$ can be used as unbiased estimating functions if Γ_{sub} and $\mathcal{K}_{sub}$ are finite dimensional. More generally (ii) holds when there exists a possibly expanded model $M(\mathcal{K}_{\exp} \times \Gamma_{\exp})$ with $\mathcal{K} \subseteq \mathcal{K}_{\exp}, \Gamma \subseteq \Gamma_{\exp}$ such that $M(\mathcal{K}_{\exp} \times \Gamma_{\exp})$ is mutually convex in κ and γ. This is so because, in model $M(\mathcal{K}_{\exp} \times \gamma), \Lambda^{\perp}_{\mathcal{K}_{\exp}}(\kappa, \gamma) = \Lambda^{\perp}_{\mathcal{K}_{\exp}}(\gamma)$ does not depend on $\kappa \in \mathcal{K}_{\exp}$ so that elements $U(\gamma)$ of $\Lambda^{\perp}_{\mathcal{K}_{\exp}}(\gamma)$ can be used as unbiased estimating functions for $\gamma \in \Gamma_{sub}$.

In this section we assume condition (ii) holds.

Example 1. (continuation). To construct $\widetilde{S}_{eff}(\theta, \widehat{\omega}, \widehat{\gamma})$, suppose $K_{sub} = \Omega_{sub} \times N$ where Γ_{sub} and Ω_{sub} are q_γ-and q_ω-dimensional parametric models. Then, although the likelihood does not factor as $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$, nonetheless $M(\mathcal{K} \times \Gamma)$ is mutually convex in κ and γ for any choice of ϕ. In particular, we can find $\widehat{\gamma} = \widehat{\gamma}(g) \in \Gamma_{sub}$ as the solution to a $q_\gamma-$dimensional estimating equation $P_n[(\frac{R}{\Phi(\gamma)} - 1)g(V)] = 0$, each component of which is in $\Lambda^{\perp}_{\mathcal{K}}(\gamma)$. Similarly we can find $\widehat{\omega} = \widehat{\omega}(s) \in \Omega_{sub}$ as the solution to $P_n[R\{Y - \omega(V)\}s(V)] = 0$, each component of which is in $\Lambda^{\perp}_{\Gamma}(\kappa)$. Recall that even though we can find consistent estimators of ω, and γ, $\widetilde{S}_{eff}(\theta, \omega, \gamma)$ is DR only if condition (i) or (ii) of page 926 holds.

Example 2. (continuation). Returning to the set-up of Section 1, for the identity or log link, we obtain a DR estimator of θ by solving $P_n[\epsilon(\theta, \widehat{\omega})\{g(R, V) - M_{est}(\widehat{\gamma}, \theta, g, \phi)\}] = 0$ for $g \in \mathcal{G}$, where $\widehat{\omega}$ is the IRLS estimator of ω under the model K_{sub} for which $\omega \in \Omega_{sub} = \{\omega; \omega(V) = \beta_0 + \beta' V\}$ and $\widehat{\gamma}$ is the MLE in the model $\Gamma_{sub} = \{\gamma; \log it\ \gamma(V) = \alpha_0 + \alpha' V\}$. Furthermore, if we take

$\Omega_{sub} = \{\omega; \omega(V) = \beta_0 + \beta' V + \varsigma M_{est}(\widehat{\gamma}, \theta, g, \phi)\}$ and $g(R, V) = R$, we also obtain a DR estimator which for the identity link is algebraically equivalent to the OLS DR estimator of Section 1. RR (2001a) show there is no DR estimator for the logit link.

Again suppose (i) and (ii) are true. In this setting, an estimation strategy we do not recommend is the following. One first performs a global lack of fit test for model $M(\mathcal{K} \times \Gamma_{sub})$; then if the test rejects, one estimates $\theta(\kappa, \gamma)$ assuming model $M(\mathcal{K}_{sub} \times \Gamma)$ is true; if it accepts then one tests the fit of $M(\mathcal{K}_{sub} \times \Gamma)$ and if it rejects, one estimates $\theta(\kappa, \gamma)$ assuming model $M(\mathcal{K} \times \Gamma_{sub})$ is true. If neither test rejects one uses the doubly robust estimation strategies described above. This preliminary test strategy can result in a RAL estimator in the union model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$ provided that, in order to insure regularity, the lack of fit tests have power zero against all parametric Pitman alternatives. However, in our view, we do not recommend this strategy because it takes too seriously the truth of the union model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$ which is really only a working model that is used because the model $M(\mathcal{K} \times \Gamma)$ is too large. We would therefore recommend that if a lack of fit test rejects either model $M(\mathcal{K} \times \Gamma_{sub})$ *or* $M(\mathcal{K}_{sub} \times \Gamma)$, one enlarges Γ_{sub} or $\mathcal{K}_{sub}$ (until neither lack of fit test rejects) and then uses the above DR estimation strategy.

We now turn to the question of how we might obtain DR estimators when no DR estimating function $U(\theta, \kappa, \gamma)$ for θ exists. We do so by extending to doubly robust estimation several well-known approaches to estimation of a parameter θ in a model where no unbiased estimating function for θ exists.

We first consider how to construct DR estimators of certain non-regular parameters (i.e., parameters that do not have finite semiparametric information bounds). Our strategy is to approximate the non-regular parameters by a regular parameter, as in Van der Laan and Robins (1998) and Bickel and Ritov (2000), because non-regular parameters do not admit unbiased estimating functions. Let $\theta(\kappa, \gamma)$ be a nonregular parameter. Suppose that $\theta_\delta(\kappa, \gamma)$ is a regular parameter such that $\theta_\delta(\kappa, \gamma)$ converges to $\theta(\kappa, \gamma)$ as $\delta \downarrow 0$. Suppose a DR estimating function $U_\delta(\theta, \kappa, \gamma) = \widetilde{U}_\delta(\theta, k(\kappa), j(\gamma))$ exists for $\theta_\delta(\kappa, \gamma)$. Then, in general, the estimator $\widehat{\theta}_{\delta(n)}(\widehat{k}, \widehat{j})$ solving $P_n[\widetilde{U}_{\delta(n)}(\theta, \widehat{k}, \widehat{j})] = 0$ will under regularity conditions be a DR consistent estimator for $\theta(\kappa, \gamma)$ if $\delta(n) \downarrow 0$ as $n \uparrow \infty$ at an appropriate rate.

Example 1. (continuation). Suppose now that Y is a continuous variable with a twice differentiable density w.r.t. Lesbesgue measure and ϕ is logistic. Let $\theta(\kappa, \gamma) = f(y; \kappa, \gamma)$ be the density of Y at y and let $\theta_\delta(\kappa, \gamma) = E_{\kappa,\gamma}\{W(\delta)\}$ where $W(\delta) = \{w((Y - y)/\delta)\}/\delta$, $w(\cdot)$ is a mean zero smooth positive kernel function and δ a suitable bandwidth. Then

$$U_\delta(\kappa, \gamma, \theta) = \widetilde{U}_\delta(\omega, \gamma, \theta) = R\Phi^{-1}(\gamma)\{W(\delta) - \theta\}$$

$$-\{R\Phi^{-1}(\gamma)-1\}E_\omega[\Phi'(\gamma)\Phi(\gamma)^{-2}\{W(\delta)-\theta\}|R=1,V]/E_\omega\{\Phi'(\gamma)\Phi(\gamma)^{-2}|R=1,V\}$$

is a DR estimating function for $\theta_\delta(\kappa,\gamma)$. Under suitable regularity conditions, and with $\delta(n)=n^{-1/5}$, $\widehat{\theta}_{\delta(n)}(\widehat{\omega},\widehat{\gamma})$ solving $P_n[\widetilde{U}_{\delta(n)}(\widehat{\omega},\widehat{\gamma},\theta)]=0$ will be $n^{2/5}$-consistent for $\theta(\kappa,\gamma)$ in model $M(\mathcal{K}\times\Gamma_{sub})\cup M(\mathcal{K}_{sub}\times\Gamma)$.

Suppose next $\theta(\kappa,\gamma)$ is a regular parameter for which no DR estimating function exists, but there exists a possibly non-regular parameter $\psi(\kappa,\gamma)$ and a function $U(\theta,\psi,\kappa,\gamma)=\widetilde{U}(\theta,\psi,k(\kappa),j(\gamma))$ that is a DR estimating function for $\theta(\kappa,\gamma)$ with $\psi(\kappa,\gamma)$ known. That is $E_{\kappa^*,\gamma}[U(\theta(\kappa^*,\gamma),\psi(\kappa^*,\gamma),\kappa,\gamma)]=E_{\kappa,\gamma^*}[U(\theta(\kappa,\gamma^*),\psi(\kappa,\gamma^*),\kappa,\gamma)]=0$. Suppose further there exists a DR $n^{1/4}$-consistent estimator $\widehat{\psi}=\widehat{\psi}(\theta(\kappa,\gamma))$ for $\psi(\kappa,\gamma)$ when $\theta(\kappa,\gamma)$ is known. Then subject to regularity conditions the estimator $\widehat{\theta}(\widehat{\psi},\widehat{k},\widehat{j})$ solving $P_n[\widetilde{U}(\theta,\widehat{\psi}(\theta),\widehat{k},\widehat{j})]=0$ will be a doubly robust RAL estimator. RR (2001a) provide a concrete example.

Finally, suppose $\theta(\kappa,\gamma)$ is a regular parameter; no DR estimating function for $\theta(\kappa,\gamma)$exists even with some other parameter known; but $\theta(\kappa,\gamma)$ is known function $b(\zeta(\kappa,\gamma),\tau(\kappa,\gamma))$of parameters $\zeta(\kappa,\gamma)$ and $\tau(\kappa,\gamma)$ that admit RAL DR estimators. In this setting we can obtain a RAL DR estimator of $\theta(\kappa,\gamma)$ by evaluating $b(\cdot,\cdot)$ at RAL DR estimators of $\zeta(\kappa,\gamma)$ and $\tau(\kappa,\gamma)$. RR (2001a) provide a concrete example. Note that the existence of DR estimating functions for $\zeta(\kappa,\gamma)$ and $\tau(\kappa,\gamma)$ does not imply that $b(\zeta(\kappa,\gamma),\tau(\kappa,\gamma))$ has a DR estimating function.

4. Generalized Double Robustness

In this section we discuss settings in which (i) exact DR estimators are difficult to compute, or (ii) no exact DR estimators exist. For such situations we propose the use of "generalized" DR estimators. Generalized DR estimators are those which have small asymptotic bias if either one of two (possibly incompatible) lower dimensional models Γ_{sub} or K_{sub} is approximately correct. Thus, a generalized DR estimator shares with a true DR estimator the crucial property of giving the analyst two chances for approximately correct inference about θ. We will illustrate a generalized DR estimator in setting (ii). RR (2001a) provide an example of a generalized estimator in setting (i).

