Semiparametric Models: Progress and Problems[1]

Jon A. Wellner
Department of Statistics, University of Washington
Seattle, Wa.98195

Semiparametric models, models which incorporate both parametric (finite-dimensional) and nonparametric (infinite-dimensional) components, have received increasing use and attention in statistics in recent years. This paper reviews developments in this very large and rich class of models which spans the middle ground between parametric and nonparametric models. Attention is devoted to a preliminary classification of such models with comments on recent work, to lower bounds for estimation, to two potentially useful methods for construction of efficient estimates, and to open problems.

1. Introduction

Models for phenomena involving randomness play a key role in statistics. If $\mathbf{P}_{all}$ denotes the collection of all probability distributions on a sample space $\mathbf{X}$ of the observations X, a model $\mathbf{P}$ is a subset of $\mathbf{P}_{all}$: thus we assume in constructing a model $\mathbf{P}$ that X has a distribution P in $\mathbf{P}$, and we write $X \cong P \in \mathbf{P}$. The sample space $\mathbf{X}$ is the set of all possible observations.

A statistician uses the observations X to make inferences about the 'true' probability distribution P, and hence about real-world phenomena in question. A common form of inference is *point estimation.* For example, if X represents the life expectancy or survival time of an individual who has been given a new medical treatment, the statistician may be interested in using a sample of such individuals to estimate $\nu(P) \equiv P(X \geq t)$, the probability of survival beyond t time units. The choice of a model $\mathbf{P}$ can have a major effect on inferences about $\nu(P)$: If the model $\mathbf{P}$ is too small, the statistician runs the risk that the model will not contain the 'true' P, and the consequent price is bias in estimation of $\nu(P)$. In this case the model is not sufficiently large to be realistic and may fail to capture the essential features of the phenomena in question. On the other hand, if the model $\mathbf{P}$ is too large, the statistician may find himself in the position of estimating too many parameters from too little data. This tradeoff

1. This is a revised version of a paper presented at the Centenary Session of the International Statistical Institute, Amsterdam 1985, and which has appeared in the proceedings of that conference (*Bull. Int. Stat. Inst. 51 (4),* 23.1.1-23.1.20). It is reproduced here by kind permission of the International Statistical Institute.

between realism and parsimony is an ever-present theme in statistics; for interesting discussions of some aspects of model-building see Chapters 2 and 4 of COX and SNELL [23] and STONE [74].

Parametric models $\mathbf{P}_0 \equiv \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$ for some d play a dominant role in classical statistical theory. Such models, with a finite-dimensional parameter space Θ, form the basis of much of classical statistics. A difficulty with such parametric models is that typically a parametric model $\mathbf{P}_0$ is a relatively small subset of $\mathbf{P}_{all}$, and hence the 'true' distribution P of X may not be contained in $\mathbf{P}_0$.

One approach to this difficulty is the completely nonparametric approach: assume only that $P \in \mathbf{P}_{all}$ or a slight restriction of $\mathbf{P}_{all}$ requiring only some smoothness or monotonicity assumptions. While this approach seems to be feasible when the dimensionality of the sample space is small, it fails to take advantage of structure in the phenomena being modeled and begins to run into difficulty when the dimensionality of the sample space (and hence of the parameter space, $\mathbf{P}_{all}$ itself) is large.

A compromise strategy which gains in model realism and the flexibility needed to make use of the larger data sets which are increasingly available is the semiparametric approach: assume that some aspects or components of the model are parametric or finite-dimensional, while other aspects or components are allowed to be nonparametric or infinite-dimensional. Then the resulting *semiparametric model* $\mathbf{P}$ is typically of the form

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

where $\Theta \subset \mathbb{R}^d$ for some d and $\mathbf{G}$ is some (large) collection of functions. We also write

$$\mathbf{P} = \{P_\theta : \theta = (\theta_1, \theta_2) \text{ with } \theta_1 \in \Theta_1 \subset \mathbb{R}^d,\ \theta_2 \in \Theta_2\},$$

where Θ_2 is a collection of functions.

This semiparametric approach has proved to be very useful in a wide range of problems, and promises to play an increasingly important role in statistics. Our object here is to survey this extremely rich and flexible class of models (Section 2), and to briefly review the developing inference methods with emphasis on lower bounds for estimation and construction of efficient estimates of the parametric component of such models (Section 3 and 4). The survey of models and review of inference methods may be read independently of one another. The final section discusses open problems.

The notion of a semiparametric model is very general, and is already being used, at least implicitly, in situations involving observations which are not independent and identically distributed (iid). For simplicity, however, we restrict attention here to the iid case: throughout this paper $X_1, \ldots, X_n$ are iid $P \in \mathbf{P}$ where $\mathbf{P}$ is a parametric or semiparametric model.

2. Classes of semiparametric models

Little effort has been made to classify or categorize semiparametric models. While such an effort may be premature, it may also help to identify related models and aid in developing methods to apply to new problems. The following scheme should be regarded as provisional and temporary.

The classification of models given here has two fundamental categories: *basic models*, and *derived models*. The basic models consist of exponential family models, group models, and transformation models. The derived models include regression models, convolution models, mixing models, censoring models, and biased sampling models. Although this scheme is both redundant and possibly incomplete, it includes all the semiparametric models with which I am now familiar. The rest of this section elaborates on these categories, and provides examples of the models of the various types with some brief comments on recent work.

2.1. Basic Models

The following basic models serve as building blocks in the construction of semiparametric models.

2.1.1. Exponential family models. (A). These are familiar parametric models with density (with respect to some measure m)

$$p(x,\theta) = c(\theta)\exp\Big(\sum_{i=1}^{k} Q_i(\theta)T_i(x)\Big)h(x)$$

for $\theta\in\Theta\subset\mathbb{R}^k$, $x\in\mathbf{X}\subset\mathbb{R}^d$. While these are themselves completely parametric (finitely dimensional) models, they serve as building blocks for many interesting semiparametric models.

2.1.2. Group models. (B).

(1). The classical parametric model of this type is obtained as follows: suppose that $Y\cong G\equiv P_0$, a fixed distribution on $\mathbf{X}$, and let $\mathbf{V}$ denote a group of (one to one) transformations on $\mathbf{X}$ parametrized by $\theta\in\Theta\subset\mathbb{R}^k$. If $v_\theta\in\mathbf{V}$, let $X\equiv v_\theta(Y)\cong P_\theta$ for $\theta\in\Theta$.

Examples:

(a) Location. $\mathbf{X}=\mathbb{R}^d$, $v_\theta(x)=x+\theta$ with $\theta\in\mathbb{R}^d$, and $P_\theta=P_0(\cdot-\theta)$.

(b) Elliptic distributions. $\mathbf{X}=\mathbb{R}^d$, $v_\theta(x)=\theta^{-1/2}x$ where θ is positive definite and symmetric; $G\equiv P_0$ is spherically symmetric on $\mathbb{R}^d$. Then $\mathbf{P}=\{P_\theta:\theta\in\Theta\}$ is the P_0-family of elliptic distributions.

(c) Two-sample models. $\mathbf{X}=\mathbf{X}_0\times\mathbf{X}_0$, $\mathbf{V}=\mathbf{V}_0\times\mathbf{V}_0$ where $\mathbf{V}_0$ is a group of transformations on $\mathbf{X}_0$, $\theta=(\mu,\nu)\in\Theta_0\times\Theta_0\equiv\Theta$, $Y=(W,Z)$ with $W,Z\cong P_0$ independent, and $X=(v_\mu(W),v_\nu\circ v_\mu(Z))$.

(2). By letting the distribution P_0 in (1) range over some large class of probability distributions $\mathbf{G}$ small enough to still allow identification of θ, or at least some important functions of θ, yields a semiparametric model

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

Examples:

(a) If $\mathbf{X}=\mathbb{R}^1$ in 1(a) above and $\mathbf{G}$ is the family of distributions symmetric about 0, $\mathbf{P}$ is the classical symmetric location family.

(b) If $\mathbf{X}$ and Θ are as in 1(b) above and $\mathbf{G}$ is the family of all spherical symmetric distributions, then $\mathbf{P}$ is the family of all elliptic distributions; see e.g. BICKEL [6].

