Statistica Neerlandica (1995) Vol. 49, nr. 1, pp. 111–125

Semiparametric models: an evaluation

We review ten years of work on semiparametric theory in statistics on the occasion of the recently published book by Bickel, Klaassen, Ritov and Wellner.

1 Introduction

Semiparametric models have been a hot topic in research in statistics and related fields for roughly ten years. The book 'Efficient and Adaptive Estimation for Semiparametric Models' (BKRW) by BICKEL, KLAASSEN, RITOV and WELLNER is the first large monograph on this topic. Though it appeared only in 1993, preliminary versions of it were referenced in papers as early as 1988. In this paper we review key concepts of semiparametric theory and at the same time review the book BKRW.

The question as to what constitutes a semiparametric model is actually not easy to answer. Different terms that are meant to cover roughly the same area, but put different emphasis, are parametric-nonparametric, BEGUN, HALL, HUANG and WELLNER (1983), general statistical model, PFANZAGL and WEFELMEYER (1983), and model with large parameter space, VAN DER VAART (1988).

In the situation where the observations consist of a random sample from a common distribution P on some sample space, the situation uniquely considered in BKRW, the *model* is simply the set $\mathscr{P}$ of all possible values of P: a set of probability measures. Then a semiparametric model might be described as not parametric and not nonparametric. Here a *parametric model* could be defined as a model $\{P_\theta: \theta \in \Theta\}$ indexed by a parameter θ ranging over a subset of Euclidean space, such that the total variation distance $\|P_{\theta_1} - P_{\theta_2}\|$ is of the order $\|\theta_1 - \theta_2\|^\alpha$ for some $\alpha > 0$ as $\theta_1 - \theta_2 \to 0$. Strictly speaking the only *nonparametric model* is the set of all probability measures, but models $\mathscr{P}$ that are restricted only in a qualitative manner, such as finite moments or absolute continuity, are also considered nonparametric. Semiparametric models are the intermediate models: larger than parametric, but smaller than nonparametric.

Often interest in semiparametric models is focused on values of some function $\nu: \mathscr{P} \to \mathbb{R}^k$ on the model, in which case the remainder of the model is called a *nuisance parameter*. In particular, the model may have a natural parametrization $(\theta, G) \to P_{\theta,G}$, where $\theta = \nu(P_{\theta,G})$ is Euclidean and G runs itself through a nonparametric class of distributions. This gives a semiparametric model in the strict sense. Examples, such as given in Section 2, make these distinctions clearer.

At the present time there seems to be a reasonable understanding of 'information bounds' for estimating 'smooth' functions of P. Such 'lower bounds' are asymptotic in nature with the number of observations tending to infinity. They extend the theory for classical, smooth parametric models, which goes back to FISHER (1922), CRAMER (1946) and HAJEK (1970). Five out of the six major chapters of BKRW are concerned with information bounds.

Even after ten years the extent to which the information bounds are sharp is unclear and general methods to construct efficient estimators are largely undeveloped. It seems certain that the classical theory for parametric models, as developed for instance by LE CAM (1960), has no easy parallel for semiparametric models. Only the last chapter of BKRW, still almost a quarter of the book, is devoted to the construction of estimators; it is the least polished part of the book. Many new developments are to be expected in this area, a number of which have already taken place since BKRW went in print (e.g. VAN DER LAAN, 1993, MURPHY, 1992, VAN DER VAART, 1992).

Section 1.2 of BKRW presents a list of subjects that are not covered. This includes statistical testing within models, goodness-of-fit testing, robustness considerations, boot-strapping and non-i.i.d. observations. Of these goodness-of-fit and robustness are the most important to be further developed. Since semiparametric models are much larger than parametric ones, they will general fit better on given data-sets. However, even a semiparametric model can only yield an approximation to reality. Checking for goodness-of-fit should be a standard ingredient in applying a semiparametric model. Giving up some efficiency in order to allow for small deviations from the model may be worthwhile. Being concerned with information bounds in more than three quarters of the book, the authors of BKRW remark that even if the prime interest is not in going for full efficiency, the information bounds are still of much interest. They show exactly how much efficiency might be sacrificed by using for instance an easy-to-apply or robust method.

The book BKRW is restricted to the estimation of 'smooth' functions of the model. This concerns situations in which optimal estimators converge at a $n^{-1/2}$-rate and are asymptotically normally distributed. In the case of parametric models such situations vastly outnumber the 'nonregular' ones, such as estimating the support boundaries of uniformly distributed observations. In contrast, the case may be made that in semiparametric theory the nonregular functionals outnumber the regular ones. Then the emphasis on regular functionals might still be defended by claiming that these are the more interesting ones. It is fair to say that mathematically they are the easier ones. Nonregular functions include for instance most objects of study in the smoothing literature, such as densities and regression functions in infinite dimensional settings. More interestingly, nonregular functions occur next to smooth functionals in almost every semiparametric model, arising as natural functions of parameters that describe the model. These situations, where the map from the underlying distribution of the observations into the parameter is not differentiable, are sometimes called *inverse problems*. They may be among the most interesting areas for semiparametric research in the near future. The theory of both lower and upper bounds for estimation is largely undeveloped.

2 Some examples

Semiparametric models have been put forward by researchers from such diverse fields as biostatistics, econometrics, demography and spectroscopy. The following list of examples shows some of the scope of semiparametric theory. The book BKRW describes many more examples and the level of detail explains in part the length of the book.

ı bounds are sharp is unclear and
undeveloped. It seems certain that
ɔr instance by LE CAM (1960), has
chapter of BKRW, still almost a
nators; it is the least polished part
this area, a number of which have
DER LAAN, 1993, MURPHY, 1992,

ot covered. This includes statistical
onsiderations, boot-strapping and
tness are the most important to be
larger than parametric ones, they
miparametric model can only yield
should be a standard ingredient in
ency in order to allow for small
erned with information bounds in
RW remark that even if the prime
bounds are still of much interest.
rificed by using for instance an

ooth' functions of the model. This
: a $n^{-1/2}$-rate and are asymptotically
ch situations vastly outnumber the
ındaries of uniformly distributed
niparametric theory the nonregular
ıasis on regular functionals might
interesting ones. It is fair to say
lar functions include for instance
; densities and regression functions
nregular functions occur next to
lel, arising as natural functions of
here the map from the underlying
differentiable, are sometimes called
g areas for semiparametric research
r bounds for estimation is largely

ırchers from such diverse fields as
ɔy. The following list of examples
book BKRW describes many more
gth of the book.

In the description of the examples X denotes a typical observation. Random vectors Y, Z, e and f are used to describe the model, but are not necessarily observed. The parameters θ and ν are always Euclidean. The following descriptions do not include censoring mechanisms of the data, though this would be natural in many examples.

EXAMPLE 1 (REGRESSION). Let (Y, Z) and e be independent random vectors satisfying the relationship $Y = \mu(Z, \theta) + \sigma(Z, \theta)e$ for given functions μ and σ that are known up to θ. The observation is the pair $X = (Y, Z)$. If the distribution of e is known to belong to a certain parametric family, such as the family of $N(0, \sigma^2)$ distributions, and the Z are modelled as constants, this is just a classical regression model, allowing for heteroscedasticity. A semiparametric version is obtained by letting the distribution of e range over all distributions on the real line, or, alternatively, all distributions that are symmetric about zero.