Example 1. (continuation). Consider Example 1 with ϕ logistic, $\alpha\neq 0$, except with Y continuous. Suppose we wish to estimate the parameter $\theta(\kappa,\gamma)$ of a given marginal parametric model $f(Y;\theta)$. For concreteness, we use a normal model with mean θ_1 and variance θ_2. As noted earlier the parameters κ and γ, defined as before, determine the marginal law of Y. However, in contrast to our previous discussion of Example 1, the model is no longer non-parametric, and when $\alpha\neq 0$ the set of parameters (κ,γ) compatible with the parametric model $f(Y;\theta)$ is no longer a product space. We can make $\mathcal{R}$ a product space by choosing the following new parameterization. We let $\Gamma=\{\gamma=\gamma(\cdot)\}$ remain unchanged, but now let κ

parametrize the law $f(Y,V)$ rather than the laws $f(Y|V,R=1)$ and $f(V)$. Thus, we now take $\mathcal{K}$ to be $\mathcal{K}=\{\kappa=(\theta,\omega);\omega\in\Omega,\theta=(\theta_1,\theta_2),\theta_1\in R^1,\theta_2\in(0,\infty)\}$, where Ω is the set of all densities for V given Y. In model $M(\mathcal{K}\times\Gamma)$, the IF space for θ is $\{U(\theta(\kappa),\kappa,\gamma,c);c\in\mathcal{C}\}$, with $\mathcal{C}$ the set of all functions Y, and

$$U(\theta,\kappa,\gamma,c)=\frac{R\widetilde{C}(\theta)}{\Phi(\gamma)}-\left\{\frac{R}{\Phi(\gamma)}-1\right\}\frac{E_{\kappa,\gamma}\left[e^{-\alpha Y}\widetilde{C}(\theta)|R=1,V\right]}{E_{\kappa,\gamma}[e^{-\alpha Y}|R=1,V]},$$
$$\widetilde{C}(\theta)=c(Y)-\int c(Y)f(Y;\theta)\,dY.$$

Note that because of the redefinition of κ, $\theta(\kappa,\gamma)=\theta(\kappa)$, and $f(Y|R=1,V;\kappa,\gamma)$ is now a function of both γ and κ. RR (2001a) show that no DR estimating function for $\theta(\kappa)$ exists in model $M(\mathcal{K}\times\Gamma)$ with respect to the new parametrization (κ,γ) for any submodels $\Gamma_{sub}\subset\Gamma$ and $\mathcal{K}_{sub}\subset\mathcal{K}$ when $\alpha\neq 0$. However a "generalized" DR estimator $\widehat{\theta}(\widehat{\tau},\widehat{\gamma},c)$ is obtained by solving $P_n[\widetilde{U}(\theta,\widehat{\tau}(\theta),\widehat{\gamma},c)]=0$ with $\widetilde{U}(\theta,\widehat{\tau}(\theta),\widehat{\gamma},c)=R\widetilde{C}(\theta)\Phi(\hat{\gamma})^{-1}-\{R\Phi(\hat{\gamma})^{-1}-1\}b(V;\widehat{\tau}(\theta))$, where $b(V;\tau(\theta))$ is a user specified model for the ratio $E_{\kappa,\gamma}[e^{-\alpha Y}\widetilde{C}(\theta)|R=1,V]/E_{\kappa,\gamma}[e^{-\alpha Y}|R=1,V]$ indexed by a finite dimensional parameter τ, and $\widehat{\tau}(\theta)$ is the $e^{-\alpha Y}-$weighted non linear least squares regression estimator of τ solving $P_n[e^{-\alpha Y}R(\widetilde{C}(\theta)-b(V;\tau))\partial b(V;\tau)/\partial\tau]=0$. The theoretical difficulty with this approach is that the model $b(V;\tau)$ for $E_{\kappa,\gamma}[e^{-\alpha Y}\widetilde{C}(\theta)|R=1,V]\Big/E_{\kappa,\gamma}[e^{-\alpha Y}|R=1,V]$ will often be incompatible with the model $M(\mathcal{K}\times\Gamma)$, in the sense that there does not exist a joint distribution that satisfies both. In such case, $\widehat{\theta}(\widehat{\tau},\widehat{\gamma},c)$ of course cannot be a DR RAL estimator in model $M(\mathcal{K}\times\Gamma)$. However, this theoretical difficulty does not seem to us to be a practical difficulty. After all, as discussed in Section 1, even for models that admit DR estimators, the chosen low dimensional models $\mathcal{K}_{sub}$ and Γ_{sub} are practically (although not logically) certain to be misspecified; thus our best hope is that one of the two submodels is nearly correct, so the bias of the DR estimator will be small. In precise analogy if either model Γ_{sub} for γ or model $b(V;\tau(\theta))$ for $E_{\kappa,\gamma}[e^{-\alpha Y}\widetilde{C}(\theta)|R=1,V]/E_{\kappa,\gamma}[e^{-\alpha Y}|R=1,V]$ is nearly correct, the bias of $\widehat{\theta}(\widehat{\tau},\widehat{\gamma},c)$ for $\theta(\kappa)$ will be small.

Discussion: Heretofore we have been studying union models $M(\mathcal{K}\times\Gamma_{sub})\cup M(\mathcal{K}_{sub}\times\Gamma)$ that possess a non-empty intersection submodel $M(\mathcal{K}_{sub}\times\Gamma_{sub})$. A consequence of this fact is that, by Theorem 1, any unbiased estimating function $U(\theta,\kappa,\gamma)$ for θ in $M(\mathcal{K}\times\gamma)\cup M(\kappa\times\Gamma)$ must satisfy $U(\theta(\kappa,\gamma),\kappa,\gamma)\in IF(\kappa,\gamma)$ in model $M(\mathcal{K}\times\Gamma)$. It is this consequence that underlies many of the results presented in this discussion. RR (2001a) discuss double robustness in union models with empty intersections.

As far as we are aware Brillinger (1983) was the first to call attention to and provide examples of DR-like estimators. Other examples are given by Ruud (1983, 1986), Duan and Li (1987, 1991), Newey (1990), Robins, Mark and Newey (1992), Ritov and Robins (1997), Lipsitz and Ibrahim (1999). All these examples have $\theta(\kappa, \gamma) = \theta(\kappa)$ and likelihood factorization $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$; thus they are all special cases of the general model treated in Theorem 2 above. Scharfstein et al. (1999) and Robins (2000) went beyond individual examples to provide a general theory of double robustness in missing data and counterfactual causal inference models in which the data was coarsened at random (CAR). Robins et al. (2000) extended these latter results to cover all models with $\theta(\kappa, \gamma) = \theta(\kappa)$ and likelihood factorization $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$; they stated and proved Theorem 2. Scharfstein et al. (1999) treated Example 1 which is the only previous example we have found in the literature in which $\theta(\kappa, \gamma)$ depends on both (κ, γ) and the likelihood does not factor. At present, Theorem 2 seems to be our most significant practical result in the sense that the set of models that are known to admit DR estimators, but that do not satisfy the suppositions of Theorem 2, is still quite small.

Acknowledgement

This work was partially completed while Andrea Rotnitzky was visiting the Department of Economics at Di Tella University, Buenos Aires. James Robins and Andrea Rotnitzky were partially funded by grants from the National Institutes of Health.

Department of Epidemiology and Biostatistics, Harvard School of Public Health, 655 Huntington Ave. Boston, MA 02115, U.S.A.

E-mail: robins@hsph.harvard.edu

Department of Economics, Di Tella University, Minones 2159. Bouenos Aires, Argentina.

E-mail: andrea@hsph.harvard.edu

COMMENTS

Xiaotong Shen and Bing Li

The Ohio State University and The Penn State University

Bickel and Kown are to be congratulated for their insights into many important issues in semiparametric and nonparametric inferences, and for sharing

their view on further research. In particular, they have listed six areas (A)-(F) for further growth. Bickel and Kown describe a systematic method for finding the efficient influence function and the information bound for semiparametric and nonparametric models in area (D). Because the efficient influence function is defined as a solution to a functional equation in a functional space, we have felt that a major part of the work would involve "guessing and checking" — that is, first guess a solution (sometimes available as a "good estimator") and then check if the relevant functional equation can be satisfied. Now Bickel and Kwon have substantially reduced the guess work in this.

Our discussion comprises two parts. The first part concerns the frequency properties of semiparametric and nonparametric Bayes procedures. The second part establishes a connection between "calculus of information" and the theory of estimating equations, which focuses on area (D).

1. Frequency Properties of Bayes Procedures

In recent years, advances in computer technology make it much easier to compute posterior distributions over large parameter spaces. As a result, Bayesian methodology has been widely implemented in many semiparametric and nonparametric models, which motivates extensive studies on computational aspects of semiparametric and nonparametric Bayes procedures. Yet, frequency aspects of these Bayes procedures have not received a lot of attention.

Frequency properties of semiparametric and nonparametric Bayes procedures are important not only to Bayesians but also to frequentists. Often, Bayes procedures are used by frequentists, especially when it is difficult to implement a desired frequentist procedure. Based on our intuition and experience with parametric models, we expect that semiparametric and nonparametric posterior distributions behave like a parametric posterior. Limited numerical evidence also indicates this. However, due to the difficulty of evaluating an infinite-dimensional distribution, it is very difficult or impossible to make a conclusive statement about semiparametric and nonparametric posterior distributions based on simulations. Recent theoretical investigation suggested that special care is necessary and that posterior distributions may behave quite differently than a parametric posterior distribution. In what is to follow, we use a simple example to illustrate a number of important aspects of semiparametric and nonparametric posterior distributions.

Consider a simple version of the nonparametric regression example of Cox (1993). Let

$$Y(x) = \theta(x) + n^{-1/2}N(x), \tag{1}$$

where $N(x)$ is the one-dimensional Brownian motion. Model (1) is closely related to the usual nonparametric regression; c.f., Cox (1993) and Brown and Low

(1996). Now expand $\theta(x)$ and $Y(x)$ in terms of an orthonormal basis $\{\psi_i\}$ with $(\theta_1, \theta_2, \ldots)$ and $(Y_1, Y_2, \ldots)$ being the corresponding coefficients in the expansion. Then (1) becomes $Y_i = \theta_i + n^{-1/2}\epsilon_i$, where the ϵ_i's are i.i.d. $N(0,1)$. Here we estimate infinitely many normal means $\{\theta_i\}$, which is equivalent to estimating $\theta(x)$. In this example, $\theta(\cdot)$ is smooth and belongs to a Sobolev space $W_2^p[0,1]$, where p is the degree of smoothness, measured by the usual L_2-metric. Now, define $\Theta(x)$ to be $\sum_{j=1}^{\infty} \alpha_j B_j(x)$, where $\{B_j\}_{j=1}^{\infty}$ is the Fourier basis on $[0,1]$, and $\{\alpha_j\}$ are independent normal random variables distributed according to $N(0, j^{-2d})$, with $d \in R^1$ satisfying $\sum_{j=1}^{\infty} j^{2(p-d)} < \infty$. By the Three Series Theorem, the sample paths of θ have pth derivatives $\theta^{(p)}$ in L_2 if and only if $\sum_{j=1}^{\infty} j^{2(p-d)} < \infty$. Roughly, $d > p + 1/2$ and $p > 1/2$. Consequently, Θ induces a probability measure on $W_2^p[0,1]$; c.f., Kuo (1975). By the Karhunen-Loéve expansion, this prior is equivalent to that of Cox (1993).

