Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

ESTIMATING FUNCTIONS: NONPARAMETRICS AND ROBUSTNESS

Pranab K. Sen
University of North Carolina at Chapel Hill

ABSTRACT

In nonparametric and robust inference, estimating functions are based on suitable implicitly or explicitly defined statistical functionals. The interplay of robustness and asymptotic efficiency properties of such typically nonlinear estimators is appraised here by reference to some standard as well as nonstandard problems that arise in statistical applications.

Key words : Adaptive estimation; alignment principle; asymptotic optimality; conditional functionals; estimable parameters; GEE; GLM; Hadamard differentiability; influence functions; L-, M-, R- and WLS estimators; statistical functional: trimmed mean; U-process; U-statistics.

1 Introduction

In parametrics, estimable parameters generally appear as algebraic constants associated with the underlying distribution function(s) of assumed functional form(s). In nonparametrics, estimable parameters are defined as functionals of the underlying distribution(s) that may not have known functional form(s). This formulation shifts emphasis to validity for a broad class of distributions wherein efficiency and robustness properties dominate the scenario. The *U-statistics* are the precursors of such nonparametric estimators; they are of the *kernel estimator* type, and enjoy good efficiency (and unbiasedness) properties but may not be generally very robust. Moreover not all parameters in a nonparametric setup are *estimable* or *regular functionals* in the Hoeffding (1948) sense; the median or a percentile of a distribution belonging to a broad class is a classical example of such a nonregular functional. Significant developments in nonparametric and robust estimation theory covering both kernel type and *estimating equation* (EE) type estimating functions (EF) have taken place in the recent past. Three important classes of estimators are the following

(i) *L-estimators* based on linear functions of order statistics,

(ii) *M-estimators* allied to the maximum likelihood estimators (MLE),

and (iii) *R-estimators* based on suitable rank statistics.
There has also been a drive to unify these estimators in terms of *differentiable statistical functionals*, though that fails to cover all such estimators. A good deal of discussion of such estimators with due attention to their robustness and asymptotic properties has appeared in various contemporary research publications; we may refer to Jurečková and Sen (1996) for an up to date treatise of this subject matter. Whereas L-estimators are explicitly defined statistical functionals, M- and R-estimators are defined implicitly (as is generally the case with the parametric MLE). *Conditional statistical functionals* particularly arising in *mixed-effects* and *multivariate models* have added new dimensions to the scope of study of (asymptotic) properties of such nonparametric estimators. In this quest, even the very linearity of the model has been challenged (on the grounds of validity and robustness), and hence, *nonparametric regression functions* have emerged in a better footing than before. Yet, the study of their robustness properties needs further scrutiny with adequate emphasis on their finite sample behavior. In the current study, due emphasis will be placed on such estimating function(al)s.

The basic motivation for estimating functions in some simple nonparametric models is presented in Section 2. We examine the picture in a more general (linear model) setup in Section 3. Some nonstandard models arising in functional estimation problems are introduced in Section 4. The concluding section deals with some general remarks.

2 Nonparametric Estimating Functions

We may remark that the very formulation of the *Fisher-consistency* criterion of statistical estimators brings the relevance of functionals of emprical measures in estimation theory. In a simple setup, given n independent and identically distributed random variables (i.i.d.r.v.) $X_1, \ldots, X_n$ from a distribution function (d.f.) F, we may introduce a statistical parameter $\theta = \theta(F)$ as a functional of F, and a natural (i.e., plug-in) estimator of this parameter is the corresponding sample counterpart

$$T_n = T(X_1, \ldots, X_n) = \theta(F_n), \tag{2.1}$$

where F_n is the sample (*empirical*) d.f. The empirical d.f. F_n is known to possess some optimality properties in a traditional setup, and if $\theta(F)$ is a linear functional, these properties are shared by T_n as well. The ingenuity of Hoeffding (1948) lies in covering a more general class of functionals where $\theta(F)$ can be expressed as

$$\begin{aligned} \theta(F) &= E_F\{g(X_1, \ldots, X_m)\} \\ &= \int \cdots \int g(x_1, \ldots, x_m) dF(x_1) \cdots dF(x_m), \end{aligned} \tag{2.2}$$

where $g(.)$ is a *kernel* of (finite) *degree* $m(\geq 1)$, and without loss of generality, we assume that $g(.)$ is a symmetric function of its m arguments. Hoeffding introduced the *U-statistic* as an (unbiased) estimator of $\theta(F)$:

$$U_n = \binom{n}{m}^{-1} \sum_{\{1 \leq i_1 < \cdots < i_m \leq n\}} g(X_{i_1}, \ldots, X_{i_m}), \; n \geq m. \tag{2.3}$$

A closely related estimator is the von Mises (1947) functional

$$\begin{aligned} V_n &= \theta(F_n) = \int \cdots \int g(x_1, \ldots, x_m) dF_n(x_1) \cdots dF_n(x_m) \\ &= n^{-m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} g(X_{i_1}, \ldots, X_{i_m}). \end{aligned} \tag{2.4}$$

For $m = 1$, U_n and V_n are the same (and average of i.i.d.r.v.'s). But for $m \geq 2$, they are generally not the same; whereas U_n is an unbiased estimator, V_n is generally not so. Nevertheless, under quite general regularity conditions, $|V_n - U_n| = O_p(n^{-1})$, so that asymptotically they share the same optimality properties, studied in detail by various researchers; Sen (1981) contains a systematic account of this work.