EXAMPLE 2 (PROJECTION PURSUIT REGRESSION). Once again let (Y, Z) and e be independent random vectors and let $X = (Y, Z)$. Now assume that $Y = r(\theta^T Z) + e$ for a function r ranging over a set of (smooth) functions and e having a $N(0, \sigma^2)$-distribution. In this model θ and r are confounded, but the direction of θ is estimable up to its sign.

EXAMPLE 3 (LOGISTIC REGRESSION). Given a vector Z let the random variable Y take the value 1 with probability $e^{r(Z)}/(1 + e^{r(Z)})$ and be 0 otherwise. Let $Z = (Z_1, Z_2)$ and let the function r be of the form $r(z_1, z_2) = \tau(z_1) + \theta^T z_2$. Observed is the pair $X = (Y, Z)$.

EXAMPLE 4 (PAIRED EXPONENTIAL). Given a variable Z with completely unknown distribution let $X = (X_1, X_2)$ be a vector of independent exponentially distributed random variables with parameters Z and $Z\theta$. The interest is in the ratio θ of the conditional hazard rates of X_1 and X_2.

EXAMPLE 5 (ERRORS-IN-VARIABLES). The observation is a pair $X = (X_1, X_2)$ where $X_1 = Z + e$ and $X_2 = \alpha + \beta Z + f$ for a bivariate normal vector (e, f) with mean zero and unknown covariance matrix. Thus X_2 is a linear regression on a variable Z which is observed with error. The distribution of Z is unknown.

EXAMPLE 6 (TRANSFORMATION REGRESSION). Suppose that $X = (Y, Z)$ where the random vectors Y and Z are known to satisfy $\tau(Y) = \theta^T Z + e$ for an unknown map τ and independent random vectors e and Z with known or parametrically specified distributions. The transformation τ ranges over an infinite dimensional set.

EXAMPLE 7 (TRANSFORMATION). Suppose that $X = (Y, Z)$ where the conditional distribution of $\tau(Y)$ given Z belongs to a parametric model $\{P_\theta(\cdot|Z): \theta \in \Theta\}$. The unknown transformation τ ranges over an infinite dimensional set.

EXAMPLE 8 (COX). The observation is a pair $X = (T, Z)$. The distribution of Z is unknown and the conditional cumulative hazard function of T given Z is of the form $e^{\theta^T Z}\Lambda(t)$ for Λ being a completely unknown cumulative hazard function. If Z_i is a $0-1$ variable then e^{θ_i} can be interpreted as the ratio of the hazards of two individuals who have $Z_i = 1$ and $Z_i = 0$, respectively, but who are identical otherwise.

EXAMPLE 9 (COPULA). The observation X is two-dimensional with cumulative distribution function of the form $C_\theta(G_1(x_1), G_2(x_2))$ for a parametric family of cumulative hazard functions C_θ on the unit square with uniform marginals. The marginal distribution functions G_i may be completely unknown or known.

EXAMPLE 10 (FRAILTY). The conditional cumulative hazard function of Y given (Z, W) is of the form $We^{\theta^T Z}\Lambda(y)$. The random variable W possesses a Gamma (ν, ν) distribution and is independent of the variable Z which possesses a completely unknown distribution. The observation is $X = (Y, Z)$.

EXAMPLE 11 (INTERVAL CENSORING). A 'death' that occurs at time T is only observed to have taken place or not at a check-up time C. The observation is $X = (C, 1\{T \le C\})$ and T and C are assumed independent with completely unknown or partially specified distributions.

3 Tangent spaces

The *tangent space* is a key concept in semiparametric and a substantial part of BKRW is devoted to computing and studying properties of tangent spaces for specific examples. Information bounds are expressed in terms of tangent spaces.

It is convenient to develop this concept first for parametric models. Assume that a parametric model with parameter θ ranging over an open subset of $\mathbb{R}^k$ is described by densities p_θ with respect to some measure μ. In BKRW the model is called *regular* if for all θ there exists a vector-valued measurable map $\dot{l}_\theta$ such that as $h \to 0$

$$\int [p_{\theta+h}^{1/2} - p_\theta^{1/2} - \tfrac{1}{2} h^T \dot{l}_\theta p_\theta^{1/2}]^2 \, d\mu = o(\|h\|^2),$$

$$\int \|\dot{l}_{\theta+h} p_{\theta+h}^{1/2} - \dot{l}_\theta p_\theta^{1/2}\|^2 \, d\mu = o(1),$$

$$I_\theta = E_\theta \dot{l}_\theta(X) \dot{l}_\theta^T(X) \quad \text{is nonsingular.}$$

In most situations $\dot{l}_\theta$ is the vector of classical score functions of the model and its value at x can be computed as the gradient

$$\dot{l}_\theta(x) = \frac{\partial}{\partial \theta} \log p_\theta(x).$$

The matrix I_θ is the Fisher information matrix. The point of taking a derivative in mean square (of the root density, not the logarithm) in the definition of a regular model is that this is exactly right for the theory of asymptotic efficiency.

Define an estimator $T_n = T_n(X_1, \ldots, X_n)$ as a measurable map of the observations (an i.i.d. sample of size n). Then an asymptotically efficient estimator sequence for the parameter in a regular parametric model must have the property that $\sqrt{n}(T_n - \theta)$ is asymptotically normal $N(0, I_\theta^{-1})$ distributed under θ. This statement may be split into a 'lower bound' assertion, saying that the limit distribution cannot be 'better' than the given normal one, and an upper bound or attainability assertion, saying that there exist estimators with this normal limit

ısional with cumulative distribution
amily of cumulative hazard functions
inal distribution functions G_i may be

ard function of Y given (Z, W) is of
; a Gamma (v, v) distribution and is
npletely unknown distribution. The

rs at time T is only observed to have
ion is $X = (C, 1\{T \leq C\})$ and T and
ı or partially specified distributions.

and a substantial part of BKRW is
ngent spaces for specific examples.
spaces.
parametric models. Assume that a
ı subset of $\mathbb{R}^k$ is described by densities
is called *regular* if for all θ there exists

ctions of the model and its value at x

t of taking a derivative in mean square
of a regular model is that this is exactly

able map of the observations (an i.i.d.
nator sequence for the parameter in a
$\sqrt{n}(T_n - \theta)$ is asymptotically normal
split into a 'lower bound' assertion,
n the given normal one, and an upper
ist estimators with this normal limit

distribution. Rigorous statements of the lower bound, in terms of a local asymptotic minimax and convolution theorem, are due to CHERNOFF (1956), HAJEK (1970), HAJEK (1972) and LE CAM (1972), but an informal statement goes back to Fisher (FISHER, 1922). Fisher also claimed the asymptotic efficiency of the maximum likelihood estimator, which was later rigorously proved under regularity conditions by Cramér (CRAMÉR, 1946), among others. Le Cam (LE CAM, 1956) showed the existence of asymptotically $N(0, I_\theta^{-1})$ estimators in regular parametric models under the only further (necessary) condition that the parameter is identifiable. (Actually the second requirement in the definition of a regular parametric model in BKRW is necessary neither for lower bound nor for attainability.)