Model (1) has been extensively studied by frequentists. A variant of maximum likelihood estimator (MLE) such as penalized and sieve MLEs achieves the desired frequency properties. Naturally, one would expect that the nonparametric posterior distribution possesses the desired frequency properties of a parametric posterior distribution. Unfortunately, Cox (1993) showed that a 95% posterior confidence region defined by the usual L_2-norm has zero asymptotic frequency coverage for almost all $\theta \in W_2^p[0,1]$ and any integer $p \geq 2$. A more detailed explanation of the phenomenon of Cox (1993) has been given in a recent paper by Wasserman (1998), and by Diaconis and Freedman (1999). This negative result is in contrast to the well established Bernstein-Von Mises Theorem, which says that this phenomenon typically does not occur for a parametric posterior distribution, except for pathological examples.

To gain insight into the structure of the problem, we first examine the rates of convergence of the posterior mean and the posterior distribution, measured by the L_2-norm. As shown in Shen and Wasserman (2001), the exact rate of convergence of the posterior mean and that of the posterior distribution are of order of n^{-b}, where b belongs to $(\frac{p}{2d}, \frac{1}{2}(1 - \frac{1}{2d})]$, depending on θ. This means that, the rates are faster for some θ's, and slower for others, as compared to the optimal rate $n^{-\frac{p}{2p+1}}$. In contrast, the optimal rate $n^{-\frac{p}{2p+1}}$ is attainable by a variant of MLE for all $\theta \in W_2^p[0,1]$; c.f., Pinsker (1980). A similar phenomenon in the minimax sense was also observed in Zhao (2000). In the setting of (1), the prior impedes the performance of the posterior distribution, and that of the posterior mean, because the prior assigns small probability to any neighborhood of the true parameter, which explains the phenomenon of Cox (1993).

It is natural to ask why the phenomenon of Cox (1993) occurs in this simple example with a Gaussian prior. It is interesting to note that the restriction of $\sum_{j=1}^{\infty} j^{2(p-d)} < \infty$ prevents us from choosing the value of $d = p + 1/2$ that yields

the optimal rate $n^{-\frac{p}{2p+1}}$, where $(\frac{p}{2d}, \frac{1}{2}(1-\frac{1}{2d})] = \{\frac{p}{2p+1}\}$. In other words, the prior impedes the performance of the posterior if we insist a prior assign positive probability to all smooth functions in the parameter space $W_2^p[0,1]$. On the other hand, even if the prior with $d = p + 1/2$ is used, it remains unclear whether the Bernstein-Von Mises Theorem holds for this prior. A similar phenomenon occurs for a Dirichlet process prior in estimating a continuous distribution function, where the Dirichlet prior assigns zero probability to all continuous distribution functions. However, the Dirichlet priors does not impede the performance of the posterior, and the Bernstein-Von Mises Theorem holds in this situation.

Philosophically, it seems reasonable to use a prior that puts positive probability on a parameter space. For this reason, a mixture of Dirichlet priors that assigns positive probability to all continuous distribution functions is preferable, as opposed to the Dirichlet prior. So consider a hierarchical prior that assigns positive probability to all smooth functions in $W_2^p[0,1]$. The hierarchical prior, called the sieve prior, is defined on a sequence $\{\Theta_k\}$ of nested approximating spaces (sieve) as follows: $\pi(\cdot) = \sum_{k=1}^{\infty} \lambda_k \pi_k(\cdot)$, where $\sum_k \lambda_k = 1$, $\lambda_k \geq 0$, and π_k is a prior on Θ_k. This prior is essentially a two-stage hierarchical prior. As shown in Shen and Wasserman (2001), it recovers the optimal rate of convergence of the posterior distribution and the posterior mean for any θ in $W_2^p[0,1]$. In addition, as shown in Zhao (2000), it also gives the optimal minimax rate of convergence of the posterior mean. Furthermore, in a closely related problem, Huang (2001) used this type of sieve prior to adaptively obtain the optimal rate of convergence of the posterior distribution without knowing the degree p of smoothness. However, it still remains an open problem as to whether the Bernstein-Von Mises Theorem holds for this type of sieve prior. Some preliminary results in Genovese, Shen and Wasserman (2001) suggest that it may not hold for the above sieve prior in the setting of (1). Further investigation is necessary.

2. Estimating Equations

In this section we explore the connection between the "calculus of information", as described in the paper, and the theory of optimal estimating equations. Though the subsequent discussion can be made much more general we focus on the simplest case — i.i.d. with real-valued parameters — to illustrate the ideas. A detailed and general development can be found in a book in preparation Li (2001).

In the theory of estimating equations we usually start with a class of candidate estimating equations, and find the optimal one that minimizes the asymptotic variance of the solutions. The optimal estimating equation turns out to be the projection of the true score on the closed spanned of the class of estimating equations. Usually, the form of the true score is unknown — that is, it

cannot be described by the set of moment conditions that specifies the class of estimating equations. However, the projection of the score *can* be described by the mentioned set of moment conditions.

To fix the ideas, assume $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. copies of (X, Y) whose conditional mean and conditional variance belong to parametric families $E(Y|x) = \mu(\theta x)$ and $\mathrm{Var}(Y|x) = V(\theta, x)$ for some known functions μ and V. Suppose we are interested in estimating the regression parameter θ. Consider the class of square-integrable unbiased estimating equations linear in Y, that is,

$$\mathcal{G} = \{g(\theta, X, Y) = a(\theta, X)(Y - \mu(\theta Y)) : E_\theta g^2(\theta, X, Y) < \infty\}. \tag{2}$$

With appropriate conditions for completeness, this is a Hilbert space with inner product $\langle g, h\rangle_\theta = E_\theta(gh)$. Note that the inner product, and hence the Hilbert space, is completely specified by the functions μ and V.

Let $p_\theta(x, y)$ be the true density. Then the projection of $\partial \log p_\theta(x, y)/\partial\theta$ onto $\mathcal{G}$ is $\dot\ell(\theta) = x\dot\mu(x\theta)(y - \mu(\theta x)/V(\theta, x)$, where $\dot\mu$ denotes the derivative of μ. Note that $\dot\ell$ is specified by μ and V even though the form of p_θ is unknown. It can be shown by a simple application of the Cauchy-Schwarz inequality that the solution to $\dot\ell(\theta) = 0$ has the smallest asymptotic variance among the solutions to any estimating equation in $\mathcal{G}$. See Heyde (1997, Chapter 2).

We now formulate the optimality of the quasi score in terms of the "calculus of information." Consider the family of probability distributions $\mathcal{P}$ whose conditional mean is described by a parametric family. That is, each P in $\mathcal{P}$ has a probability density function of the form $f(y|x)g(x)$, with g being the marginal density of X that is fixed among the family, and f being the conditional density of Y given X that satisfies the moment constraint $E(Y|x) = \mu(\theta x)$ for some θ and for all x. In symbols, for some fixed g, $\mathcal{P} = \{p(x, y) = f(y|x)g(x) : \int y f(y|x)dy = \mu(\theta x)$ for some θ and for all $x\}$. Let $p_0 = f_0(y|x)g(x)$ be the true probability density that generates (X, Y), and let θ_0 correspond to f_0. Let $V_0(x)$ be the conditional variance of Y under the true conditional density $f_0(y|x)$.

Define $\dot\nu = \{x\dot\mu(\theta_0 x)(y - \mu(\theta_0 x)/V_0(x)\}/E\{X^2\dot\mu^2(\theta_0 X)/V_0(X)\}$. It is shown in Li (2001) that $\dot\nu$ is the efficient influence function with respect to the family $\mathcal{P}$. Now it is easy to see that the solution to the quasi score equation $\dot\ell(\theta) = 0$ is an asymptotically linear estimator with $\dot\nu$ as influence function, where $V_0(x)$ is taken to be $V(\theta_0 x)$ in the previous section. Hence the quasi likelihood estimator is optimal among all estimators that are regular with respect to $\mathcal{P}$.

This also gives the quasi score a new interpretation — it is the tangent of the least favorable curve in $\mathcal{P}$ that passes through p_0.

To compare the methods of optimal estimating equation and calculus of information, we note that the first method is simpler than the second, but the second gives a stronger conclusion than the first. The demonstration of the optimality

in the first section essentially only involves an application of the Cauchy-Schwarz inequality, whereas the second method requires the Riesz representation of a Frechet derivative in the tangent space of $\mathcal{P}$, as well as the use of the Convolution Theorem. However, the second optimality result is stronger: the class of linear estimating equations $\mathcal{G}$ can roughly be identified with the class of regular, asymptotically linear estimators whose influence functions are linear in y (Li (2001)), but the class of regular estimators with respect to $\mathcal{P}$ include, in addition, asymptotically nonlinear regular estimators.

The two methods proceed in opposite ways. The first starts with a class of estimating equations $\mathcal{G}$ and seeks the optimal one among that class; the second starts with a family of distributions $\mathcal{P}$ and seeks the least favorable path in that family. In the end, both give rise to the quasi score (or what is proportional to it). Intuitively, the more distributions we put into the family $\mathcal{P}$, the harder it is for an estimator to be regular with respect to it, and consequently the efficient estimator with respect to $\mathcal{P}$ is optimal among a smaller class. Conversely a larger class of estimating equations corresponds to more moment conditions, which are satisfied by fewer probability distributions, resulting in an estimator that is optimal in a wider class.

Department of Statistics, The Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210, U.S.A.

E-mail: xshen@stat.ohio-state.edu

Department of Statistics, The Penn State University, University Park, 410 Thomas Building, U.S.A.

E-mail: bing@stat.psu.edu

COMMENTS

Cun-Hui Zhang

Rutgers University

The authors are to be congratulated for pointing out interesting research directions in semiparametric inference. I would like to complement the paper with a theorem of M-estimators based on $\widehat{b}^{(n)}$. I will also describe an extension of the semiparametric efficiency to the estimation of sums of random variables and comment on minimaxity, both related to my recent work.