(c) If $\mathbf{X}$ and Θ are as in 1(c) and $\mathbf{G}$ is arbitrary, then ν is still identifiable; see STEIN [71] or PFANZAGL [64].

(3). Classical nonparametric statistical theory uses transformation groups which are not parametrizable by a Euclidean space; for example, all continuous monotone transformations from $\mathbb{R}$ to $\mathbb{R}$. See LEHMANN [51] page 24 and 25 for 'semiparametric subgroups' of the large group and note that examples 2(a) and 2(b) are of this type. A wealth of other 'semiparametric group' families are undoubtedly possible.

2.1.3 Transformation models. (C). These models typically map $(\theta,P)\to P_\theta$ where $\theta\in\Theta\subset\mathbb{R}^k$ and $P\in\mathbf{G}$, a collection of probability distributions on $\mathbf{X}$. The key feature is that the map $P_\theta=\psi(\theta,P)$ acts on P, or some function that is one-to-one with P, rather than on X as in the case of a group model.

The classical example of this type of model is that of a family of 'Lehmann alternatives' defined as follows (see LEHMANN [50]): Let $\mathbf{X}=\mathbb{R}^1$, suppose that $Y\cong G$ and let $\{B(\cdot,\theta):\theta\in\Theta\subset\mathbb{R}^k\}$ be a family of monotone transformations from $[0,1]$ to $[0,1]$ with $B(0,\theta)=0$, $B(1,\theta)=1$ for all $\theta\in\Theta$. Then $X\cong P_{\theta,G}$ has df (distribution function) $F_{\theta,G}(x)=B(G(x),\theta)$. Here are some particular cases.

Examples:

(a) $B_a(\mu,\theta)=1-(1-\mu)^\theta$ with $0<\theta<\infty$. This yields the *proportional hazards model*: $\Lambda_F(x)=\theta\Lambda_G(x)$ where Λ_F is the cumulative hazard function corresponding to F; see LEHMANN [50] and COX [22].

(b) $B_b(\mu,\theta)=\dfrac{\theta\mu}{\theta\mu+(1-\mu)}=\dfrac{\theta\mu(1-\mu)^{-1}}{1+\theta\mu(1-\mu)^{-1}}$ with $0<\theta<\infty$. This yields the *proportional odds model*

$$\frac{F(x)}{1-F(x)}=\theta\frac{G(x)}{1-G(x)};$$

see BENNETT [2].

(c) $B_c(\mu,\theta,\nu)=1-[1-\nu\theta\log(1-\mu)]^{-1/\nu}$, $0<\nu<\infty$, $\theta>0$. This yields the *semiparametric Pareto model* suggested by CLAYTON and CUZICK [19]. Note

that $B_c(\mu,\theta,\nu)\to B_a(\mu,\theta)$ as $\nu\to 0$ while Bennett's B_b is related to Clayton and Cuzick's B_c by

$$B_c(1-\exp(-\frac{\mu}{1-\mu}),\theta,1)=B_b(\mu,\theta).$$

These three models can all be written in the form

$$h(X)=-\log(\theta)+\epsilon \tag{1}$$

where $h(x)\equiv\log\Lambda_G(x)=\log[-\log(1-G(x))]$ and ϵ has the distribution:
(a) $F(x)=1-\exp(-e^x)$ (extreme value);
(b) $F(x)=1/(1+e^{-x})$ (logistic);
(c) $F(x)=1-1/(1+\nu x)^{1/\nu}$ (Pareto).

Because of the generality allowed for the transformations h, rank methods and partial likelihoods play an important role in analyzing these models. Note that (1) yields a transformation family linear model if $\theta=\exp(\gamma z)$, and shows that these models can be viewed as special cases of a type of model involving smooth transformations of both X and z considered by BREIMAN and FRIEDMAN [11]; see 2.2.1 below and DOKSUM [24].

2.2 Derived models

The following classes of models are all derived from the basic models given above.

2.2.1 Regression models. (D). Given a basic model of one of the three types described above, there is a straightforward recipe for constructing related regression models:

1. Start with an exponential family, group or transformation model $\mathbf{P}=\{P_{\theta,G}:\theta\in\Theta,G\in\mathbf{G}\}$ where θ is the finite-dimensional Euclidean component of the model and G is the nonparametric or infinite-dimensional component of the basic model.
2. Suppose that $Z\cong H$ on $\mathbb{R}^d$.
3. Given $Z=z$, replace θ (or a component thereof) in the basic model by a semiparametric regression function $r(\gamma,z)$ taking values in Θ where $\gamma\in\Gamma\subset$ some $\mathbb{R}^k$. Different forms for r ranging from parametric to nonparametric regression models, with many interesting intermediate semiparametric forms, are possible. For example:
   (a) Linear model: $r(\gamma,z)=\gamma z$;
   (a') Exponential linear model: $r(\gamma,z)=\exp(\gamma z)$;
   (b) Nonlinear: $r(\gamma,z)=r_0(\gamma,z)$ for a fixed known nonlinear function r_0;
   (c) Nonparametric: $r(\gamma,z)=r(z)$, with r smooth;
   (d) Semiparametric: $r(\gamma,z)=\gamma z_1+r(z_2)$, where $z=(z_1,z_2)$, and r is smooth;

(e) Projection pursuit: $r(\gamma,z)=r(\gamma z)$ where $|\gamma| = 1$ and $r:\mathbb{R}^1\to\mathbb{R}^1$ is smooth;

(f) Signal-noise: $r(\gamma z)$ where $r:\mathbb{R}^1\to\mathbb{R}^1$ is periodic with period 1 so that γ is a frequency parameter.

Combining various types of regression functions illustrated by (a) - (f) with the basic models A, B or C yields a rich collection of regression models, including parametric, semiparametric, and nonparametric models. STONE [74] gives an interesting survey and further references. A few selected examples with brief comments concerning recent work follow.

Examples:

(a) Combining basic model A with the regression model D(a) yields linear exponential family regression models; see e.g. LEHMANN [51] Chapter 3, pages 196 - 207.

(b) Combining the basic model B1(a) where P_0 is normal with D(a) yields classical parametric normal theory regression models; the extension to B2(a) yields semiparametric linear regression models with arbitrary (symmetric) error distributions.

(c) The basic model B1(a) (with P_0 a fixed distribution on $\mathbb{R}^1$; e.g. normal) combined with the semiparametric regression model D(d) leads to a very interesting class of regression models introduced by ENGLE, GRANGER, RICE and WEISS [26] to study effects of weather on electricity demand, and by WAHBA [79]. This model has one nonparametric component, the smooth regression function r. Generalizations with two nonparametric components by allowing the error distribution to be arbitrary are also of interest. A special case has been studied by SCHICK [70], while STONE [74] discusses a spectrum of related regression models.

(d) Combining B2(a) with D(e) leads to a model related to projection-pursuit regression; see FRIEDMAN and STUETZLE [27], STONE [74], and HUBER [37].

(e) Combining C1(a) with D(a′) yields Cox's (1972) proportional hazards model. Many variants on this model are possible and deserve further exploration. Replacement of the exponential with some other (fixed) non-negative function has been considered by PRENTICE and SELF [67], while C1(c) combined with D(a′) has been explored by CLAYTON and CUZICK [19]. TIBSHIRANI [76] considers a version of Cox's model with the linear function in $\exp(\gamma z)$ replaced by a sum of smooth but otherwise arbitrary functions $\sum_{i=1}^{k} r_i(z_i)$. See 2.2.2 below for related mixture models involving unobserved covariates.

(f) Combination of B1(a) or B2(a) with D(f) yields a semiparametric 'signal plus noise' model which extends classical parametric signal plus noise models. For the latter, see IBRAGIMOV and HAS'MINSKII [38]. MCDONALD [56] has some interesting preliminary work on semiparametric extensions. These models are of interest in astrophysical applications; see e.g. LAFLER and KINMAN [44] or STELLINGWERF [72].