While lower bound results have been extended to semiparametric models, a generalization of the attainability proved in LE CAM (1956), or even CRAMÉR (1946), is lacking.

The tangent space at a fixed element P_θ of a parametric model is defined as the linear space $\{h^T \dot{l}_\theta : h \in \mathbb{R}^k\}$ spanned by the score functions. A general statistical model $\mathcal{P}$ can be viewed as a union of finite dimensional 'submodels' and its tangent space is defined as a union of finite dimensional tangent spaces. More precisely, in BKRW the *tangent set* $\dot{\mathcal{P}}_0$ at a fixed element P_0 of the model is defined as the union of the tangent spaces (at P_0) of all one-dimensional regular submodels passing through P_0. The *tangent space* $\dot{\mathcal{P}}$ (at P_0) is defined as the closure of the linear span of the tangent set. Here the closure is taken in the Hilbert space $L_2(P_0)$ of functions with finite second moment, equipped with norm and inner product given by

$$\|h\|_{P_0} = \left(\int h^2 \, dP_0\right)^{1/2}; \quad \langle h_1, h_2 \rangle_{P_0} = \int h_1 h_2 \, dP_0.$$

Because of the insistence on regular submodels the definition of a tangent space in BKRW differs from that used by others LEVIT (1978), PFANZAGL and WEFELMEYER (1983), VAN DER VAART (1988), among others in not admitting one-sided derivatives as scores. The benefits of this difference are not clear to me, even more so since one-sided derivatives are important later on (See page 306).

Just as the definition of tangent spaces, lower bounds for estimation in semiparametric models are based on finite dimensional submodels. Estimation of some aspect of P is clearly not easier when knowing that P belongs to the model $\mathcal{P}$ than when knowing that P belongs to a given parametric submodel. Therefore the supremum of the lower bounds resulting from all (regular) submodels yields a lower bound for the semiparametric model. There is nothing in this simple argument, usually attributed to Stein (STEIN, 1956), that suggests that the lower bound obtained in this manner is sharp. In general it is not, but in many interesting cases it has been shown to be sharp by explicit construction of estimators that have asymptotic variance equal to the bound.

We close this section with two examples of tangent spaces.

EXAMPLE 12 (NONPARAMETRIC MODELS). Suppose $\mathcal{P}$ consists of all probability laws on the sample space. Then the tangent space consists of all measurable functions g satisfying $\int g \, dP_0 = 0$ and $\int g^2 \, dP_0 < \infty$.

It suffices to exhibit suitable one-dimensional submodels. For a bounded function g consider for instance the exponential family $p_\theta(x) = c(\theta) \exp(\theta g(x)) p_0(x)$ or, alternatively, the model

$p_\theta(x) = (1 + \theta g(x))p_0(x)$. Both have score function $g - \int g \, dP_0$ at $\theta = 0$. Both submodels are of the form $p_\theta(x) = c(\theta)\psi(\theta)\psi(\theta g(x))p_0(x)$ for a nonnegative function ψ with $\psi(0) = \psi'(0) = 1$. The function $\psi(x) = (1 + e^{-2x})^{-1}$ is bounded and can be used with any g.

EXAMPLE 13 (INFORMATION LOSS MODELS). Suppose the common distribution of the observations is the distribution of a measurable transformation $X = m(Y)$ of an (unobservable) variable Y. Assume that the form of m is known and that the distribution of Y is known to belong to a class $\mathcal{G}$. This yields a natural parametrization $G \to P_G$ of the model. A nice property of differentiability in quadratic mean is that it is preserved under 'censoring' mechanisms of this type. If $\theta \to G_\theta$ is a (regular parametric) submodel of $\mathcal{G}$, then the induced submodel $\theta \to P_{G_\theta}$ satisfies the first two requirements of a regular parametric model of $\{P_G : G \in \mathcal{G}\}$. The score function h (at $\theta = 0$) for the induced model $\theta \to P_{G_\theta}$ is related to the score function b (at $\theta = 0$) of the model $\theta \to G_\theta$ by

$$h(x) = E_{G_0}(b(Y)|X = x).$$

If the scores b and h are considered the carriers of information about θ in $Y \sim G_\theta$ and $X \sim P_{G_\theta}$, respectively, the intuitive meaning of the conditional expectation operator should be clear.

Given a tangent set $\dot{\mathcal{G}}$ for the model $\mathcal{G}$, it follows that the set $\{A_{G_0} b : b \in \dot{\mathcal{G}}\}$, where A_G is the conditional expectation operator $b \to E_G(b(Y)|X = x)$, is contained in the tangent space $\dot{\mathcal{P}}$. The set of distributions of Y may itself have one of many possible structures, parametric, nonparametric or semiparametric, which leads to a variety of possible tangent spaces $\dot{\mathcal{P}}$.

EXAMPLE 14 (COX MODEL). The density of an observation in the Cox model takes the form

$$(t, z) \to \exp(-e^{\theta^T z}\Lambda(t))\lambda(t)\, e^{\theta^T z} g(z).$$

Differentiable submodels varying θ, λ and g yield score functions

$$z - z\, e^{\theta^T z}\Lambda(t), \quad a(t) - e^{\theta^T z}\int_0^t a \, d\Lambda, \quad b(z),$$

where a and b are the derivatives with respect to λ and g. The tangent space contains the linear span of these functions. Note that the scores for Λ can be found as an 'operator' working on functions a.

EXAMPLE 15 (TRANSFORMATION REGRESSION MODEL). If the transformation τ is increasing and e has density ϕ, then the density of the observation can be written $\phi(\tau(y) - \theta^T z)\tau'(y)g(z)$. Scores for θ and τ take the form

$$-z\frac{\phi'}{\phi}(\tau(y) - \theta^T z), \quad \frac{\phi'}{\phi}(\tau(y) - \theta^T z)a(y) + \frac{a'}{\tau'}(y),$$

where a is the derivative for τ. If the distribution of e or Z is (partly) unknown, there are additional scores corresponding to their distributions. Again scores take the form of an operator acting on a set of functions.

dP_0 at $\theta = 0$. Both submodels
nonnegative function ψ with
ed and can be used with any g.

mon distribution of the obser-
$Y = m(Y)$ of an (unobservable)
ie distribution of Y is known to
P_G of the model. A nice property
nder 'censoring' mechanisms of
en the induced submodel $\theta \to P_{G_\theta}$
nodel of $\{P_G: G \in \mathcal{G}\}$. The score
to the score function b (at $\theta = 0$)

about θ in $Y \sim G_\theta$ and $X \sim P_{G_\theta}$,
pectation operator should be

the set $\{A_{G_0} b: b \in \dot{\mathcal{G}}\}$, where A_G
), is contained in the tangent
ne of many possible structures,
to a variety of possible tangent

the Cox model takes the form

ctions

tangent space contains the linear
und as an 'operator' working on

ansformation τ is increasing and
written $\phi(\tau(y) - \theta^T z)\tau'(y)g(z)$.