1. M-estimators

Let $\mathcal{P} \subseteq \mathcal{M}$ be a submodel. Consider a finite- or infinite-dimensional parameter $\theta : \mathcal{P} \to \mathcal{B}_1$ for a Banach space $(\mathcal{B}_1, \|\cdot\|_1)$. Suppose conditions (A1), (A2) and (A3) hold for the nonparametric estimator $\widehat{b}^{(n)}$. Let $\Theta \equiv \{\theta(b) : b \in \mathcal{P}\}$ and $\{B_\theta, \theta \in \Theta\}$ be a family of bounded linear operators from $\mathcal{B}$ to a Banach space $(\mathcal{B}_2, \|\cdot\|_2)$ satisfying the Fisher consistency condition $B_{\theta(b)}(b) = 0 \ \forall\, b \in \mathcal{P}$. Let $\widehat{\theta}^{(n)}$ be an M-estimator of $\theta \equiv \theta(b)$ satisfying

$$B_{\widehat{\theta}^{(n)}}(\widehat{b}^{(n)}) = r_n \epsilon_2^{(n)}, \qquad \widehat{\theta}^{(n)} \in B_1, \tag{1}$$

with $\|\epsilon_2^{(n)}\|_2 \to 0$ in $P_b^{(n)}$ for all $b \in \mathcal{P}$. We provide sufficient conditions for the asymptotic linearity and efficiency of $\widehat{\theta}^{(n)}$.

Let $b_0 \in \mathcal{P}$ be the true value of b. Set $\theta_0 = \theta(b_0)$. Suppose

$$Z^{(n)} \equiv (\widehat{b}^{(n)} - b_0)/r_n \xrightarrow{\mathrm{D}} Z \text{ in } (\mathcal{B}, \|\cdot\|), \quad \|\widehat{\theta}^{(n)} - \theta_0\|_1 = o_P(1), \tag{2}$$

under $P_{b_0}^{(n)}$. Here $\xrightarrow{\mathrm{D}}$ means weak convergence in the sense of Hoffmann-Jøgensen (1984,1991). Suppose B_θ is strongly continuous in θ at θ_0:

$$\lim_{\|\theta - \theta_0\|_1 \to 0} \|B_\theta(b) - B_{\theta_0}(b)\|_2 = 0, \quad \forall\, b \in \mathcal{B}. \tag{3}$$

Suppose further that, for certain bounded linear operator $A_{\theta_0} : \mathcal{B}_2 \to \mathcal{B}_1$,

$$\left.\begin{array}{c} \|r_n^{-1}\{B_{\theta_n}(b_0) - B_{\theta_0}(b_0)\} - g_2\|_2 \to 0 \\ \text{and } \|\theta_n - \theta_0\|_1 \to 0 \end{array}\right\} \Rightarrow \|r_n^{-1}(\theta_n - \theta_0) - A_{\theta_0} g_2\|_1 \to 0 \tag{4}$$

for all $\theta_n \in \Theta$ and $g_2 \in \mathcal{B}_2$. It follows directly from the continuous mapping theorem and the asymptotic tightness of $Z^{(n)}$ that the Fisher consistency of B_θ, (1), (2), (3) and (4) imply the asymptotic linearity and normality of $\widehat{\theta}^{(n)}$ in the sense that

$$r_n^{-1}(\widehat{\theta}^{(n)} - \theta_0) = -A_{\theta_0} B_{\theta_0} Z^{(n)} + \epsilon_1^{(n)} \xrightarrow{\mathrm{D}} -A_{\theta_0} B_{\theta_0} Z \quad \text{ in } (\mathcal{B}_1, \|\cdot\|_1), \tag{5}$$

where $\|\epsilon_1^{(n)}\|_1 = o(1)$ in $P_{b_0}^{(n)}$. Since $-A_{\theta_0} B_{\theta_0}$ is the influence operator of $\widehat{\theta}^{(n)}$, the asymptotic efficiency of $\widehat{\theta}^{(n)}$ at b_0, with respect to any $\Theta^* \subseteq \mathcal{B}_1^*$, is equivalent to $B_{\theta_0}^* A_{\theta_0}^* \theta^* \in \mathcal{B}_0^*$ and $T B_{\theta_0}^* A_{\theta_0}^* \theta^* \in \mathcal{P}'(b_0)$ for all $\theta^* \in \Theta^*$, where $\mathcal{B}_0^*$ and T are as in (A1) and (A2) and $\mathcal{P}'(b_0)$ is the tangent space of $\mathcal{P}$ at b_0. By the Fisher consistency and (3) and (4), $-A_{\theta_0} B_{\theta_0}$ is the derivative of $\theta(b)$ in the sense that $\{\theta(b_n) - \theta_0\}/r_n \to -A_{\theta_0} B_{\theta_0} g$ when $(b_n - b_0)/r_n \to g$ and $\theta(b_n) \to \theta_0$.

The above theorem is a straightforward translation of the methods in Vardi and Zhang (1992), Gu and Zhang (1993), and Tsai and Zhang (1995), where

nonparametric maximum likelihood estimators were considered in specific, but quite different, incomplete-data models. In these papers, the Fisher consistency and (1) are, respectively, equivalent to the population and sample versions of self-consistency equations via the EM algorithm, while the crucial implicit differentiability condition (4) was verified via the identity

$$B_\theta(b_0) - B_{\theta_0}(b_0) = R_\theta(\theta - \theta_0), \tag{6}$$

with a family of linear operators $R_\theta : \mathcal{B}_1^0 \to \mathcal{B}_2$, $\{\Theta - \theta_0\} \subset \mathcal{B}_1^0 \subset \mathcal{B}_1$, and via the continuous invertibility of R_θ at $\theta = \theta_0$ under the strong topology of $\mathcal{B}_1$ and $\mathcal{B}_2$. The approach with (4) and (6) seems to lead to sharper results, compared with others, since it allows careful choice of the spaces $\mathcal{B}$, $\mathcal{B}_1$, $\mathcal{B}_2$ and $\mathcal{B}_1^0$. The identity (6) was extended to a general convex model by van der Laan (1995), and the methodology was extended recently by Zhan (1999) to certain general models with i.i.d. observations. Zhang and Li (1996) showed that $R_{\theta_0} = A_{\theta_0}^{-1}$ is an information operator when B_{θ_0} is a score operator with censored data. Theory of M-estimators (Z-estimators) was considered in Van der Vaart and Wellner (1996) among others, and is closely related to the compact differentiability approach of Gill (1989) and Gill and Van der Vaart (1993).

2. Estimation of Sums of Random Variables

Let (X, θ), (X_j, θ_j) be i.i.d. random vectors with an unknown joint distribution F, $F \in \mathcal{F}$. Let $\{u(x, \vartheta; F), F \in \mathcal{F}\}$ be functions satisfying certain mild smoothness conditions. Semiparametric information bounds can be extended to the estimation (or prediction) of the sum

$$S_n \equiv S_n(F) \equiv \sum_{j=1}^{n} u(X_j, \theta_j; F)$$

based on $X_1, \ldots, X_n$. The problem is closely related to the estimation of the mean $\mu(F) \equiv E_F u(X, \theta; F)$. Let $\psi_*(x; F_0)$ be the efficient influence function for the estimation of $\mu(F)$ at F_0. Let $\overline{u}(x; F) \equiv E_F[u(X, \theta; F)|X = x]$ and $u_*(x; F_0)$ be the $L_2(P_{F_0})$ projection of $\overline{u}(x; F_0)$ to the tangent space at F_0. An estimator $\widehat{S}_n$ of $S_n(F)$ is asymptotically efficient in contiguous neighborhoods of F_0 if, under P_{F_0},

$$\widehat{S}_n = n\mu(F_0) + \sum_{j=1}^{n} \phi_*(X_j; F_0) + o_P(\sqrt{n}).$$

The efficient influence function $\phi_*(x; F_0)$ for the estimation of $S_n(F)$ is related to the efficient influence function for the estimation of $\mu(F)$ via

$$\phi_*(x; F_0) = \psi_*(x; F_0) + \overline{u}(x; F_0) - \mu(F_0) - u_*(x; F_0).$$

Based on this criterion, certain "plug-in" empirical Bayes estimators are asymptotically efficient in parametric models, while the so-called "u,v" estimators of Robbins (1988) are asymptotically efficient in nonparametric mixture models.

Example 1. Given a pool of n motorists and an integer $k \geq 0$, an asymptotically efficient estimator for $\sum_{j=1}^{n} \theta_j I\{X_j = k\}$, the total intensity of those motorists with k traffic accidents, is $(k+1)\#\{i \leq n : X_j = k+1\}$, the total number of accidents of those individuals with $k+1$ accidents. Here, X_j, the number of traffic accidents for the j-th individual, is assumed to have the Poisson distribution with mean θ_j conditionally on θ_j, and θ_j are assumed to be i.i.d. variables with a completely unknown distribution.

Example 2. Let $X|\theta \sim N(\theta, \sigma^2)$ and $\theta \sim G$. Suppose G is a normal distribution with mean τ. The number of "above-average" individuals, $\#\{j \leq n : X_j > \bar{X}\}$, is an efficient estimator of the number of above-mean individuals $S_n \equiv \#\{j \leq n : X_j > \tau\}$. The estimator $n/2$ is efficient for the estimation of $E_\tau S_n = n/2$, but not S_n.

Example 3. Suppose G is completely unknown in Example 2. Then, $n/2$ is an efficient estimator of $\#\{j \leq n : X_j > \theta_j\}$.

The estimator in Example 1 was proposed by Robbins (1977, 1988). The asymptotic efficiency in Example 1 was established in Robbins and Zhang (2000), and was extended to the general case in Zhang (2001).

3. Minimaxity and Super-efficiency

For the estimation of regular parameters in regular models, super-efficient estimators (i.e. $\sqrt{n}(\widetilde{\theta}_n - \theta_0) \to 0$) are neither regular nor minimax locally asymptotically. However, this is no longer true for the estimation of *irregular* parameters.

Consider the estimation of a regression function f in the nonparametric regression model with uniform design, or equivalently, in the white noise model. Certain block empirical Bayes estimators $\widehat{f}_n$ are exactly adaptive minimax:

$$\lim_{n\to\infty} \frac{\sup_{f\in\mathcal{F}} E_f^{(n)} \int_0^1 (\widehat{f}_n - f)^2}{\inf_{\widetilde{f}_n} \sup_{f\in\mathcal{F}} E_f^{(n)} \int_0^1 (\widetilde{f}_n - f)^2} = 1, \quad \forall \mathcal{F} \in \mathcal{C}, \tag{7}$$

for a suitable collection $\mathcal{C}$ of sets $\mathcal{F} \subset L_2[0,1]$, and are also everywhere super-efficient:

$$\lim_{n\to\infty} \frac{E_f^{(n)} \int_0^1 (\widehat{f}_n - f)^2}{\inf_{\widetilde{f}_n} \sup_{f\in\mathcal{F}} E_f^{(n)} \int_0^1 (\widetilde{f}_n - f)^2} = 0, \quad \forall f \in \mathcal{F} \in \mathcal{C}. \tag{8}$$

If $\mathcal{C}$ is the class of all Sobolev balls of different smoothness indices and radii, both (7) and (8) hold for the James-Stein type block empirical Bayes estimators of Efromovich and Pinsker (1984) with the Fourier basis. If $\mathcal{C}$ is the class of all Besov balls of different smoothness indices, shapes indices (p, q), $q < \infty$, and radii, both (7) and (8) hold for the block general empirical Bayes estimators of Zhang (2000) with wavelet bases.