2.2.2 Mixture models. (E). Mixture models can usually be viewed as the result of unobserved heterogeneity as follows: suppose that $X=(Y,Z)$ has a distribution of the form

$$P_{\theta,G,H}(Y\in A,Z\in B)=\int_B P_{\theta,G}(Y\in A \mid Z=z)dH(z).$$

Then if we can only observe Y, the observations have the *mixture distribution*

$$P_{\theta,G,H}(Y\in A)=\int P_{\theta,G}(Y\in A \mid Z=z)dH(z).$$

Examples:

(a) Paired exponentials. Suppose that $(Y\equiv(Y_1,Y_2)\mid Z=z)\cong$ (exponential (z),exponential(θz)):

$$f(y\mid z)=\theta z^2\exp(-(zy_1+\theta zy_2))1_{[0,\infty)}(y_1)1_{[0,\infty)}(y_2)$$

and suppose $Z\cong H$ on $\mathbb{R}^+$. Then

$$f(y)\equiv f_{\theta,H}(y)=\int_0^\infty \theta z^2\exp(-z(y_1+\theta y_2))dH(z);$$

see e.g. LINDSAY [53]. Here θ is a parametric component and H a nonparametric component of the model, and the mixed distribution is parametric while the mixing distribution is nonparametric. Generalizations of this model, including regression type models, have been studied and advocated for use in modeling micro-economic data by HECKMAN and SINGER [35].

(b) Dependent proportional hazards or fraility models. Suppose that $(Y\equiv(Y_1,Y_2)\mid Z=z)$ has joint survival function

$$P_G(Y_1\geqslant y_1,Y_2\geqslant y_2\mid Z=z)=[1-G_1(y_1)]^z[1-G_2(y_2)]^z$$

with $G=(G_1,G_2)$ and suppose that $Z\cong\text{Gamma}(\nu,\lambda)$. Then with $\theta=(\nu,\lambda)$,

$$P_{\theta,G}(Y_1\geqslant y_1,Y_2\geqslant y_2)=\frac{\lambda^\nu}{[\lambda+\Lambda_1(y_1)+\Lambda_2(y_2)]^\nu}$$

where $\Lambda_i\equiv-\log(1-G_i)$, $i=1,2$. In this case the mixed distribution is nonparametric while the mixing distribution is a parametric family. This model, which serves as an alternative to (a), has been studied by CLAYTON [16] and OAKES [63], and has been generalized by GILL [28]. Related regression models are discussed by RIDDER and VERBAKEL [68] and ELBERS and RIDDER [25].

(c) Errors in variables models. Suppose that $X=(Y,Z)$ with

$$Y_1=Z+\epsilon_1$$

$$Y_2=\alpha+\beta Z+\epsilon_2$$

where $Z \cong H$ (non-Gaussian) and $\epsilon \equiv (\epsilon_1, \epsilon_2) \cong N(0, \Sigma)$. The resulting mixture model is an *errors in variables regression* model. Consistent maximum likelihood estimates were obtained by KIEFER and WOLFOWITZ [42], but lower bounds for estimation of (α, β) together with asymptotically efficient estimates attaining the bounds were first obtained by BICKEL and RITOV [9].

(d) If $(Y \mid Z = z) \cong$ exponential (z) and $Z \cong H$, then

$$P_H(Y \geqslant y) = \int_0^\infty \exp(-yz) dH(z).$$

Estimation of H via nonparametric maximum likelihood methods in this and more general situations has been considered by LAIRD [45] and JEWELL [39]. While the estimates are known to be consistent, little is known about the efficiency of the estimates or their rate of convergence.

Other results concerning mixing models and efficient estimation have also been obtained by LAMBERT and TIERNEY [46], [47], and by HAS'MINSKII and IBRAGIMOV [34].

2.2.3 Censoring models. (F). These models are derived from other models of one of the above types as follows: Suppose that $X \cong P_{\theta,G} \in \mathbf{P}$, and suppose that T is a many-to-one function on the sample space $\mathbf{X}$ of X. Then we can observe only $X^* \equiv T(X) \cong P^*_{\theta,G}$.

Examples:

(a) Mixing. The mixing models of E are censoring models with $X^* \equiv T(Y, Z) = Y$.

(b) Random right censorship. In this type of censoring, which has received much use in survival analysis, $X^* \equiv (X_1^*, X_2^*) \equiv T(X_1, X_2) \equiv (X_1 \wedge X_2, 1_{[X_1 \leqslant X_2]})$. Random right censoring meshes extremely well with Cox's proportional hazards regression model as discussed in D(e). On the other hand, however, this type of censoring can make estimation quite difficult. For example, estimation for the linear regression model D(b) with arbitrary right censoring of the dependent variable has been considered by MILLER [61] and by BUCKLEY and JAMES [13]; see HALPERN and MILLER [60]. RITOV [69] has, in spite of the difficulties, computed information lower bounds and produced asymptotically efficient estimators achieving the bounds. TIBSHIRANI [75] considered a version of this censored regression model with the linear (parametric) regression function replaced by a smooth regression function.

(c) Convolution. Here $X^* \equiv T(X_1, X_2) \equiv X_1 + X_2$ where X_1 and X_2 are independent. The traffic model of BRANSTON [10] is a model which results from this convolution type of censoring combined with a simple mixture model.

2.2.4 Biased sampling models. (G). Suppose that $X \cong P_{\theta,G} \in \mathbf{P}$, a semiparametric model. Then suppose that $K_i(x)$, $i = 1, \ldots, s$ is a collection of known non-negative *biasing kernels* and that λ_i, $i = 1, \ldots, s$ is a probability distribution on $\{1, \ldots, s\}$. Then the *biased sampling distribution* corresponding to $P_{\theta,G}$, $\underline{K} = (K_1, \ldots, K_s)$, and $\underline{\lambda} = (\lambda_1, \ldots, \lambda_s)$ is

$$P_{\theta,G,\lambda}(X \in A, I = i) = \frac{\int_A K_i(x) P_{\theta,G}(dx)}{\int_{\mathbf{X}} K_i(x) P_{\theta,G}(dx)} \lambda_i \tag{2}$$

for $i = 1, \ldots, s$. Here are some examples of this type of model.

Examples:

(a) Vardi's selection bias model. Suppose that $P_{\theta,G} = G$ and $K_1, \ldots, K_s$ are biasing functions with $\int K_i dG < \infty$ for $i = 1, \ldots, s$, and $\lambda_i \geq 0$ satisfy $\sum_{i=1}^{s} \lambda_i = 1$. Then

$$P_{G,\lambda}(X \in A, I = i) = \frac{\int_A K_i dG}{\int_{\mathbf{X}} K_i dG} \lambda_i, \quad i = 1, \ldots, s.$$

VARDI [78] gives a condition which guarantees existence of the non-parametric maximum likelihood estimate of G. The particular case with $\mathbf{X} = \mathbb{R}^1$, $K_1(x) = 1, K_2(x) = x$, which involves the length-biased distribution $\int_0^x y dG(y) / \mu$ corresponding to G was studied by VARDI [77], and the further special case with $\lambda_1 = 0 = 1 - \lambda_2$ was considered earlier by COX [21]. Consistency, asymptotic normality, and efficiency of Vardi's non-parametric maximum likelihood estimator are addressed in a forthcoming paper by GILL and WELLNER [29].

(b) Choice-based sampling models. Suppose that $X \equiv (\underline{Y}, Z)$, where $Z \cong H$ is a vector of covariates, and $(\underline{Y} \mid Z = z) \cong \text{Multinomial}_k(1, \underline{p}(\theta, z))$ (where k denotes the number of cells and the number of trials is 1); we will write $[Y = y]$ for the event that outcome y occurs, $y = 1, \ldots, k$. A frequently used model for the p's is the multinomial - logit model with

$$P_\theta(Y = y \mid Z = z) = p_y(\theta, z) = \frac{\exp(\theta_y z)}{\sum_{y'=1}^{k} \exp(\theta_{y'} z)},$$

but in any case this part of the model is parametric; the nonparametric part of the model is G. To get a 'choice-based sampling model', let $K_i(x) \equiv K_i(y, z) = 1_{D_i}(y)$, $i = 1, \ldots, s$ where $D_1, \ldots, D_s$ are known subsets of $\{1, \ldots, k\}$. Then the biased sampling model (2) becomes

$$P_{\theta,G}(Y = y, Z \in B, I = i) = \frac{\int_B 1_{D_i}(y) P_\theta(Y = y \mid Z = z) dG(z)}{\int \sum_{y=1}^{k} 1_{D_i}(y) P_\theta(Y = y \mid Z = z) dG(z)} \lambda_i$$

This type of model has received considerable use in econometrics; see COSSLETT [20] for some history and further references. Estimation for this model was considered by MANSKI and LERMAN (*Econometrika 45* (1977), 1977-1988). The efficiency of their estimators of θ and generalizations were treated by COSSLETT [20]. In general the 'choice functions' or biasing kernels may depend on both y and z; see MANSKI and MCFADDEN (*Structural Analysis of Discrete Data* (1981), MIT Press).

