SEMIPARAMETRIC INFERENCE AND MODELS

Peter J. Bickel, C. A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner

September 5, 2005

Abstract

"We review semiparametric models and various methods of inference – efficient and inefficient – for such models."

Outline

1. Introduction.

- **1.1.** What is a Semiparametric Model?
- **1.2.** Selected Examples.
- **1.3.** What is Semiparametric Inference?
- **1.4.** Organizational Schemes and Approaches.

2. Information Bounds: Orthogonality and Projections

- **2.1.** Basic Parametric Theory.
- **2.2.** Bounds for Semiparametric Models.

3. Estimation Methods.

- **3.1.** Efficient Estimates and Achieving Information Bounds.
- **3.2.** Inefficient Estimates.

4. Testing.

- **4.1.** Testing a Parametric Component within a Semiparametric Model.
- **4.2.** Testing a Nonparametric Component within a Semiparametric Model.
- **4.3.** Testing Goodness-of-Fit of a Semiparametric Model.

5. Extensions.

- **5.1.** Beyond i.i.d., LAN, and Differentiable Functions.
- **5.2.** Robust Estimation and Semiparametric Models.

1. Introduction.

Definitions and examples of semiparametric models, information bounds and estimation methods are discussed in sections 1, 2, and 3 respectively. An elaboration of these topics may be found in Bickel, Klaassen, Ritov, and Wellner (1993) [henceforth BKRW (1993)]. Section 4 presents extensions to testing, and section 5 discusses non - i.i.d. cases and robustness issues.

1.1. What is a Semiparametric Model?

Models of random phenomena are basic for statistics. We use the term *model* in most of this article to mean a set $\mathcal{P}$ of probability distributions for the i.i.d. observed data. In the case of classical *parametric* models the collection $\mathcal{P}$ is parametrized by a subset Θ of a finite-dimensional Euclidean space, R^k say, and we write $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. At the other extreme, a *nonparametric* model $\mathcal{P}$ is one consisting of all probability measures on the sample space for the observations (which we often denote by $\mathcal{M}$) or a large subset of $\mathcal{M}$ with only qualitative smoothness or moment constraints: $\mathcal{P} = \mathcal{M}_{qual} \subset \mathcal{M}$. Thus *semiparametric models* are intermediate between parametric and nonparametric models: they are larger than parametric models, but smaller than nonparametric models. More technical definitions of the term "semiparametric model" are possible and sometimes useful; see Section 2.2 for another definition in terms of the *tangent space* of the model.

Frequently semiparametric models have a (smooth) parametrization in terms of a finite - dimensional parameter $\theta \in \Theta \subset R^k$ and an infinite - dimensional parameter $G \in \mathcal{G}$ where $\mathcal{G}$ is some space of functions:

$$\mathcal{P} = \{P_{\theta,G} : \theta \in \Theta,\ G \in \mathcal{G}\}.$$

One way of obtaining natural semiparametric models of this type is via relaxation of one (or more) hypotheses in a classical parametric model. This is illustrated in the following list of examples.

1.2. Selected Examples.

A wide variety of semiparametric models have been suggested for applied problems in astronomy, biostatistics, demography, econometrics, epidemiology, genetics, psychology, and spectroscopy. [Searching in the on-line version of the *Current Index to Statistics* produces 4 citations with the topic "semiparametric" in 1985, and 44 and 34 citations in 1992 and 1993 respectively.] The following list of examples illustrates some of the scope of semiparametric models. As in our book (BKRW, 1993), we use X to denote a typical observation, and random vectors $Y, Z, W, \epsilon, ...$ to describe the structure of X, even though they are not all

observed. The parameters θ, η, and ν will always be finite-dimensional parameters, while F, G, τ, and r will denote infinite-dimensional parameters (unknown functions).

Example 1. (Symmetric Location). Let $X = \theta + \epsilon$ where $\epsilon \sim G \in \mathcal{G}_s$ where $\mathcal{G}_s$ denotes the collection of all distributions on R with a density g with respect to Lebesgue measure λ which is symmetric about 0. Thus the model $\mathcal{P}$ is given by

$$\mathcal{P} = \{P_{\theta,G} : \ \frac{dP_{\theta,G}}{d\lambda}(x) = g(x-\theta),\ \theta \in R,\ G \in \mathcal{G}_s\}.$$

The classical normal location-scale model $\mathcal{P}_0 \subset \mathcal{P}$ is the submodel with $\epsilon \sim N(0, \eta^2)$.

Example 2. (Regression). Suppose we observe $X = (Y, Z) \sim P_{\theta,G}$ where

$$Y = \mu(Z, \theta) + \sigma(Z, \theta)\epsilon,$$

where ϵ and Z are independent, the functions μ and σ are known – up to the finite-dimensional parameter θ, and $\epsilon \sim G \in \mathcal{G}$, the collection of all absolutely continuous distributions on R. With m a $\sigma-$finite measure, the model becomes

$$\mathcal{P} = \{P_{\theta,G} : \frac{dP_{\theta,G}}{d\lambda \times m}(y, z) = g(\frac{y - \mu(z,\theta)}{\sigma(z,\theta)})h(z),\ \theta \in R^m,\ G \in \mathcal{G}\}.$$

The classical (homoscedastic) linear regression model with normal errors $\mathcal{P}_0 \subset \mathcal{P}$ is the submodel with $\epsilon \sim N(0, 1)$, $\theta = (\nu, \eta) \in R^m \times R^+$, $\sigma(Z, \theta) = \eta$, and $\mu(Z, \theta) = \nu^T Z$.

Example 3. (Projection Pursuit Regression with Arbitrary Errors). If we relax the assumption of a parametric regression model in Example 2, replacing $\mu(Z, \theta)$ by $r(\nu^T Z)$ where $r : R \to R$ is some (smooth but unknown) function in a class of functions $\mathcal{R}$, then

$$Y = r(\nu^T Z) + \epsilon,$$

where $\epsilon \sim G$ and $r \in \mathcal{R}$, and so the model becomes:

$$\mathcal{P} = \{P_{\nu,r,G} : \frac{dP_{\nu,r,G}}{d\lambda \times m}(y, z) = g(y - r(\nu^T z))h(z),\ \nu \in R^m,\ r \in \mathcal{R},\ G \in \mathcal{G}\}.$$

The parameter ν is no longer identifiable in this model as presently formulated, but typically $\nu/|\nu|$ is identifiable (under reasonable assumptions on the distribution H of Z).

Example 4. (Partially Linear Logistic regression). Suppose that $X = (Y, Z)$ where conditionally on $Z = (Z_1, Z_2)$ the $0-1$ random variable $Y \sim$Bernoulli$(p(Z))$ with $p(Z) = \exp(r(Z))/(1 + \exp(r(Z)))$ and $r(Z) = \theta^T Z_1 + \tau(Z_2)$ for τ in some class of (smooth) functions.

Example 5. (Errors in Variables Regression). Suppose that $X = (Y, Z)$, with $Z = Z' + \epsilon$ and $Y = \alpha + \beta Z' + \delta$ where (δ, ϵ) is bivariate normal with mean zero and unknown covariance

matrix, and the distribution H of Z' is completely unknown. This models the situation of a linear regression with normal errors in which the covariate Z' is observed with error.

Example 6. (Paired Exponential Mixture). Suppose that conditionally on a positive random variable Z with completely unknown distribution G, the components of $X = (X_1, X_2)$ are independent and exponentially distributed with parameters Z and θZ respectively.