Acknowledgement

This research was partially supported by the National Science Foundation.

Department of Statistics, Rutgers University, Piscataway, NJ 08854, U.S.A.
E-mail: czhang@stat.rutgers.edu

REJOINDER

Peter J. Bickel and Jaimyoung Kwon

We thank the discussants for their varied and thought provoking responses although their number and variety makes a short response difficult. One of our goals, amply fulfilled, was to have extensive discussion of the many questions we raised but did not try to answer. In particular, Fan, Klaassen, and McNeney and Wellner all addressed questions A and B which we respond to further below.

Another goal of ours was to have additional important issues raised and here again, we were far from disappointed. Fan addressed the testing question. Klaassen, van der Laan and Yu, and Robins and Rotnitzky raised the important issue of robustness. Shen and Li discuss the estimating equation paradigm, relating the information calculus approach to estimating equations. Zhang gives a careful new abstract M estimation theorem.

Finally, McNeney and Wellner and Greenwood, Schick and Wefelmeyer directly address question C, the general information calculus we discuss. The latter in particular, in a discussion full of examples, correctly find that our calculations, though not our calculus, were flawed in one of the examples we discuss.

We respond briefly to the individual discussions. We begin with general responses to comments on robustness and bandwidth selection.

1. Robustness

Klaassen, van der Laan and Yu, and Robins and Rotnitzky correctly remind us of the critical role of uniformity of convergence and robustness. The optimality theory we have reviewed applies only to estimates which converge in law on the $\sqrt{n}$ scale uniformly on compact subsets of parametric submodels of the semiparametric model considered. Other types of uniformity requirements, like convergence on the $\sqrt{n}$ scale uniformly on bounded (in total variation distance) subsets of the semiparametric model for example, can be more compelling.

We can interpret such results in terms of robustness, or how stable the behavior of the estimate is, for fixed n, under perturbations of the underlying distribution defined by the uniformity class. See Hampel, Ronchetti, Rousseeuw and Stahel (1986) and Huber (1981) as primary sources, and Rieder (1994) and Shen (1995) for a discussion amplifying our comments. Stable here is in terms of the worst that can happen for perturbations of a given magnitude. Robustness is evidently desirable but, like all statistical criteria which involve worst case analyses, not determinative. We are often luckier than we deserve.

The type of uniformity of convergence one uses can also be guided by the simplicity and intuitive nature of the results one obtains. For example, requiring uniformity on compact subsets of parametric submodels of the big semiparametric model is pretty weak, yet necessary for exclusion of the super efficiency phenomenon. For estimation in functional models, requiring uniformity on weak compacts in the nonparametric (nuisance) part of the parametrization leads to clean results, see Bickel and Klaassen (1982), while requiring only pathwise uniformity leads to pathologies (Pfanzagl (1982)). Similar restrictions of uniformity on weak compacts proposed by Ritov and Robins (1997) lead us to not try to adapt to high dimensional nuisance parameters.

2. Choice of Bandwidth or Regularization Parameter

Fan, Klaassen and McNeney and Wellner all comment on our question A and B, bandwidth choice. We can summarize our presentation and their comments as follows: There are at least three distinct situations where one needs to estimate irregular (non-differentiable) parameters.

(i) Estimation of objects, such as the density function, regression function, or hazard rate, for their own sake and at minimax rate as "adaptively" as possible, in the sense of nonparametric function estimation literature. This is, of course, widely discussed in the work of Donoho and Johnstone (1995), Birgé and Massart (1999), and others.

(ii) Estimation of the above objects as necessary intermediaries in the estimation of regular Euclidean parameters, such as θ in the semiparametric regression

model

$$E(Y|R,V) = \omega(V) + \theta R + \epsilon, \tag{2.1}$$

where ϵ is independent of R and V, ω is "arbitrary", and $X = (R, V, Y)$ is observed. Here $E(Y|V)$ and $E(R|V)$ need to be estimated. More generally, we can consider regular parameters which can be written in the form $\theta(P, \eta(P))$, where η ranges over a function space and $(P, \eta) \to \theta(P, \eta)$ is smooth but the map $P \to \eta(P)$ is irregular. Thus, in (2.1),

$$\theta(P,\eta) = \frac{\int (y - a(v))(r - b(v))dP(y,r,v)}{\int (r - b(v))^2 dP(r,v)}$$

for $\eta = (a, b)$ and $\eta(P) = (E_P(Y|V = v), E_P(R|V = v))$.

Estimation of the normalized asymptotic variance $I^{-1}(P) = [\int (f'/f)^2 dP]^{-1}$ where f is the density of P in, say the symmetric location model, is another example of (ii) since one can write the parameter as $\theta(P, \eta) = \int \eta^2 dP$ with $\eta(P) = f'/f$. As McNeney and Wellner point out, this example is relevant to our question B as well.

(iii) Estimation of irregular parameters such as the density in an "adaptive" minimax way by $\hat{f}$, such that when $\hat{f}$ is plugged into $\int h\hat{f}$, the resulting estimate of $\int hf$ is efficient.

As Fan points out, situation (ii) can be dealt with by treating estimation of η as an intermediate step, by regularizing less than one needs to for minimax estimation of η, and then estimating η separately with appropriate regularization.

Regarding case (iii), McNeney and Wellner refer to the work of Bunea (2000) and Fan to that of Ruppert and Carroll et al. (1997), and both suggest that by working harder, one can achieve both goals simultaneously. This is, we believe, a consequence of special orthogonality properties in the examples they consider. In fact, Bickel and Ritov (2000b) show that the goal is in general unattainable in their discussion of the "plug in" principle. In particular, unattainability appears to be the case for our second example, estimation of $I(f)$ above, judging from the work of Laurent (1997).

In any case, persuasive criteria for choosing regularizing parameters in these situations (and perhaps generally!) are not really available. We believe there is great room for the advocacy and use of cross validation methods.

3. Response to Fan

Fan's study of generalized likelihood ratio tests has indeed provided us with a number of interesting and potentially useful tools for testing parametric, semiparametric or nonparametric hypotheses against non- or semi-parametric alternatives. The situations he considers in Sections 1 and 2 are all models where the

co-dimension of the hypothesis (the dimension of the orthocomplement of the tangent space to the hypothesis) is infinite. The likelihood ratio test statistics are asymptotically normal under the null hypothesis — chi square with an increasing number of degrees of freedom as n increases. As Fan points out, they may inherit minimax rate testing optimality in the sense of Ingster (1993) from the corresponding estimation minimax results for estimates of function-valued parameters used to construct them. This testing minimaxity can be viewed as the limiting version of the optimality among all invariant tests of the likelihood ratio test for a linear hypothesis in the Gaussian shift model with known covariance matrix. This gives equal non trivial asymptotic power against nearby alternative on scales larger than $n^{-1/2}$.

When considering nonparametric alternatives however, as Bickel, Ritov and Stoker (2001) point out, it may be important to tailor tests to directions which a priori appear important and save some power for grossly divergent alternatives in other directions, rather than having negligible power in all directions.

As usual in testing theory, there is no simple prescription at the $n^{-1/2}$ scale as there is in estimation.

4. Response to Greenwood, Schick and Wefelmeyer

In a long series of examples, Greenwood, Schick and Wefelmeyer explore the situations in which the traditional approach is contrasted, in the context of Markov chain models, to the approach we have advanced.

In their particularly intriguing Example 6 based on Müller, Schick and Wefelmeyer (2001b), they note that neither approach works simply directly. We are not entirely convinced, admittedly, at the level of heuristics. For simplicity, we consider the case where $d = 1$ and the restriction is given by $\int z db = 0$. As they argue, the tangent space is indeed the set of all functions of the form $h(X_1, X_2) - a_* z_{\mathcal{H}}$ where $a_* = \|z_{\mathcal{H}}\|^{-2}\langle h, z_{\mathcal{H}}\rangle$, $z_{\mathcal{H}}(X_1, X_2)$ is the representer of $z(X_1, X_2)$ and $E(h(X_1, X_2)|X_1) = 0$. As they point out, it is natural to use

$$\frac{1}{n-1}\left(\sum_{i=1}^{n-1} k(X_i, X_{i+1}) - \hat{a}_* \sum_{i=1}^{n-1} z(X_i, X_{i+1})\right),$$

to estimate $\int k db$ efficiently in the restricted model, as in the i.i.d. case. But, if Ta is the representer of $a(X_1, X_2)$ then, by construction,

$$\begin{aligned}\|Ta\|^2 &= AsVar(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} a(X_i, X_{i+1})) \\ &= \mathrm{Var}\,(a(X_1, X_2)) + 2\sum_{i=1}^{\infty} \mathrm{Cov}\,(a(X_1, X_2), a(X_{i+1}, X_{i+2})),\end{aligned}$$

$$\langle Ta, Tb\rangle = \operatorname{Cov}\left(a(X_1, X_2), b(X_1, X_2)\right) + \sum_{i=1}^{\infty}[\operatorname{Cov}\left(a(X_1, X_2), b(X_{i+1}, X_{i+2})\right) + \operatorname{Cov}\left(a(X_{i+1}, X_{i+2}), b(X_1, X_2)\right)].$$

Thus, estimating a_* is equivalent to estimating the asymptotic variance of $n^{-1/2}\sum_{i=1}^{n-1} z(X_i, X_{i+1})$ and the asymptotic covariance of $n^{-1/2}\sum_{i=1}^{n-1} k(X_i, X_{i+1})$ and $n^{-1/2}\sum_{i=1}^{n-1} z(X_i, X_{i+1})$, which are, respectively, the spectral density of the process $\{z(X_i, X_{i+1})\}$ at 0 and the cross spectral density of $\{k(X_i, X_{i+1})\}$ and $\{z(X_i, X_{i+1})\}$ at 0. For these, standard consistent estimates, say based on smoothing (cross-)periodograms, exist. This can easily be generalized to $d > 1$.