(c) Truncated regression models. Suppose that $X=(Y,Z)$ with $Y=\theta Z+\epsilon$ where $\epsilon\cong G$ with density g and $Z\cong H$ are independent. Thus the basic semiparametric model is a linear regression model with unknown error distribution G. If $s=1$ and $K(x)=K(y,z)=1_{(-\infty,y_0]}(y)$ where y_0 is a fixed constant, then

$$P_{\theta,G}(Y\in A,Z\in B)=\frac{\int_B\int_{(-\infty,y_0]\cap A}g(y-\theta z)dydH(z)}{\int\int_{(-\infty,y_0]}g(y-\theta z)dydH(z)}.$$

This truncated regression model has been investigated by BHATTACHARYA, CHERNOFF, and YANG [5]. Motivated by a controversy in astronomy concerning Hubble's law, they constructed $\sqrt{n}$-consistent estimators of the regression parameters θ. Further results for this model have been obtained by JEWELL [41], who also gives additional examples. JEWELL [40] has also considered estimation for generalizations of this model with $s\geqslant 2$ corresponding to stratified sampling on the dependent variable Y.

3. BOUNDS FOR ESTIMATION

Lower bounds for the variances of estimators play an important role in statistical theory, setting a baseline or standard against which estimators can be compared. In their classical form such bounds assert that *any* unbiased estimator θ_n of θ has variance no smaller than $(nI(\theta))^{-1}\equiv b(\theta)/n$:

$$\mathrm{Var}_\theta[\hat{\theta}_n]\geqslant\frac{b(\theta)}{n}$$

In other words $b(\theta)/n$ is the smallest variance we can hope for in an unbiased estimator $\hat{\theta}_n$ of θ. If $\hat{\theta}_n^b$ is an estimator which asymptotically *achieves* the bound (in the sense that $\sqrt{n}(\hat{\theta}_n^b-\theta)\to_d N(0,b(\theta))$), then we say that $\hat{\theta}_n^b$ is *asymptotically efficient.* If the statistician uses an estimator $\hat{\theta}_n^a$ which is inefficient, then he has not used the data to best advantage and is essentially wasting observations. Hence if $\hat{\theta}_n^a$ is another estimator with $\sqrt{n}(\hat{\theta}_n^a-\theta)\to_d N(0,a(\theta))$ where $a(\theta)\geqslant b(\theta)$ necessarily, then the limiting ratio of sample sizes which yields equal standard deviations (and hence also equal variances) of $\hat{\theta}_n^b$ and $\hat{\theta}_n^a$ is called the *asymptotic relative efficiency* $e_{a,b}$ of $\hat{\theta}_n^a$ with respect to $\hat{\theta}_n^b$; evidently $e_{a,b}=b(\theta)/a(\theta)\leqslant 1$. If the estimator $\hat{\theta}_n^a$ has asymptotic relative efficiency $1/2$ relative to an (efficient) estimator $\hat{\theta}_n^b$ and the estimator $\hat{\theta}_n^b$ requires $n_b=100$

observations to yield a given variance, then $n_a = 200$ observations will be needed to achieve the same variance using the inefficient estimator $\hat{\theta}_n^a$; half the data are 'wasted' by the use of $\hat{\theta}_n^a$. Thus in the search for 'good' estimators and other inference procedures, statisticians are interested in answers to the questions: A. How well can we do? What are the lower bounds for estimation in the model at hand? B. How can we construct efficient estimates, i.e. estimates which achieve the bounds?

Our aim in this section is to briefly survey classical (Cramér - Rao) and modern (Hájek - Le Cam) bounds for estimation in 'regular' parametric models. The Hájek - Le Cam approach has led to the development of lower bounds for estimation in nonparametric and semiparametric models. Bounds of this type have been established by BERAN [3], KOSHEVNIK and LEVIT [43], LEVIT [52], MILLAR [57], [58], [59], PFANZAGL [64], and BEGUN et al. [1]. We give a brief introduction to these bounds for semiparametric models at the end of this section. A thorough treatment will be given in the forthcoming monograph by BICKEL, KLAASSEN, RITOV, and WELLNER [7].

3.1. Cramér - Rao lower bounds

First consider the case of a 'regular' parametric model: suppose that $X_1, \ldots, X_n$ are iid $P_\theta \in \mathbf{P} \equiv \{P_\theta : \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^d$ is open, that $\mathbf{P}$ is dominated by a (sigma-finite) measure μ on $\mathbf{X}$, and let $p(\cdot, \theta) \equiv \frac{dP_\theta}{d\mu}$ for $\theta \in \Theta$. Then the classical *log-likelihood* of an observation X is

$$l(\theta, X) \equiv \log p(X, \theta),$$

the *scores vector* $\dot{l}$ is

$$\dot{l}(\theta, X) \equiv \nabla l(\theta, X) = \frac{1}{p(X,\theta)} \left(\frac{\partial}{\partial \theta_1} p(X, \theta), \ldots, \frac{\partial}{\partial \theta_d} p(X, \theta) \right)^\top,$$

and the *Fisher information matrix* for θ is

$$I(\theta) = E_\theta[\dot{l}(\theta, X)\dot{l}(\theta, X)^\top].$$

Assume that $I(\theta)$ is positive definite so that $I(\theta)^{-1}$ exists.

One form of the classical Crámer-Rao inequality for unbiased estimates $a^\top \hat{\theta}_n$ of $a^\top \theta$, where a is a fixed vector in $\mathbb{R}^d$, is:

$$n \mathrm{Var}_\theta[a^\top \hat{\theta}_n] \geq a^\top I(\theta)^{-1} a = \sup_{b \in \mathbb{R}^d} \frac{(a^\top b)^2}{b^\top I(\theta) b}. \tag{1}$$

If we focus on estimation of the first component $\theta_1 \in \mathbb{R}^1$ of θ, it follows immediately from (1), the definition of $I(\theta)$, and standard L_2-projection or

regression theory that

$$n\mathrm{Var}_\theta[\hat{\theta}_1] \geqslant \sup_{b \in \mathbf{R}^d} \frac{b_1^2}{b^{\mathsf{T}} I(\theta) b} \equiv I^{11}(\theta) \tag{2}$$

$$= \frac{1}{\inf_{c \in \mathbf{R}^d, c_1 = 1} E_\theta[\dot{l}_1 - c_2 \dot{l}_2 - \ldots - c_d \dot{l}_d]^2}$$

$$= \frac{1}{I_{11}(\theta) - I_{12}(\theta) I_{22}^{-1}(\theta) I_{21}(\theta)} \equiv \frac{1}{I_{11}^*(\theta)}$$

where

$$I(\theta) \equiv \begin{bmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{bmatrix}, \quad I(\theta)^{-1} = \begin{bmatrix} I^{11}(\theta) & I^{12}(\theta) \\ I^{21}(\theta) & I^{22}(\theta) \end{bmatrix}$$

denote the partitions of $I(\theta)$ and $I(\theta)^{-1}$ corresponding to the partition of $\theta = (\theta_1, \underline{\theta}_2^{\mathsf{T}})^{\mathsf{T}}$ with $\underline{\theta}_2 = (\theta_2, \ldots, \theta_d)^{\mathsf{T}}$. Thus when θ_1 is the parameter of interest and $\underline{\theta}_2 = (\theta_2, \ldots, \theta_d)^{\mathsf{T}}$ are nuisance parameters, the *effective information* $I_{11}^*(\theta)$ for θ_1 is