),

Z is (partly) unknown, there are
ain scores take the form of an

4 Lower bounds

For simplicity we restrict our discussion of lower bounds to the estimation of real-valued functions on the model. Chapters 3 and 4 of BKRW give discussions of the vector-valued and infinite dimensional cases.

For a regular one-dimensional parametric model $\mathcal{P} = \{P_\theta: \theta \in \Theta\}$ the Cramér-Rao lower bound for the variance of unbiased estimators of a differentiable function $\psi(\theta)$ equals

$$\frac{\dot{\psi}(\theta)^2}{I_\theta}.$$

The information for estimating $\psi(\theta)$ could be defined as the inverse of this quantity.

Consider a real-valued function $v(P)$ of the underlying distribution in a general model $\mathcal{P}$. Clearly the supremum of the expression in the preceding display over all regular parametric submodels for which $\psi(\theta) = v(P_\theta)$ is differentiable gives a lower bound for the variance of unbiased estimators. A submodel for which the supremum is taken is called a *least favourable* submodel. In this model (which does not necessarily exist) estimation of $v(P)$ is hardest. The supremum can be given an attractive form if the function $v(P)$ is *pathwise differentiable*. In BKRW this is defined to be the case if there exists an element $\dot{v}_{P_0}$ in $L_2(P_0)$ such that for every one-dimensional, regular parametric submodel passing through P_0,

$$\frac{\partial}{\partial \theta}\Big|_{\theta=0} v(P_\theta) = \langle \dot{v}_{P_0}, h \rangle_{P_0},$$

where h is the score function of the submodel (at $\theta = 0$). This requires both that the function $\psi(\theta) = v(P_\theta)$ is differentiable (at $\theta = 0$) and that the derivative can be written as an inner product of the score and some fixed *gradient* or *influence function*. The latter is not restrictive: in Van der Vaart (1991) it is shown to be necessary for the supremum that we wish to calculate to be finite. For a pathwise differentiable function the lower bound for the asymptotic variance takes the form

$$\sup \frac{\langle \dot{v}_{P_0}, h \rangle_{P_0}^2}{\langle h, h \rangle_{P_0}^2},$$

where the supremum is taken over all (regular) parametric submodels. Equivalently, since the expression depends on the scores h only, the supremum may be taken over the tangent space.

Theorem 1. *The supremum in the preceding display is equal to*

$$I^{-1}(P_0|v, \mathcal{P}) = \|\Pi_{P_0}(\dot{v}_{P_0}|\dot{\mathcal{P}})\|_{P_0}^2.$$

The notation is taken from BKRW. The left side is a formal notation for the inverse of the information for estimating v given the model $\mathcal{P}$ evaluated at true underlying distribution P_0. The right side is the square expectation of the orthogonal projection of $\dot{v}_{P_0}$ on the tangent space. In general the notation $\Pi_{P_0}(h|L)$ in BKRW is the orthogonal projection of h in $L_2(P_0)$ onto a given closed subspace L: it is the element of L that minimizes $\|l - h\|_{P_0}^2$ when l varies over L.

Even though the Cramér–Rao bound was a good starting point for motivating the preceding definition of information, its restriction to unbiased estimators is not satisfying, in particular for semiparametric models. It is better to give the lower bound an asymptotic interpretation. Among asymptotic lower bound statements the local asymptotic minimax theorem and the convolution theorem are the most popular. In BKRW only the convolution theorem is developed. Here we choose to include the minimax theorem.

THEOREM 2. *Assume that the tangent set $\dot{\mathcal{P}}_0$ contains a convex cone with closed, linear span equal to the tangent space $\dot{\mathcal{P}}$. For every element h in this cone let $\{P_{0,h}: |\theta| < 1\}$ be a (regular) parametric submodel with score h. Then for any estimator sequence $\{T_n\}$ and function $l: [0, \infty) \to [0, \infty)$*

$$\sup_I \liminf_{n\to\infty} \sup_{h\in I} E_{P_{1/\sqrt{n},h}} l(\sqrt{n}|T_n - v(P_{1/\sqrt{n},h})|) < \int l(|x|)\, dN_{0,I^{-1}(P_0|v,\mathcal{P})}(x),$$

where the first supremum is taken over all finite subsets I of the tangent set.

At first sight the local asymptotic minimax theorem looks somewhat complicated. The problem is that expressions of the type $E_P l(\sqrt{n}(T_n - v(P)))$ can converge to zero for a given P. Only the maximum risk over certain neighbourhoods can be bounded below in a nontrivial manner. This is why the two suprema on the left side appear. Since the variational distance $\|P_{1/\sqrt{n},h} - P_0\|$ converges to zero as $n \to \infty$, for every h, the left side of the theorem is for every $\theta > 0$ smaller than

$$\liminf_{n\to\infty} \sup_{\|P - P_0\| < \delta} E_P l(\sqrt{n}|T_n - v(P)|).$$

The theorem implies that this asymptotic local maximum risk cannot fall below the corresponding risk of a given normal distribution. A popular interpretation is that the best possible limit distribution of the sequence $\sqrt{n}(T_n - v(P_0))$ is normal with mean zero and variance $I^{-1}(P_0|v, \mathcal{P})$.

EXAMPLE 16 (PARAMETRIC MODEL). Consider estimation of a differentiable, real function $\psi(\theta)$ in a regular parametric model. For a given vector h the one-dimensional submodel $\{P_{\theta_0 + th}: |t| < \epsilon\}$ (well-defined for sufficiently small $\epsilon > 0$) possesses score function $h^T \dot{l}_{\theta_0}$ at $t = 0$. If $\dot{\psi}_\theta$ denotes the gradient of ψ at θ (the row vector of partial derivatives), then

$$\frac{\partial}{\partial t}_{|t=0} \psi(\theta_0 + th) = \dot{\psi}_{\theta_0} h = E_{\theta_0}(\dot{\psi}_{\theta_0} I_{\theta_0}^{-1} \dot{l}_{\theta_0}(X))(h^T \dot{l}_{\theta_0}(X)).$$

This shows that the function $v(P_\theta) = \psi(\theta)$ is pathwise differentiable with influence function

$$\dot{v}_{P_{\theta_0}} = \dot{\psi}_{\theta_0} I_{\theta_0}^{-1} \dot{l}_{\theta_0}.$$

Since this is already contained in the tangent space, the projection operation is unnecessary. This leads to the lower bound $\dot{\psi}_{\theta_0} I_{\theta_0}^{-1} \dot{\psi}_{\theta_0}^T$ for the asymptotic variance.

; point for motivating the preceding
ıtors is not satisfying, in particular
ound an asymptotic interpretation.
mptotic minimax theorem and the
only the convolution theorem is
em.

x cone with closed, linear span equal
ıe let $\{P_{0,h}: |\theta| < 1\}$ *be a (regular)*
ıator sequence $\{T_n\}$ *and function*

$|x|)\, dN_{0,I^{-1}(P_0|\mathcal{V},\mathcal{P})}(x)$,

of the tangent set.

looks somewhat complicated. The
))) can converge to zero for a given
ın be bounded below in a nontrivial
pear. Since the variational distance
left side of the theorem is for every

imum risk cannot fall below the
pular interpretation is that the best
$_0$)) is normal with mean zero and

f a differentiable, real function $\psi(\theta)$
h the one-dimensional submodel
ossesses score function $h^T \dot{l}_{\theta_0}$ at $t = 0$.
partial derivatives), then

$_{\theta_0}(X)$).