We may note that the above formulation is pivoted to a suitable kernel $g(.)$, and the optimality properties, interpreted in a nonparametric fashion, rest on the adoption of the classical Rao-Blackwell theorem along with *sufficiency* and *completeness* of sample order statistics. Nevertheless, from a robustness perspective the choice of the kernel is very important. Generally, if $g(.)$ is unbounded, neither of these two estimators may be very robust. To illustrate this feature, let us consider the simple situation where $\theta(F)$ is the variance of the d.f. F. In this case, the kernel $g(.)$ is given by

$$g(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2, \; m = 2, \tag{2.5}$$

so that if F does not have a compact support, the estimators are vulnerable to error contamination, gross errors or outliers. A similar situation arises with the mean functional (where $m = 1$ and $g(x) = x$), although there it may be possible to introduce the location parameter by imposing symmetry of F around its median and bypassing some of these technical difficulties. The degree of nonrobustness is likely to be more with dispersion than location measures. Robust estimation of location (regression) and scale parameters has its genesis in this feature, and we will discuss this briefly later on. There has been some attempts to introduce more robust dispersion functionals, such as the *mean absolute deviation* and *interquartile range*, although they may not belong to the class of estimable parameters in the sense of Hoeffding

(1948) and may also lack some generality prevailing in the case of location parameters; we refer to Jurečková and Sen (1996) for some discussion.

We may also consider a variant of U- or V-statistics which merits consideration on the ground of robustness. Instead of taking an average over all possible subsample kernels $g(X_{i_1}, \ldots, X_{i_m})$, $1 \leq i_1 < \cdots < i_m \leq n$, we arrange them in an ascending order, and denote the median of these $\binom{n}{m}$ pseudovalues by $\tilde{\theta}_n$; this can be proposed as an estimator of the median of the distribution of the kernel $g(X_1, \ldots, X_m)$. Thus, whenever for this kernel distribution, the median and mean are the same, a case that holds when this distribution is symmetric, or they differ by a known constant, a robust estimator can be obtained in this manner. In general, we may define a U-process by letting

$$U_n(y) = \binom{n}{m}^{-1} \sum_{\{1 \leq i_1 < \ldots < i_m \leq n\}} I(g(X_{i_1}, \ldots, X_{i_m}) \leq y), \quad y \in \mathbf{R}. \tag{2.6}$$

Virtually what has been studied for U-statistics remains true for such U-processes, and they have better robustness perspectives. Thus (Hadamard differentiable) statistical functionals of such U-processes may be advocated on the ground of robustness as well as other optimality properties. This line of attack initiated the development of the U-processes; in a more abstract setting, Nolan and Pollard (1987), and others, have studied related asymptotics. We would not go into further depths on this topic here. There are other types of U-processes, more akin to sequential setups, and we refer to Sen (1981) for some broad coverage of them.

For both location and dispersion measures, linear functions of order statistics, known as *L-statistics*, have been used extensively in the literature; they are generally efficient, adaptable in parametric as well as nonparametric setups, and generally possess good robustness properties. The estimating equation for L-estimators of location and scale parameters in (parametric) location-scale family of distributions has its genesis in the theory of (BLUE) *best linear unbiased estimators* which incorporates the *weighted least squares* (WLS) methodology on the set of order statistics. Led by this genesis, in a nonparametric setup, a parameter $\theta(F)$ is expressed as a functional

$$\theta(F) = \int_{-\infty}^{\infty} g(x) J(F(x)) dF(x), \tag{2.7}$$

where $g(.)$ is real valued, and $J = \{J(u), u \in (0,1)\}$ is a weight-function defined on the unit interval $(0,1)$. It is easy to see that the corresponding sample counterpart $\theta(F_n)$, given by

$$n^{-1} \sum_{i=1}^{n} g(X_{n:i}) J_n(i/n), \tag{2.8}$$

is an L-estimator; here the $X_{n:i}, i = 1, \ldots, n$ stand for the order statistics, and $J_n(.)$ is a suitable version converging to $J(.)$ (as n increases) almost everwhere. The flexibility of this approach stems from the choice of $g(.)$ and $J(.)$, for which $\theta(F)$ remains invariant, in such a way that retaining robustness to a greater extent, not much is compromized on (asymptotic) efficiency of an estimator of $\theta(F_n)$. For example, in the location model, the *trimmed mean*, a member of this class of L-estimators, for a small amount of trimming (at both ends) combines robustness with good efficiency properties. The theory of (asymptoptically) ABLUE is geared to this direction; it covers both the cases of smooth weight functions and a combination of a selected number of order statistics. For this location model, whenever F is assumed to be symmetric about its median (θ), one may take $g(x) = x$ a.e., and for nonnegative $J_n(i/n)$ (adding upto 1), we have a convex combination of the order statistics as an estimator of θ. Within this class, one may like to choose the weight-function in such a way that robustness can be fruitfully combined with high efficiency properties. Generally smoother weight functions are used in this context. For the location-scale problems, the sample quantiles and interquartile range are particular cases of such L-estimators. It is often possible to express an L-statistic as a U-statistic, and in an asymptotic setup, a first order approximation for L-estimators in terms of U-statistics works out well [viz., Sen (1981, ch.7)]. A notable example in this context is the *rank-weighted mean* $T_{n,k}$ which is the average of all the subsample medians of size $2k + 1$ from the given sample of size n. As in Sen (1964), we may write this equivalently as

$$T_{n,k} = \binom{n}{2k+1}^{-1} \sum_{i=1}^{n} \binom{i-1}{k} \binom{n-i}{k} X_{n:i}, \; k \geq 0; \tag{2.9}$$

so that for $k = 0$, we have the sample mean, and for $k = [(n-1)/2]$, we have the sample median. Incidentally, this example, for a $k \geq 1$, provides an illustration for the robustness of U-statistics even when the kernel is possibly unbounded.