$$I_{11}^*(\theta) = I_{11} - I_{12} I_{22}^{-1} I_{21} = E_\theta(\dot{l}_1^{*2}), \tag{3}$$

where the *efficient score function* $\dot{l}^*$ for θ_1 is

$$\dot{l}_1^* \equiv \dot{l}_1 - I_{12} I_{22}^{-1} \underline{\dot{l}}_2 = \dot{l}_1 - \Pi(\dot{l}_1 \mid [\underline{\dot{l}}_2]) \tag{4}$$

and the *efficient influence curve* $\tilde{l}_1$ for estimation of θ_1 is

$$\tilde{l}_1 = I_{11}^*(\theta)^{-1} \dot{l}_1^*, \tag{5}$$

so that

$$E_\theta(\tilde{l}_1^2) = I_{11}^*(\theta)^{-1} = I^{11}(\theta).$$

It is easily seen that the effective information I_{11}^* for θ_1 is just the squared length of the component $\dot{l}_1^*$ of $\dot{l}_1$ which is orthogonal to $\dot{l}_2, \ldots, \dot{l}_d$ in $L_2(P_\theta)$: in other words, the efficient score function is obtained by subtracting from $\dot{l}_1$ its projection $\Pi(\dot{l}_1 \mid [\underline{\dot{l}}_2]) = I_{12} I_{22}^{-1} \underline{\dot{l}}_2$ on the space $[\underline{\dot{l}}_2]$ spanned by $\dot{l}_2, \ldots, \dot{l}_d$ in $L_2(P_\theta)$.

If the nuisance parameters $\underline{\theta}_2 = (\theta_2, \ldots, \theta_d)^{\mathsf{T}}$ are known, the bound (2) may be replaced by

$$n\mathrm{Var}_\theta[\hat{\theta}_1] \geqslant \frac{1}{I_{11}(\theta)}, \tag{6}$$

and, of course,

$$I_{11}(\theta) \geqslant I_{11}^*(\theta) = I_{11} - I_{12} I_{22}^{-1} I_{21}$$

where equality holds if and only if

$$I_{12} = I_{21}^{\mathsf{T}} = 0 \quad \text{or iff} \quad \dot{l}_1 \perp \dot{l}_2, \ldots, \dot{l}_d \text{ in } L_2(P_\theta). \tag{7}$$

Thus lack of knowledge of $\underline{\theta}_2 \equiv (\theta_2, \ldots, \theta_d)^{\mathsf{T}}$ decreases the information for θ_1 unless (7) holds; in this case the lower bounds (2) and (6) agree, suggesting that θ_1 can be estimated as well when $\underline{\theta}_2$ is unknown as when $\underline{\theta}_2$ is known. This possibility was recognized by STEIN [71] in a paper which initiated the theory of *adaptive estimation*.

3.2. Hájek - Le Cam lower bounds

Two different but closely related asymptotic formulations of the classical Cramér - Rao lower bounds have proved useful: One is the convolution-type representation theorem of HÁJEK [32] and LE CAM [48] which has been further developed and applied by BERAN [3], [4] and MILLAR [59]. The other is the local asymptotic minimax approach; see HÁJEK [33] for a nice exposition and history, MILLAR [58], and LE CAM [49] for additional remarks.

Both types of lower bounds are formulated in terms of *locally asymptotically normal families*: Suppose that $\underline{X} = (X_1, \ldots, X_n) \cong P_{n,\theta}$ has density $p_n(\cdot, \theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, and set

$$\mathbf{l}_n(\theta) = \log p_n(\underline{X}, \theta).$$

If $\theta_n \equiv \theta + hn^{-1/2}$, so that

$$\mathbf{l}_n(\theta_n) - \mathbf{l}_n(\theta) = \log[p_n(\underline{X}, \theta_n) / p_n(\underline{X}, \theta)],$$

then $\mathbf{P} \equiv \{P_{n,\theta} : \theta \in \Theta\}$ is *locally asymptotically normal* (LAN) at θ if there is a vector of $L_2(P_\theta)$ functions $\underline{\dot{l}}_n(\theta)$ and a nonsingular matrix $I(\theta)$ such that, with

$$\mathbf{l}_n(\theta_n) - \mathbf{l}_n(\theta) = \underline{\dot{l}}_n(\theta)^{\mathsf{T}} h - \tfrac{1}{2} h^{\mathsf{T}} I(\theta) h + \mathbb{R}_n(\theta, h), \tag{8}$$

it follows that, in $P_{n,\theta}$-probability,

(i) $R_n(\theta, h) \to_p 0$ uniformly on bounded h-sets, and

(ii) $\underline{\dot{l}}_n(\theta) \to_d N(0, I(\theta))$.

Thus $\mathbf{l}_n(\theta_n) - \mathbf{l}_n(\theta) \to_d N(-\frac{1}{2}\sigma^2, \sigma^2)$ with $\sigma^2 = h^{\mathsf{T}} I(\theta) h$. In 'regular families' $\mathbf{P}$ (with iid observations) $\underline{\dot{l}}_n(\theta) = n^{-1/2} \Sigma_{i=1}^n \underline{\dot{l}}(\theta, X_i)$ where $\underline{\dot{l}}$ is the scores vector (for $n = 1$) and $I(\theta)$ is the information matrix.

Because of our interest here in the parametric component θ of a semi-parametric model $\mathbf{P} = \{P_{\theta,G}\}$, we formulate versions of the convolution and asymptotic minimax bounds for the first component θ_1 of θ.

A sequence of estimators T_{1n} of θ_1 is *regular* at θ if, under P_{θ_n}

$$\sqrt{n}(T_{1n} - \theta_{1n}) \to_d T_1$$

for every $\theta_n = \theta + n^{-1/2} h$ where the distribution $\mathbf{L}(T_1)$ of T_1 does not depend on h.

THEOREM 1 (HÁJEK, 1970). *Suppose that* $\mathbf{P}$ *is LAN at* θ *and that* T_{1n} *is a regular estimator with limit distribution* $\mathbf{L}(T_1)$. *Then*

$$T_1 \cong Z_1 + W_1 \tag{9}$$

where $Z_1 \cong N(0, 1 / I_{11}^*(\theta))$, $I_{11}^*(\theta)$ *is as in (3), and* W_1 *is independent of* Z_1.

Thus any regular estimator T_{1n} of θ_1 must have a limit distribution which is at least as dispersed as $N(0, 1 / I_{11}^*(\theta))$, and it makes sense to call a regular estimator T_{1n} asymptotically efficient if it converges in distribution to Z_1; i.e. if $W_1 = 0$ in (9).

Now suppose that $w: \mathbb{R}^1 \to \mathbb{R}^+$ satisfies:

(i) $w(x) = w(-x)$ for all $x \in \mathbb{R}^1$;

(ii) $w(0) = 0$, $w(x)$ increases in $x \geqslant 0$;

(iii) $Ew(\sigma Z) < \infty$ for all $\sigma > 0$ where $Z \cong N(0,1)$.

THEOREM 2 (HÁJEK, 1972). *Suppose that* $\mathbf{P}$ *is LAN at* θ *and that* w *satisfies (i) - (iii). Then, for any estimator* T_{1n} *of* θ_1,

$$\lim_{M \to \infty} \liminf_{n \to \infty} \sup_{\sqrt{n}\,|\theta_n - \theta| \leqslant M} E_{\theta_n} w(\sqrt{n}(T_{1n} - \theta_{1n})) \geqslant Ew(Z_1) \tag{10}$$

where $Z_1 \cong N(0, 1 / I_{11}^*(\theta))$ *as in theorem 1.*

If the uniformity in h in (i) of the definition of a LAN family is relaxed to just pointwise convergence, then theorems 1 and 2 continue to hold, but the bounds may not be attainable. Furthermore, if attention is restricted to regular estimates, then (10) holds without the supremum on the lefthand side.