Example 7. (Transformation Regression). Suppose that $X = (Y, Z)$ where $\tau(Y) = \theta^T Z + \epsilon$ for some unknown (smooth) function τ and Z and ϵ independent with distributions known or in specified parametric families.

Example 8. (Cox's Proportional Hazards Model). Suppose that $X = (Y, Z)$ where conditionally on Z the (survival time) Y has cumulative hazard function $\exp(\theta^T Z)\Lambda(y)$. Assuming that Λ has hazard rate $\lambda(y) = \Lambda'(y) = g(y)/(1 - G(y))$ and that the distribution H of Z is known with density h, this model can be expressed as

$$\mathcal{P} = \{P_{\theta,G} : \frac{dP_{\theta,G}}{d\lambda \times m}(y, z) = e^{\theta^T z}(1 - G(y))^{\exp(\theta^T z)-1} g(y)h(z),\ \theta \in R^m,\ g \in \mathcal{G}\}.$$

Example 9. (Additive Hazards Model). Now suppose that $X = (Y, Z)$ where conditionally on Z the (survival time) Y has hazard (rate) function $\lambda(y) + \theta^T Z$.

Example 10. (Known Marginals Model). Suppose that $X = (Y, Z) \in R^2$ has joint distribution function F with known marginal distributions: $F(y, \infty) = G_0(y)$ for all $y \in R$ and $F(\infty, z) = H_0(z)$ for all $z \in R$ where G_0 and H_0 are known univariate distributions. This is the limiting case of a three-sample missing data model with some joint data (with both Y and Z observed) and some marginal data on each axis in the case when the size of the marginal data samples is much larger than the sample size of the joint data.

Example 11. (Copula Models). Suppose that the distribution of $X = (Y, Z)$ is of the form $C_\theta(G(y), H(z))$ where $\{C_\theta : \theta \in \Theta\}$ is a parametric family of distribution functions on the unit square $[0, 1] \times [0, 1]$ with uniform marginal distributions (a family of *copulas*), and one or both of G, H are unknown univariate distribution functions. This yields a bivariate model with parametric dependence structure but arbitrary marginal distributions.

Example 12. (Gamma Frailty Model). Suppose that we observe $X = (Y, Z)$ where the cumulative hazard function of Y conditionally on (Z, W) is of the form $W \exp(\theta^T Z)\Lambda(y)$ and where W has a Gamma(η, η) distribution. This can be viewed as the proportional hazards model with unobserved covariate $\log W$.

These examples are listed here as representatives of various classes of models which have arisen in semiparametric contexts including *group models*, *regression models*, *mixture models*, and *transformation models*. Asymptotic information bounds for most of these examples and many others are treated in BKRW (1993), Chapter 4. While efficient estimators have been constructed for the parametric part θ in many of these models, for other

models the problem of constructing efficient estimators (or even reasonable $\sqrt{n}-$consistent estimators in some cases) still remains as an unsolved problem.

1.3. What is Semiparametric Inference?

We think of *semiparametric inference* as being simply statistical inference (estimation, testing, confidence intervals or sets) in the context of some semiparametric model. The estimators or tests or procedures under consideration may be (asymptotically) efficient or inefficient, but presumably they have at least some minimal validity (such as consistency) for the semiparametric model under consideration in order to be deemed truly *semiparametric*.

Often the *parameter of interest* in a statistical problem involving a semiparametric model is the finite - dimensional parameter $\theta \in \Theta \subset R^k$ involved in defining the model or some other natural finite-dimensional functional, while the *nuisance parameter* is the infinite-dimensional parameter $G \in \mathcal{G}$. However this distinction is not rigid – and one can easily imagine situations in which G or some function thereof would be the "parameter of interest".

One way of defining a *semiparametric estimator* would be to say that it is an estimator which has desirable (asymptotic) properties (such as consistency or $\sqrt{n}$- consistency) as an estimator of a parameter of interest in some semiparametric model. In contrast to this, a *robust estimator* is one which has good efficiency properties under some parametric or semiparametric model, and desirable stability or continuity properties in some *neighborhood* of this model. Procedures, are called *adaptive* if they attain the same efficiency or information bound as if the infinite - dimensional part of a semiparametric model were known, or at least known up to a finite-dimensional parameter. Often both *robust estimators* and *adaptive estimators* are subclasses of a larger class of *semiparametric estimators*. For further discussion of robust estimators in the context of semiparametric models, see Section 5.2.

In the past few years the term *semiparametric inference* has also been applied to inference methods for nonparametric models which have been derived by consideration of some appropriate semiparametric submodel; see e.g. Olkin and Spiegelman (1987), Faraway (1990), or Roeder (1992). This is a somewhat broader interpretation of the term than we have in mind here.

1.4. Organizational Schemes and Approaches.

The examples considered above are in the subclass of models for independent and identically distributed (or *i.i.d*) data: in each case it is implicity assumed in the discussion above that the observations $X_1, \ldots, X_n$ are i.i.d. with distribution $P \in \mathcal{P}$. Of course there are many models for *dependent data* which could exhibit the key feature discussed above – namely of being neither parametric nor nonparametric. In view of the large potential variety of semiparametric models involved, it becomes clear that some thought about organizational schemes might be useful.

Here we will briefly consider approaches based on *families of models*, *estimation principles*, and *information bounds*.

Families of Models: For the moment, we will restrict attention to the small subclass of models for i.i.d. data. Within this class of models, two basic methods of generating semiparametric models stand out: *group models* emerge from a basic core model typically involving a completely unknown distribution G or density g together with a group of transformations on the underlying sample space which is parameterized by a finite-dimensional parameter. Example 1.B is a simple example of a model of this type. On the other hand, *transformation models* can often be viewed as arising via an infinite dimensional group of transformations on the sample space of some parametric family; Examples 1.2.7, 8, 11, and 12 are semiparametric models of this type. In addition to any one of these basic types of semiparametric model, complications can be added in the form of *regression* by the addition of covariates, *biased sampling*, or *missing data*. The latter type of model includes models obtained via *mixing* and *censoring*. In our book BKRW (1993), the sections in Chapter 4 are organized roughly along these lines with special attention to regression models and mixture models.

Information calculations are often similar for models of similar type and estimation methods also have similarities in comparable models. Therefore, there clearly are advantages in studying families of models instead of individual models in isolation.

For an extensive discussion of semiparametric models in econometrics, the reader is referred to Powell (1994).

Estimation Principles or Methods: In classical (parametric) models, many estimation methods have been studied; to name a few: maximum likelihood, least squares, moment estimators, minimum distance estimators. In robustness studies $M-$, $L-$, and $R-$ estimators have been treated. In contrast, there is no general approach yet to the construction of efficient estimators or even of just good estimators in semiparametric models. Variants of MLE, including sieved MLE, penalized MLE, and estimating equations have been suggested and studied in connection with various classes of models. The issues are identifiability of the parameters to be estimated and desirable properties of proposed estimators, including consistency, $\sqrt{n}-$consistency, efficiency, robustness, etc.

Information Bounds: Organizational approaches to semiparametric models based on information bounds offer several advantages: often it is somewhat easier to establish information lower bounds for estimation than to construct good estimators, and in any case in order to construct *efficient estimators*, one must know what efficiency means – i.e. what are the available (information) lower bounds? Since the calculation of information bounds for large families of models are similar, it becomes possible to treat large classes of models quite generally, and this is the approach we have taken in Chapter 4 of BKRW. Because this approach offers considerable insight into the structure of efficient and even $\sqrt{n}-$consistent estimators, we will review information bounds for semiparametric models in the next section – before talking about the construction of estimators.