In order to introduce the salient features of the estimating functions for R- and M-estimators it may be more convenient to start with the conventional MLE when the assumed (location) model is not necessarily the true one. Assume that $X_1, \ldots, X_n$ are drawn from a population with an absolutely continuous density function $f(x-\theta)$ where $f(.)$ is symmetric about the origin. Assume further that the true density function is given by $g(x-\theta)$ where $g(.)$ is also absolutely continuous and symmetric about the origin. Then the estimating function for obtaining the MLE $\tilde{\theta}_n$ based on the assumed model is given by

$$\sum_{i=1}^{n} \{-f'(X_i - \theta)/f(X_i - \theta)\} = 0. \tag{2.10}$$

Note that under the assumed conditions (on f and g),

$$\int \{-f'(x-\theta)/f(x-\theta)\}g(x-\theta)dx = 0, \tag{2.11}$$

so that the MLE based on the assumed model remains pertinent to the entire class of g satisfying the above symmetry condition. Let us denote by

$$\begin{aligned} A^2(f,g) &= \int \{-f'(x)/f(x)\}^2 g(x)dx, \\ \gamma(f,g) &= \int \{-f'(x)/f(x)\}g'(x)dx, \\ I(g) &= \int \{-g'(x)/g(x)\}^2 g(x)dx. \end{aligned} \tag{2.12}$$

Then following standard asymptotics for the MLE, it can be shown that

$$n^{1/2}(\tilde{\theta}_n - \theta) \to_{\mathcal{D}} \mathcal{N}(0,\ A^2(f,g)/\gamma^2(f,g)\ ). \tag{2.13}$$

Let us denote the MLE based on the true model by $\hat{\theta}_n$. Then we have the following result:

$$n^{1/2}(\hat{\theta}_n - \theta) \to_{\mathcal{D}} \mathcal{N}(0,\ [I(g)]^{-1}\ ). \tag{2.14}$$

Next note that by the Cauchy-Schwarz inequality,

$$\gamma^2(f,g) \le A^2(f,g)I(g), \tag{2.15}$$

where the equality sign holds only when $\{-f'(x)/f(x)\} \equiv \{-g'(x)/g(x)\}$ almost everywhere (a.e.). Thus, if both f and g belong to the same location-scale family of densities for which the log-derivative *scale-equivariant*, as is the case when g is Laplace or normal, then $\hat{\theta}_n$ and $\tilde{\theta}_n$ are isomorphic, and there is no loss of efficiency due to incorrect model assumption; this is the usual *parametric orthogonality* condition referred to in the literature. This orthogonality condition is not universal for the location-scale family of densities; the Cauchy density is a classical example toward this point. On the other hand, if f and g do not satisfy this condition, the *asymptotic relative efficiency* (ARE) of $\tilde{\theta}_n$ with respect to $\hat{\theta}_n$ is given by

$$e(f,g) = \gamma^2(f,g)/\{I(g)A^2(f,g)\}\ (\le 1). \tag{2.16}$$

This ARE can be quite low depending on the divergence of f and g. For example if g is Cauchy while f is taken to be normal, $A^2(f,g) = \infty$, and hence, $e(f,g) = 0$. In the above development, we have tacitly assumed that (2.10) holds. Sans the assumed symmetry of f and g this may not be generally true, and therefore in such a case, the MLE $\tilde{\theta}_n$ may have serious bias, and this in turn may make it inconsistent too. In any case, it is clear

that with an incorrect model, the derived MLE can not attain the Cramér-Rao information bound for its asymptotic mean square error, and hence, loses its (asymptotic) optimality properties.

The above picture turns out to be far more complex in a general parametric model where θ may not be the location parameter or the density may not be symmetric, and as such it reveals the grave nonrobustness aspects of the classical MLE to plausible model departures. The nonparametric situation is more complex in the sense that f and g may not simply differ in nuisance parameters, and their separability is generally defined in terms of more general metrics. As such, the methodology developed for parametric EF in the presence of nuisance parameters may not be of much use in this more general setup. In the classical *robust inference* setup, Huber (1964) introduced various measures of departures from the assumed model, such as the *Levi-distance*, *Kolmogorov-distance* and *Prokhorov-distance*, and has exhibited the possible lack of robustness of the classical MLE. Following his ground-breraking work, we may therefore conceive of a suitable *score function* $\psi(t), t \in \mathcal{R}$, and for the location model, consider the estimating equation:

$$\sum_{i=1}^{n} \psi(X_1 - \theta) = 0; \tag{2.17}$$

to have good robustness properties of the derived (M-)estimator, generally the *influence function* ψ is taken to be bounded a.e. In order that the above EE provides a consistent solution, we need that

$$\int \psi(x)g(x)dx = 0, \tag{2.18}$$

and a sufficient condition for this is the skew-symmetry of ψ and symmetry of g (both about 0). Within this broad class, specific choice of ψ can be made to achieve local efficiency, and we may refer to Hampel et al. (1986) and Jurečková and Sen (1996) for details.