3.3. Bounds for semiparametric models

The Hájek-Le Cam convolution and asymptotic minimax bounds stated above for a parametric model $\mathbf{P}_0$ continue to hold in a wide range of regular nonparametric and semiparametric models. All of the extensions make use, in some form, of the *tangent space* $\dot{\mathbf{P}}$ (at (θ, G)) for the model $\mathbf{P}$. For a parametric model $\mathbf{P}_0$ the tangent space $\dot{\mathbf{P}}_0$ (at $\theta \in \Theta$) is just the linear subspace $[\dot{l}_1, \ldots, \dot{l}_d]$ of $L_2(P_\theta)$ spanned by $\dot{l}_1, \ldots, \dot{l}_d$. For a semiparametric model $\mathbf{P} = \{P_{\theta,G} : \theta \in \Theta \subset \mathbb{R}^d, G \in \mathbf{G}\}$, the tangent space $\dot{\mathbf{P}} \subset L_2(P_{\theta,G})$ is simply the set of all possible score functions of one-dimensional regular parametric submodels (at (θ, G)).

For $\theta_0 \in \Theta, G_0 \in \mathbf{G}$, let $\mathbf{P}_\theta$ and $\mathbf{P}_G$ denote the submodels of $\mathbf{P}$ with $G = G_0$ and $\theta = \theta_0$ respectively:

$$\mathbf{P}_\theta \equiv \{P_{\theta,G_0} \in \mathbf{P} : \theta \in \Theta\}, \quad \mathbf{P}_G \equiv \{P_{\theta_0,G} \in \mathbf{P} : G \in \mathbf{G}\}.$$

If $\dot{\mathbf{P}}_\theta$ and $\dot{\mathbf{P}}_G$ denote the corresponding tangent spaces, then $\dot{\mathbf{P}}_\theta \oplus \dot{\mathbf{P}}_G \subset \dot{\mathbf{P}}$ and

typically equality holds. Here $\dot{\mathbf{P}}_G$ plays the role that $[\dot{l}_2, \ldots, \dot{l}_d]$ played for the parametric model $\mathbf{P}_0$, and the *efficient score function* for θ extending (4) is:

$$\dot{l}_\theta^* = \dot{l}_\theta - \Pi(\dot{l}_\theta \mid \dot{\mathbf{P}}_G) \tag{11}$$

so that $\dot{l}_\theta^* \perp \dot{\mathbf{P}}_G$ in $L_2(P_{\theta,G})$, and the *effective information* for θ in the model $\mathbf{P}$ is

$$I^*(\theta) = E_{\theta,G}(\dot{l}_\theta^* \dot{l}_\theta^{*T}). \tag{12}$$

In the special case when $\dot{l}_\theta^* = \dot{l}_\theta \perp \dot{\mathbf{P}}_G$, then $I^*(\theta) = I(\theta) \equiv E_{\theta,G}(\dot{l}_\theta \dot{l}_\theta^T)$ and *adaptation to G* is possible; this is the situation emphasized by STEIN [71] and BICKEL [6].

Now versions of theorems 1 and 2 for the parametric component θ of the semiparametric model $\mathbf{P}$ continue to hold with θ_1 replaced by θ and $1 / I_{11}^*(\theta)$ replaced by $I^*(\theta)^{-1}$ where $I^*(\theta)$ is given in (12); see KOSHEVNIK and LEVIT [43], LEVIT [52], BEGUN et al. [1], and PFANZAGL [64], [65]. A complete treatment will be given in BICKEL, KLAASSEN, RITOV, and WELLNER [7].

4. CONSTRUCTION OF ASYMPTOTICALLY EFFICIENT ESTIMATES: TWO APPROACHES

Suppose that $\mathbf{P} = \{P_{\theta,G} : (\theta, G) \in \Theta \times \mathbf{G}\} \equiv \{P_\theta : \theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2\}$ with $\Theta_2 = G$ is a 'regular' semiparametric model. A first stage in analyzing the model is to calculate scores for θ and information lower bounds as outlined in Section 2 above if possible. A second step is to construct estimators $(\bar{\theta}_n, \bar{G}_n)$ which are $\sqrt{n}$-consistent. A third stage is to find estimators $(\hat{\theta}_n, \hat{G}_n)$ of (θ, G) which are efficient in the sense that they *achieve* the information lower bounds (perhaps in the weakened sense of convergence in distribution for fixed (θ, G) rather than locally uniformly as required by the definition of regular estimates given in Section 3).

Two classical methods of constructing asymptotically efficient estimators $\hat{\theta}_n$ in regular parametric models are the methods of maximum likelihood estimation and Bayes estimation; see LEHMANN [51] and IBRAGIMOV and HAS'MINSKII [38], though, as LEHMANN makes clear, the emphasis in likelihood estimation, even in parametric models, should be on the scores and score equations rather than on maximizing likelihoods per se since the scores often lead to efficient estimates even when likelihoods themselves are unbounded.

Our aim here is to outline two useful approaches to the construction of asymptotically efficient estimates of the parametric part θ of a semiparametric model $\mathbf{P}$.

4.1. Method 1: Efficient score equation

Suppose that it is possible to calculate the *efficient score function* $\dot{l}_1^*$ for θ_1,

$$\dot{l}_1^* = \dot{l}_1 - I_{12} I_{22}^{-1} \dot{l}_2 = \dot{l}_1 - \Pi(\dot{l}_1 \mid \dot{\mathbf{P}}_{\theta_2})$$

and the *effective information*

$$I_{11}^*(\theta) = E_\theta(\dot{l}_1^{*2}).$$

Furthermore, suppose that $\bar{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ, $\sqrt{n}(\bar{\theta}_n - \theta) = O_P(1)$. Then define $\hat{\theta}_{1n}$ to be either a solution of the *efficient score equation*

$$\sum_{i=1}^{n} \dot{l}_1^*(\hat{\theta}_{1n}, \bar{\theta}_{2n}, X_i) = 0,$$

or a one-step approximation thereof:

$$\hat{\theta}_{1n} = \bar{\theta}_{1n} + \frac{\frac{1}{n}\sum_{i=1}^{n} \dot{l}_1^*(\bar{\theta}_n, X_i)}{I_{11}^*(\bar{\theta}_n)} \tag{1}$$

$$= \bar{\theta}_{1n} + \frac{1}{n}\sum_{i=1}^{n} \tilde{l}_1(\bar{\theta}_n, X_i)$$

where $\tilde{l}_1$ is the efficient influence curve for θ_1, see (3.5). Additional smoothing may also be required in forming the sums in (1), but we have omitted it here for simplicity. Once an efficient estimator $\hat{\theta}_{1n}$ of θ_1 is found, method 2 can often be used to construct an efficient estimator of θ_2.

While no general theorem yet exists, the estimator $\hat{\theta}_{1n}$ defined above (or variations thereon involving suitable smoothing and truncation) has been shown to be asymptotically efficient in several important problems, a notable example being the errors in variables models studied by BICKEL and RITOV [9]. Roughly speaking, the fact that $\dot{l}_1^*$ is orthogonal to $\dot{l}_2, \ldots, \dot{l}_d$, the scores for θ_2, permits the use of an inefficient estimator $\bar{\theta}_{2n}$ to estimate out the 'nuisance parameter' θ_2. This should be contrasted with solving (or approximating by a one-step solution)

$$\sum_{i=1}^{n} \dot{l}_1(\theta_1, \bar{\theta}_{2n}) = 0$$

for θ_1, a method which is known to produce inefficient estimates of θ_1 in general; see e.g. GONG and SAMANIEGO [30].

The main drawback of the method is that it requires calculation of the efficient score function $\dot{l}_1^*$. Thus the method depends heavily on being able to calculate projections onto $[\dot{l}_2] = \dot{\mathbf{P}}_{\theta_2} = \dot{\mathbf{P}}_G$, which often necessitates calculation of the inverse of the information operator $\dot{l}_2^T \dot{l}_2 = I_{22}$. When $\dot{l}_1^* = \dot{l}_1$ so $\dot{l}_1$ is orthogonal to $[\dot{l}_2] = \dot{\mathbf{P}}_{\theta_2}$, then 'adaptation' with respect to $\theta_2 = G$ is possible, and method 1 becomes essentially the method used to construct efficient estimates in this case; see e.g. STONE [73] and BICKEL [6].