2. Informations Bounds: Orthogonality and Projections.

In efficient estimation one needs a bound on the performance of any estimator as well as a particular estimator attaining this bound. Then both the bound and the estimator are efficient. If one of these is missing, one hardly has obtained anything. An estimator without a bound might be very inefficient, and one is insecure at least about its performance. Therefore in this section we will discuss bounds on the performance of estimators and we will focus on asymptotic bounds, in particular the so-called Hájek - Le Cam convolution theorem.

2.1. Basic Parametric Theory.

Consider a model for i.i.d. random variables $X_1, \ldots, X_n$ with a finite - dimensional (Euclidean) parameter $\theta \in \Theta \subset R^k$. Fix θ_0 and let θ_n be a local sequence in the open parameter set Θ, i.e. $\theta_n = \theta_0 + O(n^{-1/2})$ as the sample size n tends to infinity. The model is locally asymptotically normal at θ_0 if the log likelihood ratio of the θ_n and θ_0 under θ_0 has a stochastic expansion which is quadratic in $\sqrt{n}(\theta_n - \theta_0)$ with the quadratic term nonrandom and the linear term asymptotically normal. Local Asymptotic Normality (LAN) uniform in θ_0 can be shown to hold for regular parametric models. In these models, the square roots $s(x, \theta)$ of the densities of one observation with respect to a dominating measure μ are continuously Fréchet differentiable in $L_2(\mu)$. In terms of the Fréchet derivative $\dot{s}(x, \theta_0)$, the score function is defined by

$$\dot{l}(x, \theta_0) = 2\frac{\dot{s}(x, \theta_0)}{s(x, \theta_0)} 1_{[s(x,\theta_0)>0]} \tag{2.1}$$

and the Fisher information matrix by

$$I(\theta_0) = E_{\theta_0} \dot{l}\dot{l}^T(X, \theta_0). \tag{2.2}$$

In regular parametric models $I(\theta)$ is assumed to be nonsingular.

A sequence of estimators T_n of θ is called *regular* if the limit behavior of $\{\sqrt{n}(T_n - \theta_n)\}$ under $\{\theta_n\}$ is independent of the local sequence $\{\theta_n\}$. In Hájek (1970) it has been shown that under LAN the limit distribution of such a regular estimator is the *convolution* of a normal distribution with covariance matrix $I^{-1}(\theta_0)$ and another distribution. By Anderson's lemma, it is clear that optimal asymptotic behavior of the estimator sequence $\{T_n\}$ is obtained if this last distribution is degenerate; i.e. if under θ_n

$$\sqrt{n}(T_n - \theta_n) \to_d N(0, I^{-1}(\theta_0)). \tag{2.3}$$

The Hájek - Le Cam convolution theorem also shows that (2.3) happens if and only if $\{T_n\}$ is asymptotically linear in the efficient influence function

$$\widetilde{l}(x, \theta) = I^{-1}(\theta)\dot{l}(x, \theta), \tag{2.4}$$

i.e. if under θ_n

$$\sqrt{n}\left(T_n - \theta_n - \frac{1}{n}\sum_{i=1}^{n}\widetilde{l}(X_i,\theta_n)\right) \to_p 0. \tag{2.5}$$

Since uniform LAN implies for local sequences $\{\theta_n\}$ and $\{\tilde{\theta}_n\}$

$$\sqrt{n}\left\{\widetilde{\theta}_n - \theta_n + \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{l}(X_i,\tilde{\theta}_n) - \widetilde{l}(X_i,\theta_n)\right)\right\} \to_p 0, \tag{2.6}$$

it follows by a contiguity argument that (2.5) yields regularity of $\{T_n\}$.

Nevertheless, a clear drawback of the convolution theorem is that competition between estimators is restricted to just regular estimators. The Local Asymptotic Minimax (LAM) theorem does not rule out any estimator. It states that for any bowl-shaped loss function $l(\cdot)$

$$\lim_{M\to\infty}\liminf_{n\to\infty}\inf_{T_n}\sup_{\sqrt{n}|\theta-\theta_0|\le M} E_\theta l(\sqrt{n}(T_n-\theta)) \ge El(Z), \tag{2.7}$$

where Z is $N(0, I^{-1}(\theta_0))$ and where the infimum is taken over all estimators T_n of θ; cf. Hájek (1972). Because of the "inf sup", it is called a minimax theorem. Of course, the "inf" may be left out in the left hand side of (2.7), and then (2.7) still states that no estimator performs better than an estimator which is efficient in the sense of (2.5). See also section 2 of Fabian and Hannan (1982).

Both the convolution theorem and the local asymptotic minimax theorem fit into the framework of Le Cam's theory of limits of experiments. Here the limit experiment is observing a normal random vector X with mean $I(\theta_0)t$ and known covariance matrix $I(\theta_0)$. The parameter to estimate is t, and corresponds to $\sqrt{n}(\theta_n - \theta_0)$. The best (equivariant) estimator of t is $I^{-1}(\theta_0)X$; indeed, $I^{-1}(\theta_0)X - t$ is $N(0, I^{-1}(\theta_0))$ distributed.

Of course, all these results may be formulated also for estimating $q(\theta)$ instead of θ itself where $q : R^k \to R^m$ is a differentiable function. If $\dot{q}(\theta)$ denotes the total differential matrix, then the efficient influence function becomes

$$\widetilde{l}(x,\theta) = \dot{q}(\theta)I^{-1}(\theta)\dot{l}(x,\theta) \tag{2.8}$$

and (2.5) and (2.7) are still valid with $I^{-1}(\theta_0)$ replaced by the information bound

$$\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0). \tag{2.9}$$

Here the limit experiment still is observing a normal random vector X with mean $I(\theta_0)t$ and known covariance matrix $I(\theta_0)$, but the parameter of interest now is $\dot{q}(\theta_0)t$. The best estimator is $\dot{q}(\theta_0)I^{-1}(\theta_0)X$, which is normal with mean $\dot{q}(\theta_0)t$ and covariance matrix (2.9).

The convolution theorem as well as the local asymptotic minimax theorem may be viewed as asymptotic versions of the classical (Fréchet -) Cramér - Rao inequality, also called the

information inequality. Indeed, in the above setting it states that unbiased estimators T_n of $q(\theta)$ have a covariance matrix which in the ordering of positive semidefinite matrices equals at least n^{-1} times the information bound (2.9).

Consider the estimation problem with $\theta = (\theta_1^T, \theta_2^T)^T$ split into θ_1, the *parameter of interest*, and θ_2, the *nuisance parameter*, or if preferred, the limit problem with $t = (t_1^T, t_2^T)^T$, t_1 the parameter of interest and t_2 the nuisance parameter. The phenomena in this situation are basic to semiparametrics and will have their analogues in semiparametric models, where the nuisance parameter θ_2 is not Euclidean anymore. To study this situation, we introduce the notation

$$I(\theta_0) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}, \qquad I^{-1}(\theta_0) = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}, \tag{2.10}$$

where the matrices are split up in a way compatible with the splitting of θ. The information bound (2.9) for the present situation becomes I^{11}, which by block matrix manipulations can be seen to be equal to $I^{11} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}$. Similarly, it can be seen that, with the notation $\dot{l}(x, \theta_0) = \dot{l} = (\dot{l}_1^T, \dot{l}_2^T)^T$, the efficient influence function from (2.8) may be rewritten as

$$\widetilde{l} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}(\dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2). \tag{2.11}$$