ıwise differentiable with influence

projection operation is unnecessary.
totic variance.

EXAMPLE 17 (LINEAR FUNCTION). Consider estimation of the function $v(P) = \int f \, dP$ for a fixed measurable map f. The choice $f = 1_{(-\infty,c]}$ yields the distribution function of P at the point c. For a given parametric submodel we can argue that

$$v(P_\theta) - v(P_0) = \int f(p_\theta^{1/2} - p_{p_0}^{1/2})(p_\theta^{1/2} + p_0^{1/2})\, d\mu$$

$$\approx \theta \int f(\tfrac{1}{2} \dot{l}_0 p_0^{1/2}) 2 p_0^{1/2}\, d\mu + o(\theta) = \theta \langle f, \dot{l}_0 \rangle_{P_0} + o(\theta).$$

This suggests that the function f itself can be taken as the derivative $\dot{v}_{P_0}$. In BKRW this is shown to be true if there exists a constant M such that $\int f^2 \, dP < M$ for every P in the model.

Actually for many functions f the approximation in the preceding display is not valid for *all* regular submodels without some restrictive condition. Then the functional is not pathwise differentiable in the sense of BKRW. A similar problem arises in other semiparametric models, because the regularity of a submodel is in general not related to the differentiability of the functional of interest. In BKRW this problem is overcome by arguing heuristically when discussing concrete examples. See the discussion on page 71.

If the model is nonparametric, then the projection of $\dot{v}_{P_0} = f$ on the tangent space equals $f - v(P_0)$. Then the bound $\|\dot{v}_{P_0}\|_{P_0}^2$ for the asymptotic variance is the variance of $f(X)$ and the empirical estimator $n^{-1} \sum_{i=1}^n f(X_i)$ is asymptotically efficient. If the tangent space does not contain the function $f - v(P_0)$, then the bound for the asymptotic variance is smaller. Efficient estimators are known in many examples, but not in general.

Even though the information bound for pathwise differentiable functions given by the preceding theorem and obtained by KOSHEVNIK and LEVIT (1976), LEVIT (1978) and PFANZAGL and WEFELMEYER (1983), covers all possible situations, it is worthwhile to memorize the bound for special parametrizations.

Consider first the problem of estimating the parameter θ in a semiparametric model of the form $\mathcal{P} = \{P_{\theta,G}: \theta \in \Theta, G \in \mathcal{G}\}$. Suppose that the submodel $\{P_{\theta,G_0}: \theta \in \Theta\}$ in which G_0 is fixed, is a one-dimensional regular submodel with score function $\dot{l}_{\theta_0}$ at θ_0 and let $\dot{\mathcal{P}}_{\mathcal{G}}$ be the tangent space for the submodel $\{P_{\theta_0,G}: G \in \mathcal{G}\}$ in which θ_0 is fixed. Then

$$\dot{l}_{\theta_0}^* = \dot{l}_{\theta_0} - \Pi_{P_{\theta_0,G_0}}(\dot{l}_{\theta_0} | \dot{\mathcal{P}}_{\mathcal{G}}),$$

is called the *efficient score function* for θ.

THEOREM 3. *The information lower bound* $I^{-1}(P_{\theta_0,G_0}|\theta,\mathcal{P})$ *for estimating* θ *in a semiparametric model equals* $\|\dot{l}_{\theta_0}^*\|_{P_0}^{-2}$: *the inverse of the square expectation of the efficient score function.*

This theorem, proved in BEGUN, HALL, HUANG and WELLNER (1983), has an intuitive interpretation. The score function $\dot{l}_{\theta_0}$ gives the information for θ when G is known; to find the information for the semi-parametric model we must subtract the part that is also explainable by a score for G.

Next consider the problem of estimating a functional of G. Consider first the situation that G is itself a probability measure on some measurable space and that there is no Euclidean parameter. Thus the model is $\{P_G: G \in \mathcal{G}\}$ and we are interested in a function of the type

$v(P_G) = \chi(G)$. Assume that a smooth parametric submodel $\theta \to G_\theta$ induces a smooth parametric submodel $\theta \to P_{G_\theta}$, where the scores are related by an operator $A_0: L_2(G_0) \to L_2(P_{G_0})$:

$$h = A_0 b.$$

Since A_0 turns scores for the model $\mathcal{G}$ into scores for $\mathcal{P}$ it is called a *score operator*. In Example 3.13 the score operator in information loss models was seen to be a conditional expectation operator. Assume also that the function $G \to \chi(G)$ is pathwise differentiable with gradient $\dot{\chi}_{G_0}$. Then by definition $v(P_G) = \chi(G)$ is pathwise differentiable if and only if

$$\frac{\partial}{\partial \theta}_{|\theta=0} v(P_{G_\theta}) = \frac{\partial}{\partial \theta}_{|\theta=0} \chi(G_\theta) = \langle \dot{\chi}_{G_0}, b \rangle_{G_0}$$

can be written as an inner product $\langle \dot{v}_{P_{G_0}}, A_0 b \rangle_{P_{G_0}}$, for every submodel and score b (at G_0) in the model $\mathcal{G}$. The resulting equation can be rewritten in terms of the adjoint operator $A_0^*: L_2(P_0) \to L_2(G_0)$, which by definition satisfies $\langle h, A_0 b \rangle_{P_{G_0}} = \langle A_0^* h, b \rangle_{G_0}$ for every h and b. We obtain

$$A_0^* \dot{v}_{P_{G_0}} = \dot{\chi}_{G_0}.$$

Under the two assumptions we have made, the function $v(P_G) = \chi(G)$ is pathwise differentiable if and only if this equation can be solved for $\dot{v}_{P_{G_0}}$. Equivalently if and only if $\dot{\chi}_{G_0}$ is contained in the range of the adjoint A_0^*.

If it is contained in the smaller range of $A_0^* A_0$, then the equation can be solved, of course, and the solution can be written in the attractive form

$$\dot{v}_{P_{G_0}} = A_0 (A_0^* A_0)^- \dot{\chi}_{G_0}.$$

The *information operator* $A_0^* A_0$ performs a similar role as the matrix $X^T X$ in the least squares solution of a linear regression model, which is obtained by projecting the dependent vector onto the regression space.

In a semiparametric model $\{P_{\theta,G}: \theta \in \Theta, G \in \mathcal{G}\}$ the information calculation for a function of the type $v(P_G) = \chi(G)$ is slightly more complicated, because the tangent space will also contain the score l_{θ_0} for the Euclidean parameter.