The EF's for R-estimators have a greater appeal from *global robustness* perspectives. It stems from the basic fact that under suitable hypotheses of invariance, a rank statistic is genuinely distribution-free, so that whenever it has some monotinicity properties with respect to an *alignment* in the direction of alternative hypotheses, we have a robust EF. For example in the location model, assuming that the underlying d.f. G is symmetric, and incorporating suitable scores $a_n(1) \leq \cdots \leq a_n(n)$, we may consider a signed rank statistic

$$\sum_{i=1}^{n} \text{sign}(X_i) a_n(R_{ni}^+) = S_n \text{ say}, \tag{2.19}$$

where R_{ni}^+ is the rank of $|X_i|$ among $|X_1|, \ldots, |X_n|$, for $i = 1, \ldots, n$. Note that under the null hypothesis that θ is null, S_n has a known, symmetric

distribution. Moreover, if we replace the X_i by $X_i - a$, for some real a, and denote the corresponding (aligned) signed rank statistic by $S_n(a)$, then it is easy to verify that

$$S_n(a) \text{ is nonincreasing in } a \in \mathcal{R}. \tag{2.20}$$

Therefore, the EF in this case is $S_n(a)$, and the corresponding EE is

$$S_n(a) \text{ " = " } 0, \tag{2.21}$$

where, in view of the usual step-function nature of $S_n(a)$, " = " is defined precisely as follows. We let

$$\begin{aligned} \hat{\theta}_{n,1} &= \sup\{a : S_n(a) > 0\}, \quad \hat{\theta}_{n,2} = \inf\{a : S_n(a) < 0\}; \\ \hat{\theta}_n &= \frac{1}{2}[\hat{\theta}_{n,1} + \hat{\theta}_{n,2}]. \end{aligned} \tag{2.22}$$

Although in the case of the sign statistic $\hat{\theta}_n$ turns out to be the sample median, and for the Wilcoxon signed rank statistic, it is the median of the mid-ranges, in general, an algebraic expression may not be available, and an iterative solution has to be prescribed. Such R-estimators are distribution-free in the sense that they remain valid for the entire class of symmteric distributions, and unlike the case of the MLE, here the absolute continuity of the density function or a finite Fisher information need not be a part of the regularity assumptions.

Let us have a close look into the ranks R_{ni}^+, which assume the integer values $1, \ldots, n$. Thus, if an observation is moved to the extreme right (or left), it continues to have the rank n (or 1), no matter howfar it is shifted. In that way, the rank scores have good robustness properties for error contaminations and outliers. R-estimators are *translation-invariant*, robust, consistent and *median-unbiased* under very general regularity assumptions. Their asymptotic properties have been extensively studied in the literature; see for example, Jurečková and Sen (1996). Under appropriate regularity conditions, we have

$$n^{1/2}(\hat{\theta}_n - \theta) \rightarrow_{\mathcal{D}} \mathcal{N}(0, \nu^2), \tag{2.23}$$

where

$$\nu^2 = \{\int \phi^2(u)du\}/\{\int \phi(u)\{-g'(G^{-1}(u))/g(G^{-1}(u))\}du\}^2, \tag{2.24}$$

and $\phi(.)$ is the score generating function for the $a_n(k)$. As such, we may conclude that the nonparametric and robustness aspects of R-estimators prevail as long as the score functyion $\phi(.)$ is square integrable inside $(0, 1)$; unlike

the case of M-estimators, here we need not confine ourselves to bounded score functions. Within this broad class of score functions, one may choose specific members such that the ARE at a given G is a maximum, so that robustness can also be combined with local optimality.

In passing we may remark that for both M- and R-estimators, the d.f. G is largely treated as a nuisance (functional) parameter, so that the situation differs drastically from the parametric situation where one has generally a finite (and typically small) number of nuisance parameters. Further, looking at the last equation, we gather that

$$\nu^2 \geq [I(g)]^{-1}, \ \forall \phi \in \mathcal{L}_2(0,1), \tag{2.25}$$

where the equality sign holds only when $\phi(u) \equiv \{-g'(G^{-1}(u))/g(G^{-1}(u))\}$, $u \in (0,1)$. In this way, the situation is quite comparable to the case of M-estimators (sans the boundedness condition). As a matter of fact, both M- and R-estimators have certain asymptotic equivalence properties (with congruent score functions), and such general equivalence results for L-, M- and R-estimators have been studied in detail in Jurečková and Sen (1996). It follows from their general discussion that all of them are expressible in terms of *statistical functionals*; L-estimators being defined explicitly, while M- and R-estimators implicitly. For such statistical functionals, suitable modes of *differentiability* have been incorporated to provide convenient means for the study of general asymptotic properties of such estimators. Among these, the *Hadamard differentiability* property has been exploited mostly. It turns out [viz., Sen, 1996a] that U-statistics for unbounded kernels may not be Hadamard differentiable, although they possess nice (reverse) martingale properties which provide access to various probabilistic tools that can be used to study related asymptotics. Likewise for L-, M- and R-estimators in a functional mold, one needs bounded score (weight) functions to verify the desired Hadamard differentiability property. For L- and M-estimators, such a boundedness condition does not pose any serious threat (as robustness considerations often prompt bounded influence functions), but bounded scores for R-estimators exclude some important statistics (such as the classical *normal scores* and *log-rank scores* statistics), and hence, this differentiability approach for R-estimators is not totally appropriate. Fortunately, there are alterenative methodologies, studied in detail in Jurečková and Sen (1996, ch.6), which provide a better resolution, and hence, we need not be overconcerned about Hadamard differentiability of EF's for R-estimators.