Note that the components of these random vectors live in the linear space of random variables with mean 0 and finite variance under θ_0. With covariance as inner product in this Hilbert space $L_2^0(P_0)$ (P_0 corresponds to θ_0), the projection of $\dot{l}_1 = \dot{l}_1(X, \theta_0)$ onto the linear span $[\dot{l}_2]$ of the components of $\dot{l}_2$ equals

$$\Pi_0(\dot{l}_1|[\dot{l}_2]) = I_{12}I_{22}^{-1}\dot{l}_2.$$

Here, the left hand side is a vector of projections, namely the projections of the components of $\dot{l}_1$. Consequently, (2.11) may be rewritten as

$$\widetilde{l} = (E_{\theta_0} l_1^* l_1^{*T})^{-1} l_1^* \qquad \text{with} \qquad l_1^* = \dot{l}_1 - \Pi_0(\dot{l}_1|[\dot{l}_2]); \tag{2.12}$$

l_1^* is called the *efficient score function for* θ_1. Similarly, it can be seen that

$$\Pi_0(\widetilde{l}|[\dot{l}_1]) = (E_{\theta_0}\dot{l}_1\dot{l}_1^T)^{-1}\dot{l}_1, \tag{2.13}$$

i.e. the projection on the span of $\dot{l}_1$ of the efficient influence function for estimating θ_1 in the presence of the nuisance parameter θ_2 equals the efficient influence function for estimating θ_1 when θ_2 is known; cf. (2.4).

To conclude this discussion of parametric models, note that orthogonality of $\dot{l}_1$ and $\dot{l}_2$ implies that $l_1^* = \dot{l}_1$ and hence

$$\widetilde{l} = (E_0\dot{l}_1\dot{l}_1^T)^{-1}\dot{l}_1 = \Pi_0(\widetilde{l}|[\dot{l}_1]). \tag{2.14}$$

In this case the bounds for estimation of θ_1 are the same for θ_2 unknown and θ_2 known. In other words, estimation of θ_1 is asymptotically as difficult when θ_2 is unknown as when θ_2 is known.

2.2. Bounds for Semiparametric Models.

Now consider estimation of a Euclidean parameter ν in a semiparametric model. Such a model contains many parametric submodels with parameter $\theta = (\theta_1^T, \theta_2^T)^T$ and $\theta_1 = \nu$. Define the efficient score function for ν at P_0 (with $\nu = \nu_0$) as in (2.12) with $[\dot{l}_2]$ replaced by the closed linear span $\dot{\mathcal{P}}_2$ of the components of all possible choices of $\dot{l}_2$. If $\{T_n\}$ is a sequence of estimators of ν which is regular at P_0 for any parametric submodel, then the limit distribution of $\sqrt{n}(T_n - \nu_0)$ is the convolution of some distribution and the normal distribution with mean 0 and covariance matrix

$$(E_0 l_1^* l_1^{*T})^{-1} \qquad \text{with} \qquad l_1^* = \dot{l}_1 - \Pi_0(\dot{l}_1|\dot{\mathcal{P}}_2). \tag{2.15}$$

Moreover, $\{T_n\}$ is efficient if it is asymptotically linear as in (2.5) with

$$\widetilde{l} = (E_0 l_1^* l_1^{*T})^{-1} l_1^*. \tag{2.16}$$

This is a generalization of the Hájek - Le Cam convolution theorem to semiparametric models. A natural idea in constructing an information bound in a semiparametric model for estimation of a one-dimensional parameter ν is to consider the parametric bounds

$$\left\{E_0(\dot{l}_1 - \Pi_0(\dot{l}_1|[\dot{l}_2]))^2\right\}^{-1}$$

and maximize these over $[\dot{l}_2]$. Since $[\dot{l}_2] \subset \dot{\mathcal{P}}_2$, this yields (2.15) , provided the components of $\Pi_0(_1|\dot{\mathcal{P}}_2)$ can be written as the $L_2^0(P_0)-$limits of sequences of $_2$'s.

In some semiparametric models $\dot{l}_1$ is orthogonal to $\dot{\mathcal{P}}_2$ and consequently $l_1^* = \dot{l}_1$ in (2.15). In such a situation it is possible, at least in principle, to estimate ν asymptotically as well not knowing the non-Euclidean nuisance parameter as when it is known. In this case, an efficient estimator is called *adaptive* since it adapts to the unknown nuisance parameter. This phenomenon was noticed by Stein (1956) first.

Consider estimation of a Euclidean parameter ν on a parametric, semiparametric, or even nonparametric model $\mathcal{P}$ with tangent space $\dot{\mathcal{P}}$, i.e. $\nu : \mathcal{P} \to R^m$ and $\dot{\mathcal{P}}$ is the closed linear span in $L_2^0(P_0)$ of all score functions of (one-dimensional) regular parametric submodels of $\mathcal{P}$. This parameter ν is called *pathwise differentiable* at P_0 if there exists a $\dot{\nu} \in \dot{\mathcal{P}}$ such that for any one-dimensional regular parametric submodel $\{P_\eta : \eta \in R\}$ with score function $h \in \dot{\mathcal{P}} \subset L_2^0(P_0)$ (roughly $h = (\partial/\partial\eta) \log dP_\eta/d\mu$)

$$\nu(P_\eta) = \nu(P_0) + \eta\langle\dot{\nu}, h\rangle_0 + o(|\eta|), \tag{2.17}$$

where $\langle\cdot,\cdot\rangle_0$ is the inner product, i.e. covariance, in $L_2^0(P_0)$. In parametric models it can be seen that $\dot{\nu}$ equals the efficient influence function $\widetilde{l}$ from (2.4) or (2.8). It can be proved also that in semiparametric models $\dot{\nu} = \widetilde{l}$ from (2.16).

Let ν_e by an extension of ν defined on a larger model $\mathcal{P}_e$. Often, $\dot{\nu}_e$ is quite easy to compute for appropriately large (nonparametric) $\mathcal{P}_e$, and (2.17) together with the same relation for ν_e yields $\dot{\nu}_e - \dot{\nu} \perp \dot{\mathcal{P}}$; in other words

$$\dot{\nu} = \Pi_0(\dot{\nu}_e|\dot{\mathcal{P}}). \tag{2.18}$$

Therefore, this yields another method to determine efficient influence functions, which for parametric models reduces to (2.13).

Often $\mathcal{P}_e$ will be a nonparametric extension of the semiparametric model. Here it makes sense to use a technical definition of the distinction between semiparametric and nonparametric models, thus completing the discussion in section 1.1. Nonparametric models $\mathcal{P}$ have maximal tangent spaces $\dot{\mathcal{P}} = L_2^0(P_0)$. On the other hand, semiparametric models $\mathcal{P}$ have tangent spaces $\dot{\mathcal{P}}$ which are *not* finite - dimensional and are also *proper* subspaces of $L_2^0(P_0)$. For example, in the case of the symmetric location model of Example 1.2.1, at any $P_0 \in \mathcal{P}$ with finite Fisher information for location, the tangent space $\dot{\mathcal{P}}$ is the span of the usual score for location in sampling from g (translated by θ) and all the even functions (about θ) in $L_2^0(P_0)$ - which is a proper subspace of $L_2^0(P_0)$.

All of the development in this section has been restricted to differentiable functions or parameters in semiparametric models based on i.i.d. data. For an important characterization of pathwise differentiability of implicitly defined functions ν of the form $\nu(P_{\theta,G}) = \psi(G)$ for a pathwise differentiable function ψ in terms of the *score operator for* G, see Van der Vaart (1991a). For a brief discussion of extensions beyond i.i.d, LAN, and differentiable functions, see section 5.