THEOREM 4. *In a semiparametric model the gradient $\dot{v}_{P_{\theta_0,G_0}}$ of the function $v(P_{\theta,G}) = \chi(G)$ satisfies*

$$\langle \dot{v}_{P_{\theta_0,G_0}}, l_{\theta_0} \rangle_{P_{\theta_0,G_0}} = 0; \quad A_0^* \dot{v}_{P_{\theta_0,G_0}} = \dot{\chi}_{G_0}.$$

Here $\dot{\chi}_{G_0}$ is the gradient of $G \to \chi(G)$ and l_{θ_0} the score function for θ.

If $\dot{\chi}_{G_0}$ is contained in the range of $A_0^* A_0$, then the solution of the equations in the display is

$$-\langle l_{\theta_0}^*, l_{\theta_0}^* \rangle_{P_{\theta_0,G_0}}^{-1} \langle A_0 (A_0^* A_0)^- \dot{\chi}_{G_0}, l_{\theta_0} \rangle_{P_{\theta_0,G_0}} l_{\theta_0}^* + A_0 (A_0^* A_0)^- \dot{\chi}_{G_0}.$$

This formula was first found in BEGUN, HALL, HUANG and WELLNER (1983).

The preceding calculations can be generalized and extended and lead to a 'calculus of score functions', given in Chapters 5 and 6 of BKRW.

$\vartheta \to G_\vartheta$ induces a smooth paramet-
·perator $A_0: L_2(G_0) \to L_2(P_{G_0})$:

alled a *score operator*. In Example
n to be a conditional expectation
se differentiable with gradient $\dot{\chi}_{G_0}$.
if and only if

y submodel and score b (at G_0) in
n terms of the adjoint operator
$_{G_0} = \langle A_0^* h, b \rangle_{G_0}$ for every h and b.

$;) = \chi(G)$ is pathwise differentiable
ntly if and only if $\dot{\chi}_{G_0}$ is contained

equation can be solved, of course,

he matrix $X^T X$ in the least squares
y projecting the dependent vector

rmation calculation for a function
:ause the tangent space will also

$_{G_0}$ *of the function* $\nu(P_{\theta,G}) = \chi(G)$

tion for θ.

ion of the equations in the display

$A_0^* A_0)^- \dot{\chi}_{G_0}$.

ıd WELLNER (1983).
ded and lead to a 'calculus of score

5 Construction of estimators

The approach towards construction of estimators in Chapter 7 of BKRW is characterized on page 380: 'Find a tractable procedure using whatever heuristic principles are appropriate rather than sticking to an "optimal" method of estimation whose optimality can only be guaranteed under conditions which are both difficult to check and often do not apply'.

Given that maximum likelihood is a unifying idea in the theory of estimation in parametric models, it is not surprising that most methods of constructing estimators in semiparametric models are modifications of the method of maximum likelihood. Given a model $\mathscr{P}$ the maximum likelihood estimator of a function $\nu(P)$ is the value $\nu(\hat{P})$ for $\hat{P}$ maximizing the log likelihood

$$p \to \sum_{i=1}^{n} \log p(X_i) = n \int \log p \, d\mathbb{P}_n; \quad \mathbb{P}_n = \sum_{i=1}^{n} \delta_{X_i}.$$

Here p is a density of P with respect to some fixed measure and the maximization is carried out over all $\mathscr{P}$. In a large number of semiparametric models a maximum likelihood estimator exists and is asymptotically efficient. In equally many models the maximum likelihood method breaks down: there may not be natural versions of the densities p; the likelihood may be infinite; there may be many points of maximum; there may even be a unique maximum likelihood estimator which is asymptotically inconsistent.

In cases where the maximum likelihood method fails, it can often be repaired by a modification. In BKRW four types of modifications are discussed: sieves, penalization, regularization and the one-step method.

The method of *sieves* restricts maximization of the likelihood to sets $\mathscr{P}_n$ which 'converge' to the model $\mathscr{P}$ as $n \to \infty$. The idea is that for finite n the likelihood varies too much on the whole model, but may attain unique maxima on smaller sets. Popular sieves in function spaces are sets of spline functions with a fixed number of nodes, with the distance between the nodes tending to zero at a suitable rate as $n \to \infty$. In a semiparametric model sieves $\mathscr{G}_n$ in the nuisance parameter space lead naturally to sieves for the model, though it has been found convenient to use sieves that are not submodels as well.

Penalized likelihood estimators are obtained by maximization over the whole model, but the log likelihood is replaced by

$$p \to \sum_{i=1}^{n} \log p(X_i) - \lambda_n J(p),$$

for a given 'penalty function' $J(p)$. Densities for which the penalty is high are less likely to yield the maximum value. A popular penalty is the integral of the square of the derivative of a density and penalizes roughness of p. The 'tuning' constants λ_n determine the influence of the penalty term and should converge to zero as $n \to \infty$.

The method of *regularization* is based on an initial estimator $\tilde{\mathbb{P}}_n$ of P and choses as estimator the distribution that maximizes

$$p \to \int \log p \, d\tilde{\mathbb{P}}_n.$$

Choosing $\tilde{\mathbb{P}}_n$ equal to the empirical distribution $\mathbb{P}_n$ of the observations leads back to maximum likelihood, but there is a variety of other choices, such as the smoothed empirical distribution with density $(n\sigma_n)^{-1}\sum_{i=1}^n k((X_i - x)/\sigma_n)$ and other methods to 'smooth out' the observations.

Each of the three modifications of maximum likelihood depends on a tuning rate. This is the rate of decrease of the discrepancy between $\mathscr{P}_n$ and $\mathscr{P}$ in sieve estimation, and the rate at which the constants λ_n and σ_n decrease to zero in the cases of penalization and regularization. The choice of tuning constants is an important problem, both theoretically and practically. Cross validation schemes to choose optimal values based on the data have not been developed.

As usual the asymptotic analysis of these estimators consists of separate proofs of consistency and asymptotic normality. We restrict ourselves to asymptotic normality. In the terminology of BKRW the estimators discussed previously are *generalized minimum contrast estimators*: they maximize a criterion function. The asymptotic normality proof is carried through by characterizing the estimators as *generalized M-estimators*, which are defined in BKRW as estimators that solve a system of estimating equations. This system is found by differentiating the criterion function along one-dimensional submodels. If the elements of the tangent set $\dot{\mathscr{P}}_0(\hat{P}, \mathscr{P})$ (at $\hat{P}$ for the model $\mathscr{P}$) can also be obtained as pointwise limits, then maximum likelihood estimators satisfy

$$\int h \, d\tilde{\mathbb{P}}_n = 0, \quad \text{every } h \in \dot{\mathscr{P}}_0(\hat{P}, \mathscr{P}).$$

Sieve type estimators satisfy this equation for $h \in \dot{\mathscr{P}}_0(\hat{P}, \mathscr{P}_n)$ and regularized estimators satisfy the equation with $\mathbb{P}_n$ replaced by $\tilde{\mathbb{P}}_n$. The equation obtained for penalized likelihood estimators is slightly more complicated, because it includes the derivative of the penalty.