For all the types of estimators considered above there is a basic query: Treating the underlying density g to be a nuisance functional belonging to a general class $\mathcal{G}$, is it possible to formulate some EF which yields asymptotically optimal estimators in the sense of attainment of the information bound for its asymptotic mean square error in a semiparametric setup? Un-

der fairly general regulaity conditions, we have an affirmative answer. We illustrate this point with adaptive R-estimators of location (Hušková and Sen, 1986), and a similar situation holds for other functionals as well (Sen, 1996a). Let us define the Fisher-information score generating function by

$$\phi_g(t) = \{-g'(G^{-1}(t))/g(G^{-1}(t))\},\ t \in (0,1), \tag{2.26}$$

and assume that $I(g) = \int_0^1 \phi_g^2(t)dt < \infty$. Then we may consider a Fourier series expansion

$$\phi_g(t) \equiv \sum_{k\geq 0} \gamma_k P_k(t),\ t \in (0,1) \tag{2.27}$$

where the $\{P_k(.), k \geq 0\}$ form a complete orthonornal system [on $(0,1)$] and the Fourier coefficients γ_k are defined as

$$\gamma_k = \int_0^1 \phi_g(t)P_k(t)dt,\ \text{for } k = 0,1,.... \tag{2.28}$$

For adaptive R-estimators, Hušková and Sen (1985,1986) advocated the use of the *Legendre polynomial system* on $(0,1)$, where for each $k(= 0,1,2,\ldots)$,

$$P_k(t) = (2k+1)^{1/2}(-1)^k(k!)^{-1}(d^k/dt^k)\{t(1-t)\}^k,\ t \in (0,1), \tag{2.29}$$

so that $P_0(t) = 1$, $P_1(t) = \sqrt{3}(2t-1)$, $P_2(t) = \sqrt{5}\{1-6t(1-t)\}$ and so on. Thus $P_1(.)$ is a variant of the Wilcoxon scores, and when g is a logistic density, it is easy to see that $\phi(.) \equiv P_1(.)$, while for a symmetric g, it follows from the above that $\gamma_{2k} = 0$, $\forall k \geq 0$. In general the Fourier series is an infinite one, though convergent. The basic idea is to truncate this infinite series at a suitable *stopping number*, say, K_n, and for such a truncated version $\sum_{k\leq K_n} \gamma_k P_k(t)$, to estimate the Fourier coefficients γ_k by the classical Jurečková-linearity method. If we denote these estimates by $\hat{\gamma}_{k,n}$, $k \leq K_n$, then our adaptive version of the Fisher score function is

$$\hat{\phi}_{g,n}(t) = \sum_{k\leq K_n} \hat{\gamma}_{k,n} P_k(t),\ t \in (0,1). \tag{2.30}$$

One may then define a (signed-) rank statistic as in before with the scores $\hat{a}_n(k)$, $k = 1,\ldots,n$ generated by the adaptive score generating function $\hat{\phi}_{g,n}(\cdot)$, and, based on the same alignment principle, obtain adaptive R-estimators of location. These estimators are robust and asymptotically efficient for the class of densities with finite Fisher information. Hušková and Sen (1986) considered a suitable sequential version to formulate suitable stopping numbers $\{K_n\}$ for which there are suitable rate of convergence; a simpler algorathim without appealing to such a sequential scheme has been worked out in Sen (1996a).

3 Estimating Functions for Linear Models

In linear models though observations may not be i.d., the error components are assumed to be i.i.d.r.v.'s. Conventionally, we define a vector of observable random variables $\mathbf{Y}_n = (Y_1, \ldots, Y_n)'$ by letting

$$\mathbf{Y}_n = \mathbf{X}_n\boldsymbol{\beta} + \mathbf{e}_n; \quad \mathbf{e}_n = (e_1, \ldots, e_n)', \tag{3.1}$$

where $\mathbf{X}_n$ is a given (nonstochastic) $n \times p$ matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown regression parameters, and the errors e_i are assumed to be i.i.d.r.v.'s. For normally distributed errors the ML EE's are linear in $\mathbf{Y}_n$, and they agree with the LSE. Specifically, we have

$$\hat{\boldsymbol{\beta}}_n = \{\mathbf{X}_n'\mathbf{X}_n\}^{-1}\mathbf{X}_n'\mathbf{Y}_\mathbf{n}. \tag{3.2}$$

In a general parametric model with an assumed error density f, the ML EE is given by

$$\sum_{i=1}^{n} \mathbf{x}_i'\{-f'(Y_i - \mathbf{x}_i'\mathbf{b})/f(Y_i - \mathbf{x}'\mathbf{b})\} = \mathbf{0}, \tag{3.3}$$

where $\mathbf{X}_n' = (\mathbf{x}_1', \ldots, \mathbf{x}_n')$. The situation is quite comparable to the i.i.d. case presented in (2.9) through (2.15), and although the ARE properties are isomorphic, possible lack of robustness would be more accentuated here because of the nonidentity of the $\mathbf{x}_i$. Moreover, for f not belonging to an exponential family, the resulting MLE in (3.3) may not be linear estimators, and hence, a trial and error solution may be necessary.