3. Estimation Methods.

3.1. Efficient Estimates and Achieving Information Bounds.

The asymptotic information bounds derived for parametric models are achievable in the sense that for any regular model there are regular efficient estimators (Le Cam, 1986). On the other hand with semiparametric models the information bounds, even for the Euclidean parameter, are not necessarily achievable. The main difficulty is finding a good enough initial estimator of the parameters. Ritov and Bickel (1990) presented two models in which the information is strictly positive (and even infinite) but no $\sqrt{n}-$consistent estimator exists. Actually there is no algebraic rate that is attainable on compacts.

It is true, as Ritov and Bickel (1990) show, that the information bound for a Euclidean parameter can be achieved if the semiparametric model is a union of nested smooth finite-dimensional parametric models. This is the situation for which the BIC model selection criterion is appropriate. This suggests, that if we consider less restrictive semiparametric models smoothly parametrized by Euclidean and abstract parameters, but of "small entropy"

then again the information bounds for the Euclidean parameters should be achievable. This is the case in the examples of non-achievement of the bounds given by Ritov and Bickel (1990): the phenomenon vanishes when the non-Euclidean part of the parameter indexing the model is assumed sufficiently smooth.

Many authors have given constructions of efficient estimators in specific models, with the classical first cases being those treated by Stein (1956). Van Eeden (1970) was the first to explicitly construct adaptive estimators of location for Example 1.2.1 under the extra assumption that g be strongly unimodal. Her work was based on the adaptive rank test of Hájek (1962). Fully adaptive location estimators were given by Stone (1975), and by Beran (1978), ((1974) under an extra regularity condition). Bickel (1982) described a general approach to construction of adaptive estimators of finite-dimensional parameters. See also Fabian and Hannan (1982).

A general scheme for construction of efficient estimators has been described by Klaassen (1987), Schick (1986, 1987 and 1993), and BKRW (pp. 394–413). The general scheme involves, implicitly or explicitly, the following stages:

A. First one should have an estimator $\tilde{\theta}_n$ of θ which is $n^{1/2}-$consistent.

B. Then one should have an estimator $\tilde{\psi}(x, \theta; X_1, \ldots, X_n)$ of the efficient influence function such that $\sqrt{n} \int \tilde{\psi}(x, \theta) dP_{\theta,G}(x) = o_{P_{\theta,G}}(1)$ and $\int (\tilde{\psi}(x, \theta) - \psi(x, \theta))^2 dP_{\theta,G}(x) = o_{P_{\theta,G}}(1)$.

C. The efficient estimator is then constructed as $\tilde{\theta}_n$ plus an average of $\tilde{\psi}$ over the observations, essentially as a one-step Newton-Raphson approximation.

In some cases some technical modifications, such as data splitting, i.e., using different parts of the data for the construction of $\tilde{\theta}$, $\tilde{\psi}$ (this is originally due to Hájek (1962)), and truncation of the estimators to a grid (this is originally due to Le Cam (1956)) is needed for proof of the efficiency of the estimator.

In many cases more direct methods have proved effective. In particular, non-parametric maximum likelihood estimates and some variants, including profile likelihood methods (see Severini and Wong (1992) and Huang (1994a), (1994b)) have been proved to be efficient in many cases. An important example is the maximum partial likelihood estimator in the Cox proportional hazards model. Many of these results depend on the efficiency of the generalized maximum likelihood estimators of the distribution function such as the empirical distribution and the product limit estimator; see Gill (1989), Gill and Van der Vaart (1993), and Van der Vaart (1994a). For a class of "smooth" semiparametric models, Huang (1994a) has recently used empirical process methods to show that maximum likelihood estimators are asymptotically efficient, and has applied this result to the proportional hazards model, Example 1.2.8, with interval censored data. Huang (1994b) applies the same method to a proportional odds regression model. For a further example of the use of modern empirical process methods to the study of MLE's in a semiparametric model setting, see Van der Vaart (1994b).

3.2. Inefficient Estimates.

As seen in the preceeding discussion of efficient estimators, a crucial first step is often to obtain a consistent or $\sqrt{n}-$consistent preliminary estimator. Often such estimators will be of importance in their own right because of computational or robustness issues.

The variety of approaches to construction of preliminary estimators is staggeringly large, but certainly includes moment estimates, minimum distance estimators, estimates based on estimating equations, pseudo-likelihood (of several types ... since this terminology is used in several different ways), and nonparametric "principle of substitution" estimators. For an extensive discussion with connections to the econometric literature and applications, see Newey and McFadden (1994).

4. Testing and Confidence Intervals.

Unfortunately a general theory of hypothesis testing for semiparametric models does not yet exist. Nevertheless, some of the general principles from regular parametric models do carry over in many cases. Here we briefly discuss what is known (to us) concerning testing in semiparametric models.

4.1. Testing a Parametric Component within a Semiparametric Model.

Consider testing

$$H : \theta = \theta_0 \qquad \text{versus} \qquad K : \theta \neq \theta_0 \tag{4.1}$$

in the context of a semiparametric model

$$\mathcal{P} = \{P_{\theta,G} : \theta \in \Theta,\ G \in \mathcal{G}\}. \tag{4.2}$$

Different solutions for this testing problem have been given in different specific semiparametric models. For example, Hájek (1962) and Hájek and Šidák (1967), Section VII.1.6, constructed asymptotically efficient tests of (4.1) in the context of Example 1.2.2 and Example 1.2.1 respectively and Cox (1972) suggested tests of (4.1) in the context of Example 1.2.8 based on partial likelihood. Unfortunately, however, there does not yet exist a suitable general theory for semiparametric models of appropriate analogues of the trinity of tests (Wald, likelihood ratio, and score or Rao statistics) which are known to be asymptotically equivalent (and efficient in a certain sense) for local alternatives in the context of classical regular parametric models. Of course the most obvious asymptotic tests would indeed be based on Wald - type statistics of the form

$$W_n = n(\hat{\theta}_n - \theta_0)^T \hat{I}_n (\hat{\theta}_n - \theta_0) \tag{4.3}$$

where $\hat{\theta}_n$ is an asymptotically efficient estimator of θ and $\hat{I}_n$ is a consistent estimator of $I(\theta) = I(P_{\theta,G}|\theta,\mathcal{P})$, or at least of $I(\theta_0) = I(P_{\theta_0,G}|\theta,\mathcal{P})$; here $I(P_{\theta_0,G}|\theta,\mathcal{P})$ denotes the information matrix at $P_{\theta_0,G}$ for estimation of θ within the model $\mathcal{P}$ as defined by (2.15), namely $E_0 l_1^* l_1^{*T}$. This last question, construction of consistent estimators of $I(\theta)$, or even $I(\theta_0)$ for fixed θ_0, has not yet been satisfactorily resolved in general, even though methods for such constructions are clear in particular cases. To obtain robustness of validity of such a test, the approach taken by Lin and Wei (1989) for the Cox model, Example 1.2.8, is often sensible and fruitful. In the present testing context, this entails study of the asymptotic properties of the estimator $\hat{\theta}_n$ under consideration in (4.3) *off the model* $\mathcal{P}$, and then use of an appropriate "sandwich estimator" of its asymptotic variance in (4.3) rather than the simpler estimator $\hat{I}_n$ based on validity of the model.

Analogues of the Rao (or score) tests are also frequently available. Again the main additional problem (beyond those arising in estimation) is construction of consistent estimators of $I(\theta_0)$. If $\hat{I}_n$ is such a consistent estimator, then a reasonable test might be based on

$$R_n = S_n^*(\theta_0)^T \hat{I}_n^{-1} S_n^*(\theta_0) \tag{4.4}$$

where

$$S_n^*(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_\theta^*(X_i;\theta_0,\hat{G}_n) \tag{4.5}$$

where $l_\theta^* = \Pi_0(\dot{l}_\theta|\dot{\mathcal{P}}_2^\perp)$ as in (2.15) and $\hat{G}_n$ is some (consistent) estimator of G. Again, there is no general theory available yet, but only success stories in particular instances such as Example 1.2.8.