4.2. Method 2: Efficient estimation of θ_2 for known θ_1

Now suppose that an *efficient* estimate $\tilde{\theta}_{2n}$ of θ_2 is available if θ_1 is *known*. We denote this estimator by $\tilde{\theta}_{2n}(\theta_1)$ because it depends on the 'known' value of θ_1. Substitution of this estimate of θ_2 into the ordinary score for θ_1 (as if θ_2 were known and equal to $\tilde{\theta}_{2n}$ yields the 'condensed' or 'concentrated' score equation

$$\sum_{i=1}^{n} \dot{l}_1(\theta_1, \tilde{\theta}_{2n}(\theta_1), X_i) = 0$$

which we can solve for $\theta_1 \equiv \hat{\theta}_1$. Or, if $\bar{\theta}_{1n}$ is a $\sqrt{n}$-consistent estimate of θ_1, a one-step approximation thereof:

$$\hat{\theta}_{1n} = \bar{\theta}_{1n} + \frac{\frac{1}{n}\sum_{i=1}^{n} \dot{l}_1(\bar{\theta}_{1n}, \tilde{\theta}_{2n}(\bar{\theta}_{1n}), X_i)}{\frac{1}{n}\sum_{i=1}^{n} \dot{l}_1^2(\bar{\theta}_{1n}, \tilde{\theta}_{2n}(\bar{\theta}_{1n}))}; \tag{2}$$

as in the case of (1), more smoothing may be needed in forming the sums in (2), we have omitted it here for simplicity. This is a frequently used device in parametric models, but the method is equally useful for semiparametric models. While no general results concerning the estimator (2) seem to be known, this method has been used by RITOV [69] to construct efficient estimates for censored regression models.

5. PROBLEMS

Statisticians have a large, well-stocked tool-box for dealing with classical parametric models, and a growing companion set of tools for handling completely nonparametric models. The choice of tools for dealing with the rich middle ground of semiparametric models is, however, still relatively limited, and the few available tools are not all well suited for the job. Many important problems remain. Here is a partial list:

(a). *Calculation of lower bounds.* If the projection $\Pi(\dot{l}_\theta \mid \dot{\mathbf{P}}_G)$ in Section 3 can be calculated, then so can the efficient score function $\tilde{l}_\theta$, the effective information $I_{11}^*(\theta)$, and the efficient influence curve $\tilde{l}_1$. In many models this projection is simply a conditional expectation, and hence can be calculated easily; but in other models such as the dependent proportional hazards model of 2.E(b) the projection calculation is apparently intractable. More systematic methods, possibly involving iterative, numerical techniques, are needed.

(b). *Construction of efficient estimates.* HUANG [36] has made a preliminary study of method 1 outlined in Section 4, but general results concerning the asymptotic efficiency of methods 1 and 2, or variations thereof involving more smoothing, are still needed. Other methods including minimum Hellinger distance estimates, minimum Kullback-Leibler discrepancy estimators, and maximum-likelihood estimators obtained via EM-algorithms

all need further development and sharpening in the context of semiparametric models. Efficient estimates are still unknown for many of the models given in Section 2.

(c). *Identifiability and regularity criteria.* For many semiparametric models, further work on identifiability and conditions for regularity of submodels is still needed before work on estimation can get underway. For examples of such studies, see the papers by HECKMAN and SINGER [35] and ELBERS and RIDDER [25] concerning identifiability issues for the models of 2.E(b) and 2.E(c). Classical regularity investigations of translation and parametric models, which carry over to many group models are given by HÁJEK [31], [33].

(d). *Hypothesis testing.* As yet no adequate theory of hypothesis testing exists for semiparametric models. One type of testing problem concerns testing hypotheses within a nested family of semiparametric models: for example, consider testing $\Lambda_2 = \gamma\Lambda_1$ for some $0 < \gamma < \infty$ in the Clayton-Oakes model of example E(b). Or, of interest in survival analysis, test the assumption of a proportional hazards regression model against some general family of alternatives. Another rather different testing problem would involve testing non-nested semiparametric models against one another, e.g. a Cox-type regression model against a more classical linear regression model or perhaps a semiparametric mixed regression model.

(e). *Asymptotics for estimates based on smoothing.* Construction of efficient estimates for many of the models discussed above require smoothing techniques involving density or conditional expectation estimators. While the asymptotics for such smoothing processes are available, they need further development, study, and refinement to ease their systematic application to the construction of efficient estimates in a wide range of semiparametric models.

(f). *Robustness; connections and problems.* Efficient estimation in semiparametric models has many interesting connections with questions of robustness. Just as classical robustness theory has focused on neighborhoods of parametric models (often a one - sample location model), a generalization suggested by BICKEL and LEHMANN [8] concerns neighborhoods of semiparametric models, which they called 'nonparametric models with natural parameters'. For example, are the partial likelihood estimtors for the Cox proportional hazards model robust in some appropriate sense (with respect to the assumption of proportional hazards)? As more experience is gained with efficient estimates for semiparametric models, this more general type of robustness outlined by BICKEL and LEHMANN [8] can begin to be considered. Many challenging problems remain.

Acknowledgments: I have profited from several helpful discussions concerning semiparametric models with Peter Bickel. In particular, I learned of 'method 2' in Section 4 from him. I also owe thanks to Richard Gill for helpful comments concerning Sections 1 and 3. R.D. Martin suggested example 2D(f).

REFERENCES

1. J.M. BEGUN, W.J. HALL, W.M. HUANG, J.A. WELLNER (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist. 11*, 432 - 452.
2. S. BENNETT (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine 2*, 273-277.
3. R. BERAN (1977). Estimating a distribution function. *Ann. Statist. 5*, 400-404.
4. R. BERAN (1977). Robust location estimates. *Ann. Statist. 5*, 431-444.
5. P.K. BHATTACHARYA, H. CHERNOFF, S.S. YANG (1983). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist. 11*, 505-514.
6. P.J. BICKEL (1982). On adaptive estimation. *Ann. Statist. 10*, 647-671.
7. P.J. BICKEL, C.A.J. KLAASSEN, Y. RITOV, J.A. WELLNER (1986). *Efficient and Adaptive Inference in Semiparametric Models*, forthcoming monograph, Johns Hopkins University Press, Baltimore.
8. P.J. BICKEL, E.L. LEHMANN (1975). Descriptive statistics for nonparametric models. I. Introduction. *Ann. Statist. 3*, 1038-1044.
9. P.J. BICKEL, Y. RITOV (1984). *Efficient Estimation in the Errors in Variables Model*, preprint, Dept. of Statistics, University of California, Berkeley.
10. D. BRANSTON (1976). Models of single lane time headway distributions. *Transportation Science 10*, 125-148.
11. L. BREIMAN, J. FRIEDMAN (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc. 80*, 580-619 (with discussion).
12. N.E. BRESLOW, N.E. DAY (1980). *The Analysis of Case-Control Studies*, International Agency for Research on Cancer, Lyon.
13. J. BUCKLEY, I. JAMES (1979). Linear regression with censored data. *Biometrika 66*, 429-436.
14. R.J. CARROLL (1984). *Adaptation for the Slope in Simple Logistic Regression with an Intercept in the Structural Errors in Variables Model*, preprint.
15. R.J. CARROLL (1984). *A General Technique for Computing Information Bounds in Errors in Variables Structural Models*, preprint.
16. D. CLAYTON (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*, 141-151.
17. D. CLAYTON, J. CUZICK (1985). Multivariate generalizations of the proportional hazards model. *J. Roy. Statist. Soc. Ser. A. 148*, 82-117 (with discussion).
18. D. CLAYTON, J. CUZICK (1985). *An Approach to Inference for Rank-Regression Models with Right-Censored Data*, preprint.