The usual M-estimators for such linear models are based on EF's that resemble (3.3) along the same line as in Section 2: One would use a score function $\psi(.) : \mathcal{R} \to \mathcal{R}$ satisfying the same regularity conditions as in Section 2, and consider the EE:

$$\sum_{i=1}^{n} \mathbf{x}_i'\psi(Y_i - \mathbf{x}_i'\mathbf{b}) = \mathbf{0}. \tag{3.4}$$

Jurečková and Sen (1996,ch.5) contains an extensive treatment of first and second order asymptotic distributional representations for M-estimators in linear models; their treatise exploits mostly the uniform asymptotic linearity results of M-statistics (in the regression parameters). Another approach to this type of representation is based on the Hadamard differentiability of extended statistical functionals, and this has been considered in detail by Ren and Sen (1991,1994,1995). As in location models, such M-estimators have good local robustness and optimality properties, but they are not genuinely nonparametric.

R-estimators of regression parameters are based on suitable linear rank statistics, and they are globally robust and asymptotically optimal for suitable subfamilies of underlying densities. For a given $\mathbf{b} \in \mathcal{R}^p$, we define the residuals by

$$Y_i(\mathbf{b}) = Y_i - \mathbf{x}_i\mathbf{b},\ i = 1, \ldots, n, \tag{3.5}$$

and denote the aligned ranks by

$$R_{ni}(\mathbf{b})[= \sum_{j=1}^{n} I(Y_j(\mathbf{b}) \leq Y_i(\mathbf{b}))],\ i = 1, \ldots, n. \tag{3.6}$$

Then a vector of aligned rank statistics is defined by

$$\mathbf{L}_n(\mathbf{b}) = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_n) a_n(R_{ni}(\mathbf{b})),\ \ \mathbf{b} \in \mathcal{R}^p, \tag{3.7}$$

where the scores $a_n(1) \leq \cdots \leq a_n(n)$ are defined as in the location case, and $\bar{\mathbf{x}}_n = n^{-1} \sum_{i \leq n} \mathbf{x}_i$. Note that if we take the design matrix in the conventional form with the first column vector as $\mathbf{1}_n$, then β_1 is the intercept parameter. In this case, in the last equation, the contribution of the first coordinate of the $\mathbf{x}_i$ and $\mathbf{b}$ will be null (as the ranks are translation-invariant). Thus, we would have $p - 1$ elements in $\mathbf{L}_n(\mathbf{b})$, while for the first element, we need to use suitable signed rank statistics.

In the case of real valued b, for monotone scores, $L_n(b)$ is a monotone (step-) function, so that an estimator of β can be obtained by "equating $L_n(b)$ to 0"; unmiform asymptotic linearity of $L_n(b)$ in a shrinking neighborhood of β provides access to the asymptotic properties of the estimator. In the case of vector valued $\mathbf{b}$, unlike the case of LSE, these aligned linear rank statistics are not linear in $\mathbf{b}$; typically for monotone scores, for each $j(= 1, \ldots, p)$, the jth coordinate of $\mathbf{L}_n(\mathbf{b})$ is montone in b_j, but may not be so with respect to the remaining components of $\mathbf{b}$, so we do not have a linear EE. There are various approaches to motivate such EF's, and a detailed account of these is given by Jurečková and Sen (1996, ch.6). Jaeckel (1972) introduced a rank dispersion measure

$$D_n(\mathbf{b}) = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i\mathbf{b}) a_n(R_{ni}(\mathbf{b})),\ \mathbf{b} \in \mathcal{R}^p, \tag{3.8}$$

and proposed to minimize $D_n(\mathbf{b})$ with respect to $\mathbf{b}$ to obtain an estimator of $\boldsymbol{\beta}$. Since the ranks are translation-invariant, it can be shown easily that $D_n(\mathbf{b})$ is nonnegative, piecewise linear (and hence, continuous) and convex function of $\mathbf{b}$. Thus, $D_n(\mathbf{b})$ is almost everywhere differentiable with respect to $\mathbf{b}$ and $(\partial/\partial\mathbf{b})D_n(\mathbf{b}) = -L_n(\mathbf{b})$ at any point of differentiability. Therefore 'equating $\mathbf{L}_n(\mathbf{b})$ to $\mathbf{0}$' in a suitable norm (such as the $\mathcal{L}_1$ norm) yields a convenient R-estimator of $\boldsymbol{\beta}$. In this context too, the uniform asymptotic linearity

of $\mathbf{L}_n(\mathbf{b})$ in $\mathbf{b}$ in a shrinking neighborhood of the true $\boldsymbol{\beta}$ provides access to the variety of asymptotic results pertaining to robustness and asymptotic representations for R-estimators.