Analogues of the classical likelihood ratio statistics are still less - well understood, although profile likelihood methods perform well in particular cases such as Examples 1.2.7, 8, and 12. See also Rotnitzky and Jewell (1990).

4.2. Testing a Nonparametric Component within a Semiparametric Model.

Now consider testing

$$H : G = G_0 \qquad \text{versus} \qquad K : G \neq G_0 \tag{4.6}$$

for some fixed $G_0 \in \mathcal{G}$ in the context of a semiparametric model (4.2). Still less is known generally for this problem than for the problem in the previous section, but particular examples have received considerable attention. For example Cox, Koh, Wahba, and Yandell (1988) consider testing $H : \eta = 0$ versus $K : \eta > 0$ in the regression model

$$Y = \theta^T Z_1 + \eta r(Z_2) + \epsilon$$

where $r \in \mathcal{R}$, a collection of smooth functions and $\epsilon \sim N(0, \sigma^2)$ is independent of $Z = (Z_1, Z_2)$. They found the locally most powerful test of H versus K. See Simon and Smith (1991) for related problems.

4.3. Testing Goodness-of-Fit of a Semiparametric Model.

Now let $\mathcal{P}$ be a given semiparametric model, and consider testing

$$H : P \in \mathcal{P} \qquad \text{versus} \qquad K : P \in \mathcal{P}^c. \tag{4.7}$$

Perhaps the best known example of this is the case when the semiparametric model $\mathcal{P}$ is given by Example 1.2.1. Thus the testing problem is that of testing symmetry about an (unknown!) point of symmetry. This problem itself has a large literature; see e.g. Gupta (1967), Gastwirth (1971), and Boos (1982). [The easier problem of testing symmetry about a known point has a still larger literature.]

Another example with a rapidly growing literature is that of testing (4.7) when $\mathcal{P}$ is the Cox proportional hazards model, Example 1.2.8; see e.g. Gill and Schumacher (1987), Horowitz and Neumann (1992), Nagelkerke, Osting, and Hart (1984), and Therneau, Grambsch, and Fleming (1990).

For a general approach to problems of this type, see Bickel and Ritov (1992).

5. Extensions.

The theory sketched in sections 2 and 3 is in the process of being extended in several directions. Here we focus on going beyond i.i.d. and on robustness.

5.1. Beyond i.i.d., LAN, and Differentiable Functions.

Dropping the i.i.d. assumption is in principle straightforward whenever Local Asymptotic Normality (LAN) continues to hold. In fact, this was accomplished for models with finite-dimensional parameter spaces by Hájek (1970), (1972): Hájek's convolution and asymptotic minimax theorems are not restricted to the i.i.d. case, but are in force for any model with finite-dimensional parameter space for which LAN holds.

For models in which the parameter space can be taken to be a (pre-)Hilbert space H and the differentiable parameters to be estimated take values in a Banach space, generalizations of Hájek's theorem under a LAN assumption have been given by Levit (1978), Millar (1983), Van der Vaart and Wellner (1989), (1995), and Bickel (1993). These lower bound results cover a wide range of independent but non-identically distributed models, continuous state Markov models, and models for stationary ergodic processes; see Millar (1983), Bickel (1993), and Greenwood and Wefelmeyer (1989 - 1993) for many examples. The "calculus" for computing

asymptotic information bounds in these models exists, but it is still relatively difficult to use, and its utility, which has been clearly demonstrated for the i.i.d. case, remains to be established in the context of this larger class of models.

A particular class of non i.i.d. models are time series models. Most of these are constructed as follows: the present observation X_t is a location - scale transformation of the present innovation ϵ_t, where the location and scale parameters depend as well on a Euclidean parameter as on the past and possibly on exogeneous variables. The unknown distributions of the innovations and the exogeneous variables are the infinite-dimensional parameters in this semiparametric model. More explicitly,

$$X_t = \mu_t(\theta) + \sigma_t(\theta)\epsilon_t,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d., $\mu_t(\theta)$ and $\sigma_t(\theta)$ depend on the past. Time series models fitting into this framework are ARMA, ARCH, GARCH, TAR, EXPAR, etc., but also linear regressin fits in (Example 1.2.2 with $Z_1, \ldots, Z_n$ not necessarily i.i.d.). Under quite general conditions LAN has been shown to hold and adaptive estimators have been constructed by Drost, Klaassen, and Werker (1994b); adaptivity has been discussed in this context by Drost, Klaassen, and Werker (1994a). With $\sigma_t(\theta) = 1$ efficient estimators have been obtained by Jeganathan (1994) and Koul and Schick (1995).

Going beyond LAN in Le Cam's general theory has been successfully accomplished in only a few cases, the most notable being models satisfying a generalization of LAN known as "Local Asymptotic Mixed Normality" (LAMN); see Jeganathan (1980), (1982), and Le Cam and Yang (1990), Section 5.6. Still further extensions can be found in Jeganathan (1994) and Phillips (1988) in connection with multiple (co-integrated) time series models. To the best of our knowledge none of these results have been extended to honestly semiparametric contexts. Also see the examples in Van der Vaart (1991b), section 8.

Going beyond differentiable parameters is an extremely active area in nonparametric and semiparametric estimation theory at present, but there is little general theory bridging the gap between the differentiable/regular cases and the non-differentiable/irregular cases. For some examples of the non-differentiable cases, see Birgé (1983), Millar (1983), Stone (1982), (1985), (1986), and Donoho and Liu (1991a, 1991b).

5.2. Robust Estimation and Semiparametric Models.

Suppose we postulate a semiparametric model for data that we treat as an i.i.d. sample, but we suspect that the data has been contaminated by gross errors of some type. It is reasonable to look for estimators that are relatively efficient under the semiparametric model, but are robust against departures from this model.

Various ad hoc robust estimators have been suggested in the literature. Jones (1991), Sasieni (1993a), (1993b), among others, discussed different robust modifications of the partial

likelihood estimator for the Cox model. Thus Jones (1991) considers the family of estimators defined by estimating equations of the form: $W(t|q,Z) = \int_0^t q_n(s) \sum (Z_j(s) - \bar{Z}(s)) dN_j(s)$ where $Z(\cdot)$ is some predictable function which depends on the covariate $X_j(s)$ as well as on all the history up to time s. Different choices of q and Z generate different estimators with different robustness and efficiency tradeoffs.

There are essentially three points of view in the systematic asymptotic analysis of robust estimators in the parametric context:

A. The parametric model is extended to a semiparametric model incorporating gross errors in which the original parameter remains identifiable. An example of this is the original Huber extension of the Gaussian location model (Huber, 1964), in which the contamination is assumed symmetric and hence the center of symmetry is the parameter to estimate. In such a case, the usual semiparametric information bound relative to the larger model, applies to robust estimators. In that sense, the adaptive estimator of location, constructed carefully, is the solution of Huber's problem, rather than Huber's minimax M estimator. In this sense, robust estimation is simply efficient estimation in particular semiparametric models.

B. The parameter to be estimated is extended uniquely to a nonparametric model. This is to some extent the approach of Bickel and Lehmann (1975-79) and Millar (1979). In this case there is no robustness problem since we need only use the asymptotically unique regular estimator of whatever parameter we have chosen, whether defined by properties as in Bickel and Lehmann (1975–79), or by minimizing an appropriate distance between a specified model and the true completely unknown P as in Millar (1979).