Typically the elements of (a subset of) the tangent set $\dot{\mathscr{P}}_0(P, \mathscr{P})$ can be written as $A_P b$ for b running through some index set. For instance, a score operator A_P could be acting on a set of functions b. Then all types of estimators considered previously satisfy equations of the type $\mathbb{W}_n(\hat{P}_n)b = 0$. For instance, for regularized estimators take $\mathbb{W}_n(P)b = \int A_P b \, d\tilde{\mathbb{P}}_n$. The set of equations is linearized and inverted to obtain asymptotic normality. Appropriate technical assumptions are

$$\sqrt{n}(\mathbb{W}_n - W_n)(\hat{P}_n) - \sqrt{n}(\mathbb{W}_n - W_n)(P_0) \xrightarrow{P} 0$$

$$\sqrt{n}(W_n(\hat{P}_n) - W_n(P_0)) - \dot{W}_n(P_0)\sqrt{n}(v(\hat{P}_n) - v(P_0)) \xrightarrow{P} 0.$$

The functions W_n are centering functions that should ensure that the processes $\sqrt{n}(\mathbb{W}_n - W_n)$ converge in distribution to a normal distribution. Typically the true value P_0 is a zero of W_n just as $\hat{P}_n$ is a (near) zero of $\mathbb{W}_n$. The first condition is a technical regularity condition. The second requires some type of differentiation of the centering functions W_n at the true value P_0 with derivative denoted by $\dot{W}_n(P_0)$. In most situations the $\mathbb{W}_n$ and W_n are stochastic processes indexed by a set of score functions. Their convergence should take place in some functional sense.

bservations leads back to maximum
he smoothed empirical distribution
s to 'smooth out' the observations.
i depends on a tuning rate. This is
id $\mathscr{P}$ in sieve estimation, and the
in the cases of penalization and
ant problem, both theoretically and
ues based on the data have not been

rs consists of separate proofs of
ves to asymptotic normality. In the
ly are *generalized minimum contrast*
mptotic normality proof is carried
M-estimators, which are defined in
equations. This system is found by
al submodels. If the elements of the
obtained as pointwise limits, then

$_n$) and regularized estimators satisfy
d for penalized likelihood estimators
ivative of the penalty.
$\dot{\mathscr{P}}_0(P, \mathscr{P})$ can be written as $A_P b$ for
operator A_P could be acting on a set
eviously satisfy equations of the type
ke $\mathbb{W}_n(P)b = \int A_P b \, d\tilde{\mathbb{P}}_n$. The set of
ic normality. Appropriate technical

$P_0)) \xrightarrow{P} 0$.

ure that the processes $\sqrt{n}(\mathbb{W}_n - W_n)$
ally the true value P_0 is a zero of W_n
a technical regularity condition. The
ering functions W_n at the true value
ions the $\mathbb{W}_n$ and W_n are stochastic
ivergence should take place in some

THEOREM 5. *If $\mathbb{W}_n(\hat{P}_n) = W_n(P_0) + o_P(n^{-1/2})$, then $\dot{W}_n(P_0)\sqrt{n}(v(\hat{P}_n) - v(P_0))$ is asymptotically equivalent to $-\sqrt{n}(\mathbb{W}_n - W_n)(P_0) + o_P(1)$.*

This theorem is trivial: it follows immediately from the two conditions. As indicated in the quote at the beginning of this section the point of view in BKRW is that it is impossible to obtain workable results beyond the very abstract level of a theorem of this type.

If the centered process $\sqrt{n}(\mathbb{W}_n - \mathbb{W}_n)(P_0)$ is asymptotically normal and the inverse operators $\dot{W}_n^{-1}$ exist and converge in a suitable manner to a (continuous) limit, then the sequence $\sqrt{n}(v(\hat{P}_n) - v(P_0))$ converges to a normal distribution as well.

A heuristic discussion in Section 7.8 of BKRW indicates that maximum likelihood estimators should be asymptotically efficient in every case where the approach outlined above works. Furthermore, the three modifications ought to yield asymptotically efficient estimators as well, provided the effect of modifying the likelihood wears off as $n \to \infty$. In BKRW this is illustrated for several examples, but no rigorous results are formulated. Results for maximum likelihood estimators can be found in GILL (1989), GILL and VAN DER VAART (1993) and VAN DER VAART (1984).

Finally we discuss the *one-step* modification of maximum likelihood. Consider estimation of the parameter θ in a semiparametric model $\{P_{\theta,G}: \theta \in \Theta, G \in \mathscr{G}\}$. Suppose that the model $\{P_{\theta,G_0}: \theta \in \Theta\}$ with G_0 fixed is a regular parametric submodel and assume that the efficient score functions l^*_{θ,G_0} for θ satisfy the regularity condition

$$\int |l^*_{\theta,G_0} p^{1/2}_{\theta,G_0} - l^*_{\theta_0,G_0} p^{1/2}_{\theta_0,G_0}|^2 \, d\mu \to 0, \quad \theta \to \theta_0.$$

Then the one-step method yields asymptotically efficient estimators for θ provided suitable preliminary estimators for θ and $l^*_{\theta,G}$ are available.

THEOREM 6. *Suppose that there exist (function-valued) estimators $\hat{l}^*_\theta$ such that for every $\theta_n = \theta_0 + O(n^{-1/2})$*

$$\sqrt{n} \int \hat{l}^*_{\theta_n} \, dP_{\theta_n,G_0} \xrightarrow{P} 0; \quad \int |\hat{l}^*_{\theta_n} - l^*_{\theta_n,G_0}|^2 \, dP_{\theta_n,G_0} \xrightarrow{P} 0.$$

Then there exists an efficient estimator for θ provided there exists an asymptotically normal estimator. These conditions are necessary.

The proof of the theorem is based on an explicit construction. The idea is to solve an estimator $\hat{\theta}_n$ from the estimating equation

$$\sum \hat{l}^*_\theta(X_i) = 0.$$

This is similar to solving the maximum likelihood estimator from the likelihood equation. The present case is more complicated in that the estimating equation involves an *estimator* of the *efficient* score function. In case the estimating equation is ill-behaved one may use instead of a zero the estimator

$$\hat{\theta}_n = \tilde{\theta}_n - (\sum \hat{l}_{\tilde{\theta}_n}(X_i))^{-1} \sum \hat{l}_{\tilde{\theta}_n}(X_i),$$

for given preliminary asymptotically normal estimators $\tilde{\theta}_n$. This estimator is called a one-step estimator, because it is the next approximation to a zero of the estimating equation $\Sigma \hat{l}^*_{\theta}(X_i) = 0$, when running the Newton–Raphson algorithm with starting value $\tilde{\theta}_n$. Under regularity conditions the one-step method improves a given asymptotically normal (or at least $\sqrt{n}$-consistent) estimator into an efficient estimator. The preliminary estimator is usually constructed by ad-hoc methods. A number of theoretical tricks can be used to reduce the regularity conditions to those of the theorem, which is due to KLAASSEN (1987).

The theorem reduces the problem of efficient estimation of θ to estimation of the efficient score function. The two conditions that $\hat{l}^*_{\theta}$ must satisfy can be characterized as a bias and a consistency condition. The consistency is usually easy to arrange. The bias condition may be harder to achieve, because the 'bias' $\int \hat{l}^*_{\theta_n} \, dP_{\theta_n, G_0}$ is required to disappear at the rate $n^{-1/2}$. This is trivially true for an estimator of the type $\hat{l}^*_{\theta,G}$ if

$$\int l^*_{\theta,G} \, dP_{\theta,G_0} = 0, \quad \text{every } \theta, G, G_0.$$

This full unbiasedness occurs in semiparametric models that are convex linear in the nuisance parameter. Without a special structure of the model the unbiasedness condition requires that the nuisance parameter is estimable at some rate.