Greenwood et al. rightly take us to task for our Example 3, where our assertions are essentially false. Our calculations and not our calculus are at fault. Consider the nonlinear autoregressive model $X_{i+1} = g(X_i) + \epsilon_{i+1},\ 1 \le i \le n$ where ϵ_i are i.i.d. with distribution F with density f unknown except $E(\epsilon_1) = 0$ and g is unknown. At least formally, the tangent space in our sense is $\{u(X_1)l(\epsilon_2) + v(\epsilon_2) : u \in L_2(\pi), v \in V\}$, where $V = \{v \in L_2(F) : \int \epsilon v(\epsilon)dF(\epsilon) = \int v(\epsilon)dF(\epsilon) = 0\}$ and $l(\epsilon) = (f'/f)(\epsilon)$. This follows since $E(u(X_1)l(\epsilon_2)|X_1) = 0$ and $E(v(\epsilon_2)) = E(\epsilon_2 v(\epsilon_2)) = 0$ are immediate. The tangent space structure is exactly as in the nonparametric regression model with i.i.d. errors, and the analysis of Koul and Schick (1997) based on the traditional approach comes out as it must from our formulation also, i.e., proceed as if the (X_i, X_{i+1}) are (X_i, Y_i) in the i.i.d. regression model.

This should be a familiar space to at least one of us and, indeed, adaptation is not possible unless $\int l(\epsilon)v(\epsilon)dF(\epsilon) = 0$ for all $v \in V$. This happens if f is known to be symmetric about 0 but not, in general, otherwise. Thus, the "kernel plug in" estimate of $\int g(x)\lambda(x)dx$ given by Kwon (2000) is indeed efficient only in the Gaussian case, since only then is $l(\epsilon) \propto \epsilon$ so that the influence function $\epsilon\lambda(X)\pi^{-1}(X)$ belongs to the tangent space.

Thus if f is known but not Gaussian, one presumably needs to estimate g by a regularized version of the appropriate maximum likelihood, for instance by maximizing

$$\frac{1}{n-1}\sum_{i=1}^{n-1} \log f(X_{i+1} - g(X_i)) - \lambda_n \int [g'']^2(x)dx \tag{4.2}$$

for $\lambda_n \to 0$ appropriately, or by some variant of the method of sieves.

If f is unknown and we are in the adaptive case, say the distribution of ϵ symmetric about 0, it is then natural to begin by estimating g crudely as in the Gaussian case to get $\tilde{g}$, and then f by $\hat{f}$ symmetric about 0 based on the centered residuals $\hat{\epsilon}_{i+1} = \tilde{\epsilon}_{i+1} - \frac{1}{n}\sum_{j=1}^{n-1}\tilde{\epsilon}_{j+1}$ where $\tilde{\epsilon}_{i+1} = X_{i+1} - \hat{g}(X_i)$. Then, apply (4.2) with f replaced by $\hat{f}$ to get the final $\hat{g}$. Here, use of the centered residuals is

indeed necessary as Greenwood et al. point out so that the $\hat{\epsilon}_i$ empirically conform to the ϵ_i and have mean 0.

5. Response to Klaassen

As Klaassen points out, asymptotics should be used as a guide to small sample behavior. But what is small and which particular situations should we consider? Simulations can guide us but again are not determinative. To quote an anonymous observer, "There is no safety in numbers or in anything else!" His elegant Edgeworth approximation theorem is encouraging but it is not about small samples any more than first order asymptotics is.

We very much appreciate the elegant device of equal sample splitting introduced by Klaassen and Schick, and its superiority in this context to the small initial sample methods of Hájek (1962) and subsequently Bickel (1981). We believe its use in practice should be investigated further.

6. Response to McNeney and Wellner

McNeney and Wellner point us to their work on the non i.i.d. case which is, in that respect, considerably more extensive than ours. They also correct our loose talk about infinite dimensional parameters. We were, of course, aware that different topologies are possible but we think, although these determine what procedures are candidates, it's still fair to say that lower bounds on risk are determined essentially by the finite case. It was only in that sense that we would argue that the extension from finite to infinite dimensional regular parameters is relatively straightforward.

7. Response to Shen and Li

Shen and Li point to the development of Bernstein-von Mises theorems in terms of the negative examples, i.e., prior distributions that lead to posterior distributions that do not converge at the "optimal" minimax rate while penalized maximum likelihood does, and point to examples such as that of Cox (1993) where posterior confidence region based on the L_2 norm misbehave. Since penalized maximum likelihood can be viewed as a posterior mode with respect to a formal prior distribution, these arguments apply only to some priors. As the works of Shen and Wasserman (2001) and Zhao (2000) show, "satisfactory" asymptotic behavior can be achieved by priors placing "enough" mass in every neighborhood of any particular value.

This is not surprising in view of the effect of prior structure on Bayes consistency in the early work of Freedman (1965). And even the parametric Bernstein-von Mises theorem requires positive continuous densities in the neighborhood of every parameter value.

The "semiparametric" as opposed to the "nonparametric" aspect of Bayesian analysis, that is, not of the posterior of the whole object but of that of nice parameters has been studied for special priors such as Dirichlet processes where calculations are explicit, see Ferguson (1973). One would expect that the Bernstein-von Mises theorem in the classical sense would hold under less restrictive conditions. Specifically, if we are interested in $\mu(f) = \int xf(x)dx$, when is

$$E\left\{\int xf(x)dx \middle| X_1, \ldots, X_n\right\} = \bar{X} + o_p(n^{-1/2})?$$

Priors for which such result holds and for which posterior samples can be generated by MCMC might then be used to construct estimates which are efficient from a frequentist point yet, easier to calculate than using regularized maximum likelihood techniques for these large parameter spaces.

Shen and Li's discussion relating the information calculus to estimation equations is clear and attractive. Estimating equations are, however, more restrictive than may appear at first sight, at least if these as usual correspond to M estimates and are of the form

$$\int \psi(x, \theta)dP(x) = 0, \tag{7.3}$$

in which case they are linear in P. For example, if we consider the Cox proportional hazards model, we would conjecture there is no equation of form (7.3) which leads to an estimate of θ consistent whatever the unknown baseline hazard rate may be. The Cox estimate corresponds to an estimating equation that is nonlinear in P.

As a second example, consider the class of all equations based on functions $\psi(x-\theta)$ where x and θ are real and ψ is antisymmetric. These are essentially all the equations that maximum likelihood for a fixed shape lead to in the symmetric location model. If one calculates the information bound as Shen and Li suggest, we arrive at $I^{-1}(f)$ where $I(f) = \int (f'/f)^2 f$ and the optimal $\psi = -f'/f$. Unfortunately, ψ now depends on the unknown shape. Although a fixed estimating equation can give a locally efficient estimates for any given shape f, to fully adapt one needs to estimate f appropriately. This leads to estimating equations outside the framework since the ψ function is itself data-determined.

8. Response to van der Laan and Yu and Robins and Rotnitzky

These authors, as we indicated earlier, point to the fundamental issue of robustness. They present a large number of important examples where (a) there, in principle, exist asymptotically globally efficient estimates, but (b) we can expect the behavior of these to be very sensitive to slight violations of rather arbitrary

assumptions about the smoothness of the distribution of high dimensional covariates, and, on the other hand, (c) simple estimates which are locally efficient and insensitive to behavior of the covariates can be used.

At least one of us has great sympathy for this viewpoint developed by Robins, Rotnitzky, van der Laan and others in a remarkable way. Indeed, Bickel and Ritov (2000b) point out that, with the stringent Hampel view of robustness, one should knowingly limit oneself to estimation of parameters with uniformly bounded influence functions, which of course implies acceptance of the point of view that densities should be estimated only with fixed bias. Unfortunately, this point of view prevents us from talking about estimation of a real parameter θ in a model as simple as the semiparametric regression model discussed above. What Robins and collaborators isolate are remarkably large submodels of models such as (2.1) where use of low dimensional sub-submodels can produce estimating equations for θ which yield consistent asymptotically normal estimates of θ in the high dimensional submodel.

The most useful though not most general theorem (Theorem 2) cited by Robins and Rotnitzky, and also by van der Laan and Yu, which is due to Robins, Rotnitzky and van der Laan, gives simple conditions under which the following holds: if the parameter θ of interest is a function of two high dimensional (infinite dimensional) parameters κ and γ, then an estimate based on putting parametric models on both κ and γ is in fact robust under misspecification of *either* κ and γ.

A simple example of this phenomenon is estimation of θ in the semiparametric regression model (2.1), which we think is worth exhibiting explicitly. Here, assuming $R = aV + b + \epsilon'$ with ϵ' Gaussian and $\omega(V) = cV + d$ leads to the estimate

$$\hat{\theta} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(R_i - \hat{R}_i)}{\sum_{i=1}^{n}(R_i - \hat{R}_i)^2},$$

where $\hat{Y}_i$ is the fitted value of the linear regression Y on V and $\hat{R}_i$ that of R on V. It is clear that if the model for $\omega(V)$ is correct, then $E(\hat{\theta}|V_1, \ldots, V_n) = \theta$ and the estimate is $\sqrt{n}$ consistent. A slightly more involved calculation shows that if the model for R is correct, then $E(\hat{\theta}) = \theta + O(n^{-1/2})$ again because of the asymptotic orthogonality of the two factors which the conditions of Robins and Rotnitzky assure.

The theory explained in Theorems 1 and 2 as cited goes well beyond this elementary example. Of course, as Robins and Rotnitzky readily agree, if both specifications are wrong then the resulting estimate is asymptotically biased, but to quote another popular philosopher "There's no free lunch!" In any case, this work, apparently motivated by Robins' work on censored models, seems well worth pursuing.

9. Response to Zhang

Zhang's theorem, whose conditions are unfortunately missing, can be viewed as a linearization theorem for estimates of infinite dimensional parameters defined implicitly. This can be viewed as a special case of the estimates discussed in BKRW, pp. 370-371, where B is linear in b. (His $B_\theta(b)$ is $W_n(\nu, P)$ of BKRW with $\theta = \nu$ and $b = P$.) However, Zhang's theorem is evidently more "utility grade" than Theorem 7.6.2 of BKRW.

The problem of prediction in semiparametric models and the solution he discusses are intriguing. However, rather than viewing this as a problem of estimating $E_F(u(X, \theta; F))$ where (X, θ) has joint distribution $F \in \mathcal{F}$, it may be easier to simply define the parameter $F \to E_F(u(X, \theta; F)|X = \cdot)$ carrying F to functions of x and then apply the theory of Chapter 5 of BKRW.

Additional References

Abramson, I. and Goldstein, L. (1991). Efficient nonparametric testing by functional estimation. *J. Theoret. Probab.* **4**, 137-159.

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

Bhattacharya, R. and Lee, C. (1996). On geometric ergodicity of nonlinear autoregressive models. *Statist. Probab. Lett.* **22**, 311-315.