C. The parametric model is extended to a semiparametric (contamination) neighborhood of itself depending on n (for example contamination of the order $O(n^{-1/2})$), so that the parameter is asymptotically identifiable. The key point is that the contamination contributes bias which is of the same order as the standard deviation for an estimator which is regular under the parametric model. This point of view was formally espoused by Huber Carol (1970), Jaeckel (1971) and further developed by Bickel (1979) and Rieder (1978). However, its solution and underpinnings when the semiparametric neighborhood is arbitrary $O(n^{-1/2})$ contamination are in Hampel (1974) and Hampel et al. (1986). The key notion of the influence function and the basic methods for finding optimal estimators subject to bounds on the influence function are introduced there.

Within the semiparametric context only approach C leads to new points of view. Shen (1990), (1995) considers contamination neighborhoods of semiparametric models, thus pursuing this approach. Following Shen (1990) we may consider asymptotically linear estimators that are regular under the semiparametric models and have bounded influence functions. Let $\hat{\theta}_n$ be such an estimator of θ and let $\mathcal{P}_{\theta,G,\epsilon,n} = \{(1 - \epsilon n^{-1/2}) P_{\theta,G} + \epsilon n^{-1/2} H : H \in \mathcal{M}\}$; recall that $\mathcal{M}$ consists of all probability measures. Let ψ be the influence function of $\hat{\theta}_n$. Then it is clear that the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$ is $\int \psi \psi^T dP_{\theta,G}$ while its asymptotic bias is at most $\epsilon ||\psi||_\infty$. This leads Shen (1990), (1995) to analysis of the following

problem. Find ψ that minimizes $\int ||\psi||^2 dP_{\theta,G}$ among all influence functions ψ that satisfy $||\psi||_\infty < C$, $\int \psi dP_{\theta,G} = 0$, $\int \psi l_1^T dP_{\theta,G} = J$ and $\psi \perp \dot{\mathcal{P}}_2$ (the last three conditions are needed to ensure the regularity of the estimator, see (4.20) of Klaassen (1987) and Propositions 2.4.2 and 3.3.1 of BKRW (1993)). Shen proved that if there is a solution, it is essentially unique. It is not difficult to see that the robust influence function must be of the form $h_C(A\dot{l}_1 + a)$ where $h_C(x) = \min\{|x|, C\}x/|x|$, A is a matrix and $a \in \dot{\mathcal{P}}_2$. Unfortunately, finding A and $a(\cdot)$ is typically complicated since they both depend on (θ, G). In some models it can be verified that $a \equiv 0$ is the solution. This is the case in the symmetric location model and some two sample problems. More generally it is possible to calculate when we can "condition G out". Shen (1990), (1995) applies this notion to exponential family mixture models and heteroscedastic regression models.

Beran (1981) considered approach C for parametric models when the contamination is so restricted that robust and efficient estimation essentially coincide provided one truncates mildly. His approach was adapted to semiparametric models by Wu (1990). Here it is assumed that the true distribution belongs to a $\sqrt{n}-$Hellinger neighborhood of a semiparametric model $\{P_{\theta,G}\}$. The parameter to be estimated is $\tilde{\theta}(P)$ defined by $\inf_G d_H(P_{\tilde{\theta},G}, P) = \inf_{\theta,G} d_H(P_{\theta,G}, P)$ where $d_H(\cdot,\cdot)$ is the Hellinger distance. In particular $\tilde{\theta}(P_{\theta,G}) = \theta$. Wu (1990) found the same phenomenon in the semiparametric case as Beran (1981) did in the parametric. His particular examples were transformation models and exponential mixture models.

REFERENCES

Beran, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.* **2**, 63-74.

Beran, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6**, 292-313.

Beran, R. (1981). Efficient robust estimates in parametric models. *Z. Wahrsch. Verw. Gebiete*, **55**, 91 - 109.

Bickel, P.J. (1979). Quelques aspects de la statistique robuste. *École d'Été de Probabilités de St. Flour IX.* Springer Lecture Notes in Mathematics **876**, 1 - 72.

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*, 55 - 73, C. R. Rao, editor. Elsevier, Amsterdam.

Bickel, P. J., and Ritov, Y. (1992). Testing for goodness of fit: A new approach. *Nonparametric Stat. and Related Topics*; 51-57. Elsevier, A. K. Md. E. Saleh, editor.

Bickel, P.J. and Lehmann, E.L. (1975). Descriptive statistics for nonparametric models: I. Introduction. *Ann. Statist.* **3**, 1031-1045,

Bickel, P.J. and Lehmann, E.L. (1975). Descriptive statistics for nonparametric models: II. Location. *Ann. Statist.* **3**, 1045 - 1069.

Bickel, P.J. and Lehmann, E.L. (1976). Descriptive statistics for nonparametric models: III. Dispersion. *Ann. Statist.* **4**, 1139-1158.

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

Birgé, L. (1983). Approximations dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. verw. Gebiete* **65**, 181 - 237.

Boos, D. D. (1982). A test for asymmetry associated with the Hodges-Lehmann estimator. *J. Amer. Statist. Assoc.* **77**, 647-651.

Cox, D., Koh, E., Wahba, G., and Yandell, B. S. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.* **16**, 113 - 119.

Cox, D.R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Donoho, D. L. and Liu, R. C. (1991a). Geometrizing rates of convergence, II. *Ann. Statist.* **19**, 633-667.

Donoho, D. L. and Liu, R. C. (1991b). Geometrizing rates of convergence, III. *Ann. Statist.* **19**, 668-701.

Drost, F. C., Klaassen, C.A.J., and Werker, B.J.M. (1994a). Adaptiveness in time-series models. In *Asymptotic Statistics*, 203 - 212, P. Mandl and M. Hušková, editors, Physica-Verlag, New York.

Drost, F. C., Klaassen, C.A.J., and Werker, B.J.M. (1994b). Adaptive estimation in time-series models. *Technical Report* **9488**, Center for Economic Research, Tilburg University.

Fabian, V. and Hannan, J. (1982). On estimation and adaptive estimation for locally asymptotically normal families. *Z. Wahrsch. verw. Gebiete* **59**, 459-478.

Faraway, J. (1990). Implementing semiparametric density estimation. Stat. and Prob. Letters **10**, 141 - 163.

Gastwirth, J. L. (1971). On the sign test for symmetry. *J. Amer. Statist. Assoc.* **66**, 821-823.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1) *Scand. J. Statist.* **16**, 97 - 124.

Gill, R. D. and Schumacher, M. (1987). A simple test of the proportional hazards assumption. *Biometrika* **74**, 289-300.

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.

Greenwood, P.E. and Wefelmeyer, W. (1989). Efiicient estimating equations for nonparametric filtered models. In *Statistical Inference in Stochastic Processes* **1**, 107 - 141, Marcell Dekker, New York.

Greenwood, P.E. and Wefelmeyer, W. (1990). Efficiency of estimators for partially specified filtered models. *Stoch. Proc. and their Applic.* **36**, 353 - 370.

Greenwood, P.E. and Wefelmeyer, W. (1991). Efficient estimation in a nonlinear counting-process regression model. *Canad. J. Statist.* **19**, 165 - 178.

Greenwood, P.E. and Wefelmeyer, W. (1992a). Efficiency of empirical estimators for Markov chains. *Preprints in Statistics* **135**, University of Cologne.

Greenwood, P.E. and Wefelmeyer, W. (1992b). Nonparametric estimators for Markov step processes. *Preprints in Statistics* **136**, University of Cologne.