19. D. CLAYTON, J. CUZICK (1985). The semi-parametric Pareto model for regression analysis of survival times. *Bull. Int. Stat. Inst. 51, part 4,* 23.3.1 - 23.3.18
20. S.R. COSSLETT (1981). Maximum likelihood estimation for choice-based samples. *Econometrica 49,* 1289-1316.
21. D.R. COX (1969). Some sampling problems in technology. N.L. JOHNSON, H. SMITH, JR. (eds.). *New Developments in Survey Sampling,* 506-527, Wiley-Interscience, New York.
22. D.R. COX (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B 34,* 187-220.
23. D.R. COX, E.J. SNELL (1981). *Applied Statistics: Principles and Examples,* Chapman and Hall, London.
24. K. DOKSUM (1985). *Partial Likelihood Methods in Transformation Models,* preprint, University of California, Berkeley.
25. C. ELBERS, G. RIDDER (1983). True and spurious duration dependence: the identifiability of the proportional hazards model. *Review of Economic Studies 49,* 403-410.
26. R.F. ENGLE, C.W.J. GRANGER, J. RICE, A. WEISS (1983). *Nonparametric Estimates of the Relation between Weather and Electricity Demand,* preprint, Department of Economics, University of California, San Diego.
27. J.H. FRIEDMAN, W. STUETZLE (1981). Projection pursuit regression. *J. Amer. Statist. Assoc. 76,* 817-823.
28. R.D. GILL (1984). *Models for the Censored Data Matched Pairs Problem,* preprint, Centrum voor Wiskunde en Informatica, Amsterdam.
29. R.D. GILL, J.A. WELLNER (1985). *Limit Theorems for Empirical Distributions in Selection Bias Models,* preprint, Dept. of Statistics, University of Washington.
30. G. GONG, F.J. SAMANIEGO (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist. 9,* 861-869.
31. J. HAJEK (1962). Asymptotically most powerful rank order tests. *Ann. Math. Statist. 33,* 1124-1147.
32. J. HAJEK (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. verw. Gebiete 14,* 323-330.
33. J. HAJEK (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berk. Symp. Math. Statist. Prob. 1,* 175-194, University of California Press, Berkeley, California.
34. R.Z. HAS'MINSKII, I.A. IBRAGIMOV (1983). On asymptotic efficiency in the presence of an infinite dimensional nuisance parameter. K. ITO, J.V. PROKHOROV (eds.). *Probability Theory and Mathematical Statistics, Fourth USSR - Japan Symposium, Lecture Notes in Mathematics, 1021,* 95-229, Springer - Verlag, Berlin.
35. J. HECKMAN, B. SINGER (1984). A method for minimizing the impact of distributional assumptions in economic studies for duration data. *Econometrica 52,* 271-320.

36. W. HUANG (1984). *On Effective Score Estimation in Semiparametric Models*, preprint.
37. P.J. HUBER (1985). Projection pursuit. *Ann. Statist. 13*, 435-525 (with discussion).
38. I.A. IBRAGIMOV, R.Z. HAS'MINSKII (1981). *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.
39. N.P. JEWELL (1982). Mixtures of exponential distributions. *Ann. Statist. 10*, 479-484.
40. N.P. JEWELL (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika 72*, 11-21.
41. N.P. JEWELL (1985). *Least Squares Estimation of the Slope of a Truncated Regression*, preprint, Department of Biostatistics, University of California, Berkeley.
42. J. KIEFER, J. WOLFOWITZ (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist. 27*, 887-906.
43. YU.A. KOSHEVNIK, B.YA. LEVIT (1976). On a nonparametric analogue of the information matrix. *Theor. Prob. Appl. 21*, 738-753.
44. J. LAFLER, T.D. KINMAN (1965). An RR Lyrae star survey with the Lick 20 - inch astrograph II. The calculation of RR Lyrae periods by the electronic computer. *Astrophysical J., Suppl. 11*, 216-222.
45. N. LAIRD (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc. 73*, 805-811.
46. D. LAMBERT, L. TIERNEY (1984). Asymptotic efficiency of estimators of functionals of mixed distributions. *Ann. Statist. 12*, 1380-1387.
47. D. LAMBERT, L. TIERNEY (1984). Asymptotic properties of maximum likelihood estimates in the mixed Poisson model. *Ann. Statist. 12*, 1388-1399.
48. L. LE CAM (1972). Limits of experiments. *Proc. Sixth Berkeley Symp. Math. Statist. and Prob. 1*, 245-261, University of California Press, Berkeley, California.
49. L. LE CAM (1984). Review of Ibragimov and Has'minskii (1981) and Pfanzagl (1982). *Bull. (New Series) Amer. Math. Soc. 11*, 391-400.
50. E.L. LEHMANN (1953). The power of rank tests. *Ann. Math. Statist. 24*, 23-43.
51. E.L. LEHMANN (1983). *Theory of Point Estimation*, Wiley, New York.
52. B.YA. LEVIT (1978). Infinite-dimensional informational lower bounds. *Theor. Prob. Applic. 20*, 723-740.
53. B. LINDSAY (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London Ser. A 296*, 39-665.
54. B. LINDSAY (1983) The geometry of mixture likelihoods, part I. *Ann. Statist. 11*, 86-94.
55. B. LINDSAY (1983). The geometry of mixture likelihoods, part II. *Ann. Statist. 11*, 783-792.

56. J. McDonald (1983). Periodic smoothing of time series. *Project Orion Technical Report 017*, Department of Statistics, Stanford University, Stanford, California.
57. P.W. Millar (1979). Asymptotic minimax theorems for the sample distribution. *Z. Wahrsch. verw. Gebiete 48*, 233-252.
58. P.W. Millar (1983). The minimax principle in asymptotic statistical theory, *Proc. Ecole d'Ete St. Flour, Lecture Notes in Math. 976*, 75-265, Springer - Verlag, Berlin.
59. P.W. Millar (1985). Nonparametric applications of an infinite dimensional convolution theorem. *Z. Wahrsch. verw. Gebiete 68*, 545-556.
60. R. Miller, J. Halpern (1982). Regression with censored data. *Biometrika 69*, 521-531.
61. R.G. Miller (1976). Least squares regression with censored data. *Biometrika 63*, 449-464.
62. D. Oakes (1982). A model for association in bivariate survival data. *J. Roy. Statist. Soc. 44, Ser. B*, 412-422.
63. D. Oakes (1985). *Semiparametric Estimation in a Model for Association in Bivariate Survival Data*, preprint, Dept. of Statistics, University of Rochester (to appear in *Biometrika*)
64. J.Pfanzagl (1982). *Contributions to a General Asymptotic Statistical Theory, Lecture Notes in Statistics 13*, Springer - Verlag, New York.
65. J. Pfanzagl (1984). *A Remark on Semiparametric Models*, preprint, University of Cologne.
66. R.L. Prentice, R. Pyke (1979). Logistic disease incidence models and case-control studies. *Biometrika 66*, 403-411.
67. R. Prentice, S. Self (1983). Asymptotic distribution theory for Cox-type regression models with general risk form. *Ann. Statist. 11*, 804-813.
68. G. Ridder, W. Verbakel (1983). *On the Estimation of the Proportional Hazards model in the Presence of Unobserved Heterogeneity*, preprint. (to appear in *J. Numer. Statist. Assoc.*)
69. Y. Ritov (1984). *Efficient and Unbiased Estimation in Nonparametric Linear Regression with Censored Data*, preprint, Department of Statistics, University of California, Berkeley.
70. A. Schick. (1984). *On adaptive estimation*, preprint.
71. C. Stein (1965). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob. 1*, 187-195, University of California Press, Berkeley, California.
72. R.F. Stellingwerf (1978). Period determination using phase dispersion minimization. *Astrophysical J. 224*, 953-960.
73. C.J. Stone (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist. 3*, 267-284.
74. C.J. Stone (1985). Additive regression and other nonparametric models. *Ann. Statist. 33*, 689-705.

75. R. TIBSHIRANI (1982). Censored data regression with projection pursuit. *Project Orion Technical Report 013*, Department of Statistics, Stanford University.
76. R. TIBSHIRANI (1983). Non-parametric estimation of relative risk. *Project Orion Technical Report 022*, Department of Statistics, Stanford University.
77. Y. VARDI (1983). Nonparametric estimation in the presence of length bias. *Ann. Statist. 10*, 616-620.
78. Y. VARDI (1985). Empirical distributions in selection bias models. *Ann. Statist. 13*, 178-203.
79. G. WAHBA (1984). Partial spline models for the semiparametric estimation of functions of several variables. *Seminar Proceedings, Japan - USSR Joint Seminar on the Statistical Analysis of Time Series.*