R-estimators in linear models are closely related to *regression rank scores* (RRS) estimators, developed mostly due to Gutenbrunner and Jurečková (1992). To indroduce such estimators, we make use of some related EF's, due to Koenker and Bassett (1978), termed the *regression quantiles* (RQ), which possess a basic regression equivariance property and are variants of L-estimators in linear models. For a given $\alpha : 0 < \alpha < 1$, define

$$\rho_\alpha(x) = |x|\{(1-\alpha)I\{x < 0) + \alpha I(x > 0)\},\ x \in \mathcal{R}, \tag{3.9}$$

and define the α-RQ estimator $\hat{\boldsymbol{\beta}}_n(\alpha)$ by

$$\hat{\boldsymbol{\beta}}_n(\alpha) = \arg\min\{\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i'\mathbf{b})) : \mathbf{b} \in \mathcal{R}_p\}. \tag{3.10}$$

Following Koenker and Bassett (1978), it can also be shown that an $\alpha-$RQ can be characterized as optimal solution $(\boldsymbol{\beta}(\alpha))$ of the *linear programming* problem

$$\begin{aligned} \alpha\sum_{i=1}^{n} r_i^+ \ &+\ (1-\alpha)\sum_{i=1}^{n} r_i^- = \min \\ \sum_{i=1}^{n} x_{ij}\beta_j \ &+\ r_i^+ - r_i^- = Y_i,\ i = 1,\ldots,n; \\ \beta_j \in \mathcal{R},\ j = 1,\ldots,p; \qquad & r_i^+ \geq 0,\ r_i^- \geq 0,\ i = 1,\ldots,n; \end{aligned} \tag{3.11}$$

where $r_i^+(r_i^-)$ is the positive (negative) part of the residual $Y_i - \boldsymbol{\beta}'\mathbf{x}_i$, $i = 1,\ldots,n$. Thus, for a given $\alpha : 0 < \alpha < 1/2$, usually small, if we define a diagonal matrix $\mathbf{C}_n = \text{diag}(c_{n1},\ldots,c_{nn})$ by letting

$$c_{ni} = I\{\mathbf{x}_i'\boldsymbol{\beta}(\alpha) < Y_i < \mathbf{x}_i'\boldsymbol{\beta}(1-\alpha)\}, i = 1,\ldots,n, \tag{3.12}$$

then an *α-trimmed* (T)LSE of $\boldsymbol{\beta}$ can be defined as

$$\mathbf{T}_n(\alpha) = (\mathbf{X}_n'\mathbf{C}_n\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{C}_n\mathbf{Y}_n. \tag{3.13}$$

Here the EF in (3.9)-(3.10) is primarily used to obtain the matrix $\mathbf{C}_n$, so that robustness and efficiency considerations are to be related to this basic choice. We refer to Jurečková and Sen (1996, ch.4) for some detail discussion of these aspects of RQ and related TLSE. For introducing the RRS estimators, for a given $\alpha \in (0,1)$, we define the vector of regression ranks

(RR) $\hat{\mathbf{a}}_n(\alpha) = (\hat{a}_{n1}(\alpha), \ldots, a_{nn}(\alpha))'$ as the optimal solution of the linear programming problem:

$$\sum_{i=1}^{n} Y_i \hat{a}_{ni}(\alpha) = \max \quad ; \quad \sum_{i=1}^{n} x_{ij} \hat{a}_{ni}(\alpha) = (1-\alpha) \sum_{i=1}^{n} x_{ij}, \; j = 1, \ldots, p;$$
$$\hat{a}_{ni}(\alpha) \in [0,1], \qquad i = 1, \ldots, n; \; \alpha \in (0,1). \tag{3.14}$$

Let $\phi(u), u \in (0,1)$ be a nondecreasing, square integrable score generating function, and let

$$\begin{aligned} \phi_n(u) \;&=\; \phi(\alpha^*) I(0 < u \leq \alpha^*) + \phi(u) I(\alpha^* < u < 1 - \alpha^*) \\ &+\; \phi(1-\alpha^*) I(1 - \alpha^* \leq u < 1), \end{aligned} \tag{3.15}$$

where $\alpha \in (0, 1/2)$ and is usually chosen small. Then the RRS, generated by the score function ϕ, are taken as

$$\hat{b}_{ni} = -\int_0^1 \phi_n(\alpha) d\hat{a}_{ni}(\alpha), \; i = 1, \ldots, n. \tag{3.16}$$

We can then consider a regression rank measure of dispersion:

$$D_n^*(\mathbf{b}) = \sum_{i=1}^{n} (Y_i - \mathbf{b}'\mathbf{x}_i)[\hat{b}_{ni}(\mathbf{Y} - \mathbf{X}\mathbf{b}) - \bar{\phi}], \tag{3.17}$$

where the RRS $(\hat{b}_{ni}(\mathbf{Y} - \mathbf{X}\mathbf{b}), i = 1, \ldots, n)$ are computed from the aligned observations $\mathbf{Y} - \mathbf{X}\mathbf{b}$. Then the derived RRS estimator is defined as

$$\tilde{\boldsymbol{\beta}}_n = \text{ Arg. } \min\{D_n^*(\mathbf{b}) : \mathbf{b} \in \mathcal{R}^p\}. \tag{3.18}$$

A similar estimating function works out for the estimation of a subvector of $\boldsymbol{\beta}$. The interesting fact is that under fairly general regularity conditions and based on a common score function $\phi(.)$, the classical R-estimator and the RRS estimator are asymptotically equivalent upto the order $(n^{-1/2})$; we refer to Section 6.8 of Jurečková and Sen (1996) for details. Robustness aspects of both these type of rank estimators can therefore be studied in a unified manner, without requiring a bounded score generating function. We conclude this section with a note that as in the case with the semiparametric location model, here in the semiparametric linear models whenever the density f admits a finite Fisher information $I(f)$, we may construct adaptive EF's based on adaptive rank scores statistics, and these yield asymptotiocally optimator estimators of the parameters involved in the parametric part of the model [viz., Hušková and Sen, 1985; Sen, 1996a]. Thus, adaptive EF's in semiparametric linear models, though computationally more complex, yield robust and asymptotically efficient estimators.