Greenwood, P.E. and Wefelmeyer, W. (1992c). Optimality properties of empirical estimators for multivariate point processes. *Preprints in Statistics* **137**, University of Cologne.

Greenwood, P.E. and Wefelmeyer, W. (1993a). Maximum likelihood estimator and Kullback - Leibler information in misspecified Markov chain models. *Preprints in Statistics* **141**, University of Cologne.

Greenwood, P.E. and Wefelmeyer, W. (1993b). Empirical Estimators for semi-Markov processes. *Preprints in Statistics* **142**, University of Cologne.

Gupta, M. K. (1967). Asymptotically nonparametric tests of symmetry. *Ann. Math. Statist.* **38**, 849-866.

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

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Gebiete* **14**, 323 - 330.

Hájek, J. (1972). Local asymptotic minimax and admissiblity in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **1**, 175 - 194.

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 383 - 393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* John Wiley and Sons, New York.

Horowitz, J. L. and Neumann, G. R. (1992). A generalized moments specification test of the proportional hazards model. *J. Amer. Statist. Assoc.* **87**, 234-240.

Huang, J. (1994a). Efficient estimation for the Cox model with interval censoring. *Technical Report* **274**, Department of Statistics, University of Washington, Seattle. Submitted to *Ann. Statist.*

Huang, J. (1994b). Maximum likelihood estimation for proportional odds regression model with current status data. *Preprint*, University of Iowa.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73 - 101.

Huber-Carol, C. (1970). *Etude asymptotique de tests robustes.* Ph.D. Thesis, ETH Zurich.

Jaeckel, L. (1971). Robust estimates of location: symmetric and asymmetric contamination. *Am. Math. Statist.* **42**, 1020 - 1034.

Jeganathan, P. (1980). An extension of a result of L. Le Cam concerning asymptotic normality. *Sankhya Ser. A* **42**, 146-160.

Jeganathan, P. (1982). On the asymptotic theory of estimation when the limit of the log-likelihood is mixed normal. *Sankhya Ser. A* **44**, 173-212.

Jeganathan, P. (1994). Some aspects of asymptotic theory with applications to time series models. *Preprint*, University of Michigan, Ann Arbor.

Jones, M. P. (1991). Robust tests for survival data involving a single continuous covariate. *Scand. J. Statist.* **18**, 323 - 332.

Klaassen, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimates. *Ann. Statist.* **15**, 1548 - 1562.

Koul, H. L. and Schick, A. (1995). Efficient estimation in nonlinear time series models. *Preprint*, Michigan State University, East Lansing.

LeCam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1**, 129-156. Univ. California Press, Berkeley.

LeCam, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

LeCam, L. and Yang, G. (1990). *Asymptotics in statistics: Some basic concepts.* Springer-Verlag, New York.

Levit, B. Ya (1978). Infinite-dimensional informational lower bounds. *Theory Probab. Its Appl.* **23**, 388 - 394.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Amer. Stat. Assoc.* **84**, 1074 - 1078.

Millar, P.W. (1979). Asymptotic minimax theorems for the sample distribution function. *Z. Warsch. Verw. Gebiete* **48**, 233-252.

Millar, P.W. (1983). The minimax principle in asymptotic statistical theory. *École d'Été de Probabilités de St. Four XI.* Springer Lecture Notes in Mathematics **976**, 76 - 267.

Nagelkerke, N. J. D., Oosting, J., and Hart, A. A. M. (1984). A simple test for goodness-of-fit of Cox's proportional hazards model *Biometrics* **40**, 483-486. (Corr: **40**, 1217.)

Newey, W.K., and McFadden, D.L. (1994). Large Sample Estimation and Hypothesis Testing, In *Handbook of Econometrics, Volume IV*, 2111 - 2245, Engle, R. F. and McFadden, D.L., editors, Elsevier, Amsterdam.

Olkin, I. and Spiegelman, C. H. (1987). A semiparametric approach to density estimation. *J. Amer. Stat. Assoc.* **82**, 858 - 865.

Phillips, P.C.B (1988). Multiple Regression with Integrated Time Series. In, Statistical Inference from Stochastic Processes, *Contemporary Mathematics* **80**, 79 - 107.

Powell, J. L. (1994). Estimation of Semiparametric Models, In *Handbook of Econometrics, Volume IV*, 2443 -2521, Engle, R. F. and McFadden, D.L., editors, Elsevier, Amsterdam.

Rieder, H. (1978). A robust asymptotic testing model. *Ann. Statist.* **6**, 1080-1094.

Ritov, Y. and Bickel, P. J. (1990). Achieving information bounds in non and semiparametric models. *Ann. Statist.* **18**, 925 - 938.

Roeder, K. (1992). Semiparametric estimation of normal mixture densities. *Ann. Statist.* **20**, 929 - 943.

Rotnitzky, A. and Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485-497.

Sasieni, P. (1993a). Maximum weighted partial likelihood estimators for the Cox model. *J. Amer. Statist. Assoc.* **88**, 144 - 152.

Sasieni, P. (1993b). Some new estimators for Cox regression. *Ann. Statist.* **21**, 1721 - 1759.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14**, 1139 - 1151.

Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Planning Inf.* **16**, 89 - 105.

Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.* **21**, 1486 - 1521.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768 - 1862.

Shen, Z. (1990). *Robust Estimation in Semiparametric Models*. Ph.D. Thesis, Department of Statistics, University of California, Berkeley.

Shen, Z. (1995). Optimal B robust influence functions in semiparametric models. *Ann. Statist.* **23**, to appear.

Simon, P. and Smith, R. J. (1991). Distributional specification tests against semiparametric alternatives. *J. of Econometrics* **47**, 175-194.

Stein, C. (1956). Efficient nonparametric testing and estimation. In *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1**, 187 - 195, Univ. of California Press.

Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3**, 267-284.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040 - 1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689 - 705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14**, 590-606.

Therneau, T. M., Grambsch, P. M., and Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147-160.

Van der Vaart, A. W. (1991a). On differentiable functions. *Ann. Statist.* **19**, 178 - 204.

Van der Vaart, A. W. (1991b). An asymptotic representation theorem. *Int. Statist. Rev.* **59**, 97 - 121.

Van der Vaart, A. W. (1994a). Efficiency of infinite-dimensional M-estimators. *Statistica Neerlandica* **48**, 9 - 30.

Van der Vaart, A. W. (1994b). Maximum likelihood estimation with partially censored data. *Ann. Statist* **22**, 1896 - 1916.

Van der Vaart, A. W. (1995). Semiparametric models: an evaluation. *Statistica Neerlandica* **49**, 111 - 125.

Van der Vaart, A. W. and Wellner, J. A. (1989). Prohorov and continuous mapping theorems in the Hoffmann-Jørgensen weak convergence theory with applications to convolution and asymptotic minimax theorems. *Technical Report* **157**, Department of Statistics, University of Washington, Seattle.

Van der Vaart, A. W. and Wellner, J. A. (1995). *Weak Convergence and Empirical Processes.* Springer Verlag, New York, to appear.

Van Eeden, C. (1970). Efficiency robust estimation of location. *Ann. Statist.* **41**, 172-181.

Wu, C. O. (1990). *Asymptotically Efficient Robust Estimation in Some Semiparametric Models*. Ph.D. Thesis, Department of Statistics, University of California, Berkeley.

Department of Statistics
University of California
Berkeley, California 94720-4735

Department of Mathematics
University of Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
The Netherlands

Department of Statistics
Hebrew University
Jerusalem 91905, Israel

University of Washington
Statistics
Box 354322
Seattle, Washington 98195-4322