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647-671.

Bickel, P. J. (1993). Estimation in semiparametric models. In *Multivariate Analysis: Future Directions* (Edited by C. R. Rao), 55-73. North-Holland, Amsterdam.

Bickel, P. J., Götze, F. and van Zwet, W. R. (1986). The Edgeworth expansion for U-statistics of degree two. *Ann. Statist.* **14**, 1463-1484.

Bickel, P. J. and Klaassen, C. A. J. (1986). Empirical Bayes estimation in functional and structural models, and uniformly adaptive estimation of location. *Adv. Appl. Math.* **7**, 55-69.

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

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam: Research Papers in Proability and Statistics*. (Edited by David Pollard, Erik Torgersen, and Grace L. Yang), 55-88. Springer-Verlag, New York.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203-268. http://www.springer.de/link

Böhning, D. (1986). A vertex-exchnage method in D-optimal design theory. *Metrika* **33**, 337-347.

Böhning, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *J. Statist. Plann. Inference* **47**, 5-28.

Bradley, R. C. (1988a). On a theorem of Gordin. *Stochastics* **24**, 357-392.

Bradley, R. C. (1988b). On some results of M. I. Gordin: A clarification of a misunderstanding. *J. Theoret. Probab.* **1**, 115-119.

Brillinger, David R. (1983). A generalized linear model with "Gaussian" regressor variables. *A Festschrift for Erich L. Lehmann*, 97-114.

Brown, L. and Low, M. (1996). Asymptotic equivalence of non-parametric regression and white noise model. *Ann. Statist.* **24**, 2384-2398.

Bunea, F. (2000). A model selection approach to semiparametric regression. Technical Report **385**, University of Washington, Department of Statistics. http://www.stat.washington.edu/www/research/reports/#2000.

Carroll, R. J., Fan, J., Gijbels, I, and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Satist. Assoc.* **92**, 477-489.

Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998). Nonparametric estimation via local estimating equations. *Jour. Ameri. Statist. Assoc.* **93**, 214-227.

Cox, D. D. (1993). An analysis of Bayesian inference for non-parametric regression. *Ann. Statist.* **21**, 903-924.

Does, R. J. M. M. (1983). An Edgeworth expansion for simple linear rank statistics under the null hypothesis. *Ann. Statist.* **11**, 607-624.

Dominici, F. and Zeger, S. L. (2001). Smooth quantile ratio estimation. Submitted.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-1224.

Duan, Naihua, Li, Ker-Chau. (1987). Distribution-free and link-free estimation for the sample selection model. *J. Econometrics* **35**, 25-35.

Duan, Naihua, Li, Ker-Chau. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* **19**, 505-530.

Dürr, D. and Goldstein, S. (1986). Remarks on the central limit theorem for weakly dependent random variables. In *Stochastic Processes – Mathematics and Physics* (Edited by S. Albeverio, P. Blanchard and L. Streit), 104-118. Lecture Notes in Mathematics 1158, Springer, Berlin.

Efromovich, S. (1998). On global and pointwise adaptive estimation. *Bernoulli* **4**, 273-282.

Efromovich, S. and Pinsker, M. S. (1984). An adaptive algorithm of nonparametric filtering. *Automation and Remote Control* **11**, 58-65.

Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of additive and linear components for high dimensional data. *Ann. Statist.* **26**, 943-971.

Fan, J. and Huang, L. (2001). Goodness-of-fit test for parametric regression models. *J. Amer. Statist. Assoc.* In press.

Fan, J., Zhang, C. M. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.*, **29**, 153-193.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

Genovese, C., Shen, X. and Wasserman, L. (2001). Coverage of Bayesian confidence regions and sieves for infinitely many parameters. Manuscript in preparation.

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

Gill, R. D. and van der Vaart, A. W. (1993). Non- and semi-parametric maximum likelihood estimators and the von Mises method - II. *Scand. J. Statist.* **20**, 271-288.

Gill, R. D., van der Laan, M. J. and Robins, J. M. (1997). Coarsening at Random: Characterizations, Conjectures and Counter-Examples. *Proceedings of the First Seattle Symposium in Biostatistics* (Edited by D. Y. Lin and T. R. Fleming), 255–294, Springer Lecture Notes in Statistics.

Gill, R. D., van der Laan, M. J. and Robins, J. M. (2000). Locally efficient estimation in censored data models. Technical report, Division of Biostatistics, University of California, Berkeley.

Goldstein, L. and Khas'minskii, R. (1995). On efficient estimation of smooth functionals. *Theory Probab. Appl.* **40**, 151-156.

Goldstein, L. and Messer, K. (1992). Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.* **20**, 1306-1328.

Gordin, M. I. (1969). The central limit theorem for stationary processes. *Soviet Math. Dokl.*, **10**, 1174-1176.

Greenwood, P. E., McKeague, I. W. and Wefelmeyer, W. (1998). Information bounds for Gibbs samplers. *Ann. Statist.* **26**, 2128-2156.

Greenwood, P. E. and Wefelmeyer, W. (1994a). Nonparametric estimators for Markov step processes. *Stochastic Process. Appl.* **52**, 1-16.

Greenwood, P. E. and Wefelmeyer, W. (1994b). Optimality properties of empirical estimators for multivariate point processes. *J. Multivariate Anal.* **49**, 202-217.

Greenwood, P. E. and Wefelmeyer, W. (1995). Efficiency of empirical estimators for Markov chains. *Ann. Statist.* **23**, 132-143.

Greenwood, P. E. and Wefelmeyer, W. (1996). Empirical estimators for semi-Markov processes. *Math. Methods Statist.* **5**, 299-315.

Greenwood, P. E. and Wefelmeyer, W. (1999a). Reversible Markov chains and optimality of symmetrized empirical estimators. *Bernoulli* **5**, 109-123.

Greenwood, P. E. and Wefelmeyer, W. (1999b). Characterizing efficient empirical estimators for local interaction Gibbs fields. *Stat. Inference Stoch. Process.* **2**, 119-134.

Greenwood, P. E. and Wefelmeyer, W. (2001). Empirical estimators based on MCMC data. To appear in *Handbook of Statistics* **21** (Edited by D. Shanbhag). Elsevier, Amsterdam.

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

Hájek, J. (1962). Asymptotically most powerful rank-order tests. *Ann. Math. Statist.* **33**, 1124-1147.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19**, 2244-2253.

Helmers, R. (1980). Edgeworth expansions for linear combinations of order statistics with smooth weight functions. *Ann. Statist.* **8**, 1361-1374.

Heyde, C. C. (1997). *Quasi-likelihood and its Applications.* Springer, New York.

Hoffmann-Jøgensen, J. (1984). *Stochastic Processes on Polish Spaces.* unpublished.

Hoffmann-Jøgensen, J. (1991). *Stochastic Processes on Polish Spaces, Various Publication Series* **39**. Aarhus Universitet, Aarhus, Denmark.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.

Huang, T. M. (2000). Convergence rates for posterior distributions and adaptive estimation. Technical report 711, Department of Statistics, Carnegie Mellon University.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Ingster, Yu. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives I–III. *Math. Methods Statist.* **2**, 85-114; **3**, 171-189; **4**, 249-268.

Jacobsen, M. and Keiding, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.* **23**, 774-786.

Jacod, J. and Shiryaev, A. N. (1987). *Limit Theorems for Stochastic Processes.* Grundlehren der Mathematischen Wissenschaften 288, Springer, Berlin.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *J. Comp. Graph. Statist.* **7**, 310-321.

Kartashov, N. V. (1985). Criteria for uniform ergodicity and strong stability of Markov chains with a common phase space. *Theory Probab. Math. Statist.* **30**, 71-89.

Kartashov, N. V. (1996). *Strong Stable Markov Chains*. VSP, Utrecht.

Kessler, M., Schick, A. and Wefelmeyer, W. (2001). The information in the marginal law of a Markov chain. *Bernoulli* **7**, 243-266.

Klaassen, C. A. J. (1979). Nonuniformity of the convergence of location estimators. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics* 21-25 August 1978 (Edited by P. Mandl, M. Hušková), 251-258. North-Holland, Amsterdam.

Klaassen, C. A. J. (1980). Statistical Performance of Location Estimators, PhD-thesis, University of Leiden, also published in 1981 as Mathematical Centre Tracts 133, Mathematisch Centrum, Amsterdam.

Klaassen, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* **15**, 1548-1562.

Klaassen, C. A. J., Mokveld, P. J. and van Es, A. J. (2000). Squared skewness minus kurtosis bounded by 186/125 for unimodal distributions. *Statist. Probab. Lett.* **50**, 131-135.

Klaassen, C. A. J. and Lenstra, A. J. (2000). Zero information in the mixed proportional hazards model. Preprint, Korteweg-de Vries Institute, University of Amsterdam. http://preprint.beta.uva.nl/

Klaassen, C. A. J. and Putter, H. (1997). Efficient estimation of the error distribution in a semiparametric linear model. In *Contemporary Multivariate Analysis and Its Applications* (Edited by K. T. Fang and F. J. Hickernell). Hong Kong Baptist University.

Klaassen, C. A. J. and Putter, H. (2000). Efficient estimation of Banach parameters in semiparametric models. Technical Report, Department of Mathematics, University of Amsterdam.

Koshevnik, Y. (1996). Semiparametric estimation of a symmetric error distribution from regression models. *Publ. Inst. Statist. Univ. Paris* **40**, 77-91.

Koul, H. L. and Schick, A. (1997). Efficient estimation in nonlinear autoregressive time-series models. *Bernoulli* **3**, 247-277.

Künsch, H. R. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.* **12**, 843-863.

Kuo, H. H. (1975). *Gaussian measures on Banach spaces*. Lecture notes in Mathematics. **463**, Springer, Berlin.

Kutoyants, Yu. A. (1994). *Identification of Dynamical Systems with Small Noise*. Mathematics and its Applications 300, Kluwer, Dordrecht.

Kutoyants, Yu. A. (1997). Some problems of nonparametric estimation by observations of ergodic diffusion process. *Statist. Probab. Lett.* **32**, 311-320.

Kutoyants, Yu. A. (1998). On density estimation by the observations of ergodic diffusion processes. In *Statistics and Control of Stochastic Processes* (Edited by Y. M. Kabanov, B. L. Rozovskii and A. N. Shiryaev), 253-274. World Scientific, Singapore.

Kutoyants, Yu. A. (1999). Efficient density estimation for ergodic diffusion processes. *Stat. Inference Stoch. Process.* **1**, 131-155.

van der Laan, M. J. (1995). An identity for the nonparametric maximum likelihood estimator in missing data and biased sampling models. *Bernoulli* **1**, 335-341.

van der Laan, M. J. and Robins, J. M. (1998). Locally Efficient Estimation with Current Status Data and Time-Dependent Covariates. *J. Amer. Statist. Assoc.* **93**, 693–701.

van der Laan, M. J. and Robins, J. M. (2001). Unified methods for censored longitudinal data and causality. To appear.

van der Laan, M. J. and van der Vaart, A. W. (2001). Estimating a survival distribution with current status data and high-dimensional covariates. To be submitted.