In semiparametric models that are convex linear in the nuisance parameter functions of the type

$$\frac{p_{\theta,G}}{p_{\theta,G_0}} - 1 = \frac{\partial}{\partial t}_{|t=0} \log p_{\theta, tG + (1-t)G_0}$$

are score functions (for possibly one-sided submodels). The orthogonality of l^*_{θ,G_0} to the tangent space readily yields the unbiasedness of the efficient score function.

6 Concluding remarks

The work BKRW is strong in information calculations. It is less strong in the construction of estimators. While this largely reflects the current state of affairs in semiparametrics, it appears that the authors are too pessimistic about the possibility of a rigorous, general theory of the asymptotic behaviour of various types of estimators. The modern theory of empirical processes (see POLLARD, 1992) is promising, where the theorems would be formulated in terms of Donsker classes and entropy numbers. Some may regret that this will lead to further use of functional analytic concepts (which are already used in the lower bound theory), but these seem indispensable given that models are described in terms of abstract parameters rather than vectors of numbers.

Some subjects omitted from BKRW were already noted. Since every of the four authors has made important contributions to the field of semiparametrics, it is understandable that the choice of subjects and presentation is somewhat biased towards the authors' own work. One striking example of this is the omission of a reference to the paper SEVERINI and WONG (1992). This paper puts forward the nice idea that least favorable submodels may be constructed by minimizing a (smoothed) Kullback-Leibler divergence. In SEVERINI and WONG (1992) this is

This estimator is called a one-step
zero of the estimating equation
ım with starting value $\tilde{\theta}_n$. Under
asymptotically normal (or at least
preliminary estimator is usually
tricks can be used to reduce the
e to KLAASSEN (1987).
of θ to estimation of the efficient
ı be characterized as a bias and a
·range. The bias condition may be
ɔd to disappear at the rate $n^{-1/2}$.

t are convex linear in the nuisance
biasedness condition requires that

uisance parameter functions of the

orthogonality of $\tilde{l}^*_{\theta,G_0}$ to the tangent
function.

t is less strong in the construction
e of affairs in semiparametrics, it
ibility of a rigorous, general theory
s. The modern theory of empirical
rems would be formulated in terms
ɛt that this will lead to further use
the lower bound theory), but these
of abstract parameters rather than

Since every of the four authors has
trics, it is understandable that the
wards the authors' own work. One
ɛ paper SEVERINI and WONG (1992).
submodels may be constructed by
SEVERINI and WONG (1992) this is

presented as an alternative to the one-step method as discussed in BKRW. Even though in SEVERINI and WONG (1992) this approach is carried through only under somewhat forbidding regularity conditions, the paper deserves to be mentioned.

References

BEGUN, J. M., W. J. HALL, W. HUANG and J. A. WELLNER (1983), Information and asymptotic efficiency in parametric-nonparametric models, *Annals of Statistics 11*, 432–452.

BICKEL, P. J., C. A. J. KLAASSEN, Y. RITOV and J. A. WELLNER (1993), *Efficient and adaptive estimation for semiparametric models*, Johns Hopkins University Press.

CHERNOFF, H. (1956), Large sample theory: parametric case, *Annals of Mathematical Statistics 27*, 1–21.

CRAMÉR, H. (1946), *Mathematical methods of statistics*, Princeton University Press.

FISHER, R. A. (1922), Theory of statistical estimation, *Proceedings Cambridge Philosophical Society 22*, 700–725.

GILL, R. D. (1989), Non- and semi-parametric maximum likelihood estimators and the von Mises method, Part I, *Scandinavian Journal of Statistics 16*, 97–128.

GILL, R. D. and A. W. VAN DER VAART (1993), Non- and semi-parametric maximum likelihood estimators and the von Mises method, Part II, *Scandinavian Journal of Statistics 20*, 171–288.

HAJEK, J. (1970), A characterization of limiting distributions of regular estimates, *Zeitschrift Wahrscheinlichkeitstheorie under Verwante Gebiete 14*, 323–330.

HAJEK, J. (1972), Local asymptotic minimal and admissibility in estimation, *Proceedings Sixth Berkeley Symposium on Mathematical Statistics and Probability 1*, 175–194.

KLAASSEN, C. A. J. (1987), Consistent estimation of the influence function of locally asymptotically linear estimates, *Annals of Statistics 15*, 1548–1562.

KOSHEVNIK, YU. A. and B. YA. LEVIT (1976), On a non-parametric analogue of the information matrix, *Theory Probability and Applications 21*, 738–753.

LE CAM, L. (1956), On the asymptotic theory of estimation and testing hypotheses, *Proceedings Third Berkeley Symposium on Mathematical Statistics and Probability 1*, 129–156.

LE CAM, L. (1960), Locally asymptotically normal families of distributions, *University California Publications in Statistics 3*, 37–98.

LE CAM, L. (1970), On the assumptions used to prove asymptotic normality of maximum likelihood estimators, *Annals of Mathematical Statistics 41*, 802–828.

LE CAM, L. (1972), Limits of experiments, *Sixth Berkeley Symposium on Mathematical Statistics and Probability 1*, 245–261.

LE CAM, L. (1986), *Asymptotic methods in statistical decision theory*, Springer-Verlag.

VAN DER LAAN, M. (1983), *Efficient and inefficient estimation in semiparametric models*, University of Utrecht.

LEVIT, B. YA. (1978). Infinite-dimensional informational lower bounds, *Theory Probability and Applications 23*, 388–394.

MURPHY, S. (1992), Asymptotic theory for the frailty model, (preprint).

PFANZAGL, J. and WEFELMEYER, W. (1982), *Contributions to a general asymptotic theory*, *Lecture Notes in Statistics 13*, Springer Verlag.

POLLARD, D. (1990), *Empirical processes: theory and applications, NSF-CBMS Regional Conference Series in Probability and Statistics 2*, IMS, ASA.

SEVERINI, T. A. and W. H. WONG (1992), Profile likelihood and conditionally parametric models, *Annals of Statistics 20*, 1768–1862.

STEIN, C. (1956), Efficient nonparametric estimation and testing, *Proceedings Third Berkeley Symposium Mathematical Statistics and Probability 1*, 187–195.

VAN DER VAART, A. W. (1988), *Statistical estimation in large parameter spaces*, CWI, Amsterdam.

VAN DER VAART, A. W. (1991), On differentiable functions, *Annals of Statistics 19*, 178–204.

VAN DER VAART, A. W. (1992), Maximum likelihood estimation with partially censored data, (to appear).

VAN DER VAART, A. W. (1994), Efficiency of infinite dimensional M-estimators, *Statistica Neerlandica* 9–30.

AAD VAN DER VAART, Department of Mathematics and Computer Science, Free University, De Boelelaan 1081a, 1081, HV Amsterdam, The Netherlands.