Laurent, B. (1997). Estimation of integral functionals of a density and its derivatives. *Bernoulli* **3**, 181-211.

Lenstra, A. J. (1998). Analyses of the nonparametric mixed proportional hazards model. PhD-thesis, University of Amsterdam.

Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25**, 2512-2546.

Levit, B. Y. (1975). Conditional estimation of linear functionals. *Problems Inform. Transmission* **11**, 39-54.

Li, B. (2000). Nonparametric estimating equations based on a penalized information criterion. *Canad. J. Statist.* **28**, 621-639.

Li, B. (2001). *Asymptotic Analysis of Estimating Equations*. Book in preparation.

Li, B. (2001). On quasilikelihood equations with nonparametric weights. To appear in *Scand. J. Statist.*

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Amer. Statist. Assoc.* **84**, 1074-1078.

Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Amer. Statist. Assoc.* **94**, 1147-1160.

Maercker, G. (1997). *Statistical Inference in Conditional Heteroskedastic Autoregressive Models*. Shaker, Aachen.

Maigret, N. (1978). Théorème de limite centrale fonctionnel pour une chaîne de Markov récurrente au sens de Harris et positive. *Ann. Inst. H. Poincaré Probab. Statist.* **14**, 425-440.

McNeney, B. and Wellner, J. A. (2000). Application of convolution theorems in semiparametric models with non-i.i.d. data. *J. Statist. Plann. Inference* **91**, 441-480.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2001a). Plug-in estimators in semiparametric stochastic process models. To appear in *Selected Proceedings of the Symposium on Inference for Stochastic Processes* (Edited by I. V. Basawa, C. C. Heyde and R. L. Taylor). IMS Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward, California.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2001b). Improved estimators for constrained Markov chain models. To appear in *Statist. Probab. Lett.*

Müller, U. U., Schick, A. and Wefelmeyer, W. (2001c). Estimating linear functionals of the error distribution in nonparametric regression. Technical Report, Department of Mathematical Sciences, Binghamton University. http://math.binghamton.edu/anton/preprint.html

Murphy, S. A. and Van der Vaart, A. W. (1996). Likelihood inference in the errors-in-variables model. *J. Multivariate Anal.* **59**, 81-108.

Newey, W. K. (1990). Semiparametric Efficiency Bounds. *J. Appl. Econometrics* **5**, 99-135.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Penev, S. (1990). Convolution theorem for estimating the stationary distribution of Markov chains. *C. R. Acad. Bulgare Sci.* **43**, 29-32.

Penev, S. (1991). Efficient estimation of the stationary distribution for exponentially ergodic Markov chains. *J. Statist. Plann. Inference* **27**, 105-123.

Peng, H. and Schick, A. (2001). Efficient estimation of linear functionals of a bivariate distribution with equal but unknown marginals: The least squares approach. Technical Report, Department of Mathematical Sciences, Binghamton University.

Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York.

Pfanzagl, J. (1993). Incidental versus random nuisance parameters. *Ann. Statist.* **21**, 1663-1691.

Pinsker, M. S. (1981). Optimal filering of square integrable signals in Gaussian white noise (in Russian). *Problems Information Transmission* **16**, No. 2, 52-68.

Rieder, H. (1994). *Robust Asymptotic Statistics*. Springer-Verlag, New York.

Ritov, Y., and Bickel, P. J. (1987). Achieving Information Bounds in Non and Semiparametric Models. Technical Report No. 116, University of California, Berkeley, Department of Statistics.

Ritov, Y. and Wellner, J. A. (1988). Censoring, martingales, and the Cox model. In *Statistical Inference from Stochastic Processes* (N. U. Prabhu, ed.), 191-219. Contemporary Mathematics 80, American Mathematical Society, Providence, Rhode Island.

Robbins, H. (1977). Prediction and estimation for the compound Poisson distribution. *Proc. Nat. Acad. Sci. USA* **74**, 2670-2671.

Robbins, H. (1988). The u, v method of estimation. In *Statistical Decision Theory and Related Topics IV* **1** (Edited S. S. Gupta and J. O. Berger), 265-70. Springer-Verlag, New York.

Robins, J. M. and Ritov, Y. (1997). A curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statist. in Medicine* **16**, 285-319.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the Journal of the American Statistical Association*. To appear.

Robins, J. M. and Rotnitzky, A. (2001a). Double robustness in statistical models. *J. Statist. Plann. Inference*. To appear.

Robins, J. M. and Rotnitzky, A. (2001b). Testing and estimation of direct effects by reparameterized directed acyclic graphs with structural nested models. *Ann. Statist.* To appear.

Robins, J. M., Rotnitzky, A., and van der Laan, M. (2000). Comment on "On Profile Likelihood" by S. A. Murphy and A. W. van der Vaart. *J. Amer. Statist. Assoc.* **95**, 431-435.

Robbins, H. and Zhang, C.-H. (2000). Efficiency of the u,v method of estimation. *Proc. Nat. Acad. Sci. USA* **97**, 12976-12979.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *Aids Epidemiology, Methodological issues*. Birkhäuser.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41-55.

Rotnitzky, A., Robins, J. M. and Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Amer. Statst. Assoc.* **93**, 1321-1339.

Ruud, P. A. (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica* **51**, 225-228.

Ruud, P. A. (1986). Consistent estimation of limited dependent variable models despite misspecification of distribution. *J. Econometrics* **32**, 157-187.

Saavedra, A. and Cao, R. (1999). Rate of convergence of a convolution-type estimator of the marginal density of an MA(1) process. *Stochastic Process. Appl.* **80**, 129-155.

Saavedra, A. and Cao, R. (2000). On the estimation of the marginal density of a moving average process. *Canad. J. Statist.* **28**, 799-815.

Scharfstein D. O., Rotnitzky, A. and Robins J. M. (1999). Rejoinder to "Adjusting for nonignorable drop-out using semiparametric non-response models." *J. Amer. Statist. Assoc.* **94**, 1135-1146.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14**, 1139-1151.

Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.* **21**, 1486-1521. Correction and addendum **23** (1995), 1862-1863.

Schick, A. (1994). On efficient estimation in regression models with unknown scale functions. *Math. Methods Statist.* **3**, 171-212.

Schick, A. (1999a). Efficient estimation in a semiparametric autoregressive model. *Stat. Inference Stoch. Process.* **2**, 69-98.

Schick, A. (1999b). Efficient estimation in a semiparametric heteroscedastic autoregressive model. Technical Report, Department of Mathematical Sciences, Binghamton University. http://math.binghamton.edu/anton/preprint.html

Schick, A. (2001). Sample splitting with Markov chains. *Bernoulli* **7**, 33-61.

Schick, A. and Wefelmeyer, W. (1999). Efficient estimation of invariant distributions of some semiparametric Markov chain models. *Math. Methods Statist.* **8**, 119-134.

Schick, A. and Wefelmeyer, W. (2001a). Estimating joint distributions of Markov chains. To appear in *Statist. Inference Stoch. Process.*

Schick, A. and Wefelmeyer, W. (2001b). Estimating the innovation distribution in nonlinear autoregressive models. To appear in *Ann. Inst. Statist. Math.*

Schick, A. and Wefelmeyer, W. (2001c). Efficient estimation in invertible linear processes. Technical Report, Department of Mathematical Sciences, Binghamton University. http://math.binghamton.edu/anton/preprint.html

Schick, A. and Wefelmeyer, W. (2001d). Estimating invariant laws of linear processes by U-statistics. Technical Report, Department of Mathematical Sciences, Binghamton University. http://math.binghamton.edu/anton/preprint.html

Schick, A. and Wefelmeyer, W. (2001e). Root n consistent and optimal density estimators for moving average processes. Technical Report, Department of Mathematical Sciences, Binghamton University. http://math.binghamton.edu/anton/preprint.html

Severini, T. A. and Wong, W. H. (1992). Generalized profile likelihood and conditional parametric models. *Ann. Statist.* **20**, 1768-1802.

Shen, L. Z. (1995). On optimal b-robust influence functions in semiparametric models. *Ann. Statist.* **23**, 968-989.

Shen, X., and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29**, to appear.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Royal Statist. Soc. Ser. B*, **50**, 413-436.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24**, 2477-2498.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14**, 590-606.

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

Tsai, W.-Y. and Zhang, C.-H. (1995). Asymptotic properties of nonparametric maximum likelihood estimator for interval-truncated data. *Scand. J. Statist.* **22**, 361-370.

Vardi, Y. and Zhang, C.-H. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *Ann. Statist.* **20**, 1022-1039.

Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In *Practical Nonparametric and Semiparametric Bayesian Statistics.* (Edited by D. Dey, P. Müller and D. Sinha). Springer, New York.

Wefelmeyer, W. (1994). An efficient estimator for the expectation of a bounded function under the residual distribution of an autoregressive process. *Ann. Inst. Statist. Math.* **46**, 309-315.

Wefelmeyer, W. (1996). Quasi-likelihood models and optimal inference. *Ann. Statist.* **24**, 405-422.

Wefelmeyer, W. (1999). Efficient estimation in Markov chain models: an introduction. In *Asymptotics, Nonparametrics, and Time Series* (Edited by S. Ghosh), 427-459. Statistics: Textbooks and Monographs 158, Dekker, New York.

Yatchew, A. (1997). An elementary estimator of the partial linear model. *Econom. Lett.* **57**, 135-143.

Zhan, Y. (1999). Central limit theorems for functional Z-estimators. To appear in *Statist. Sinica*.

Zhang, C.-H. (2000). General empirical Bayes wavelet methods. Tech. Report 2000-007, Department of Statistics, Rutgers University, Piscataway, New Jersey, U.S.A. http://stat.rutgers.edu/~cunhui/papers/

Zhang, C.-H. (2001). Efficient estimation of sums of random variables. Tech. Rep. 2001-005, Department of Statistics, Rutgers University, Piscataway, New Jersey, U.S.A. http://stat.rutgers.edu/~cunhui/papers/

Zhang, C.-H. and Li, X. (1996). Linear regression with doubly censored data. *Ann. Statist.* **24**, 2720-2743.

Zhang, C. M. (2000). Adaptive tests of regression functions via multi-scale generalized likelihood ratios. Technical Report #1026, Department of Statistics, University of Wisconsin–Madison.

Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28**, 532-552.

Zhuo, Y. and van der Laan, M. J. (2001). Locally efficient estimation in semiparametric regression models. In progress.