4 Estimating Functionals.

In traditional and generalized linear models, as well as in other specific forms of nonlinear models, essentially the parameter space is finite dimensional,so that EF's are vector valued. Robustness and nonparametric perspectives often prompt us not to advocate such finite dimensional models; the conditional mean or quantile regression functions in a multuivariate nonparametric setup are simple and classical examples of such functionals. While the basic motivations for EF's remain essentially the same in such a functional case, technical manipulations are usually more extensive. For this reason, robustness and efficiency considerations are to be assessed in a somewhat different manner.

In a multivariate setup, let $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$, be i.i.d.r.v.'s with a d.f. F defined on $\mathcal{R}^{p+1}$, and let $G(y|\mathbf{x})$ be the conditional d.f. of Y, given $\mathbf{X} = \mathbf{x}$. Then typically a regression functional of Y on $\mathbf{x}$ is a location functional of the conditional d.f. $G(.|\mathbf{x})$, $\mathbf{x} \in \mathcal{R}^P$. Therefore as long as this conditional d.f. can be estimated consistently and efficiently in a nonparametric fashion, suitable sample counterparts of such location functionals (based on, for example, appropriate L-, M- and R-statistics) can be constructed in a robust manner. Sans multinormality of F, such regression functionals may not be generally linear (in $\mathbf{x}$), and moreover a specific nonlinear form in turn calls for a specific form of F, though in general such functionals may otherwise exhibit good smoothness properties. Therefore in a nonparametric or semiparametric setup, it seems quite plausable to incorporate appropriate smoothness conditions on such conditional functionals and estimate them in a robust, consistent and efficient manner. Of course, one has to pay a little penalty for choosing an infinite dimensional parameter space when actually a finite dimensional one prevails, but, in the opposite case, a finite dimensional model based statistical analysis may be totally inadequate when a functional parameter space prevails.

Among various possibilities, we may mention specifically the two popular approaches to this problem. They are (i) the *nearest neighbor* (NN) method, and (ii) *kernel smooting* methods. In a NN-method, corresponding to a set pivot $\mathbf{x}_o$, usually lying in a convex set $C \in \mathcal{R}^p$, we define the *pseudovariables*

$$Z_i = d(\mathbf{X}_i, \mathbf{x}_o), \; i = 1, \ldots, n, \tag{4.1}$$

where $d(.)$ is a suitable metric on R^p (which may as well be taken as the Euclidean norm), and denote the corresponding order statistics by $Z_{n:1} \leq \cdots \leq Z_{n:n}$; note that they specifically depend on the base sample as well as the chosen pivot. Also we define a nondecreasing sequence $\{k_n\}$ of positive integers such that $k_n \to \infty$ but $n^{-1}k_n \to 0$ as $n \to \infty$. Further, we set the antiranks S_i by letting $Z_{n:i}$ correspond to $\mathbf{X}_{S_i}$, for $i = 1, \ldots, n$. Then a

k_n-NN empirical d.f. at the pivot $\mathbf{x}_o$ is defined by

$$G_{n,k_n}(y|\mathbf{x}_o) = k_n^{-1} \sum_{i=1}^{k_n} I(Y_{S_i} \leq y),\ y \in \mathcal{R}. \tag{4.2}$$

Estimating functionals are then based on the entire set $G_{n,k_n}(\cdot|\mathbf{x}_o),\ \mathbf{x}_o \in \mathcal{C}$. Naturally robustness considerations dominate the choice of such functionals (viz., Sen 1996b). In a kernel method, we choose a known density $\phi(\mathbf{x})$ possessing some smoothness properties such as unimodality and symmetry around $\mathbf{0}$, compact support and differentiability upto a certain order, and define a smooth conditional empirical d.f. by integrating $\phi(\mathbf{x}-\mathbf{x}_o)$ with respect to the empirical d.f. F_n. This conditional measure is then incorporated in the formulation of suitable robust functionals. These two methods compare favorably with respect to their *asymptotic bias* and *asymptotic mean squares* and robustness pictures. There is, however, a common concern that stems from the fact that such conditional functionals span an infinite dimensional parameter space, and hence, in setting suitable confidence sets, a prescription in terms of a finite number of pivots may not suffice. Weak convergence approaches [viz., Sen 1993, 1995a, 1996b] provide viable alternatives, yet retaining robustness to a certain extent.

5 General Remarks

In the context of (generalized) linear models (GLM),possibly involving nuisance parameters, EF's have received a good deal of attention, and these developments constitute a major advancement in the research literature. There is, however, a point worth mentioning: The very motivation of retaining the flavor of exponential family of densities by skillful choice of (canonical) *link functions* yields (generalized) GEE's that share nonrobustness properties with MLE's and other parametric estimators. In most biomedical applications the response variate is nonnegative and has typically a positively skewed distribution, so often the Box-Cox type transformation is used to induce more symmetry (if not normality); however, this may also distort the inherent linearity or other parametric structure of the underlying dose-response relations. Hence GLM [viz., McCullagh and Nelder, 1989] may not be universally advocated in such studies. Considerations of model robustness naturally call for nonparametrics or semiparametrics , and as such L-, M- and R- EF's along with their siblings come into the picture. In this context, the dimension of nuisance parameters is often large if not infinity, and the estmation parameter space may also be large. In a quasi-parametric setup, under the coverup of semiparametrics, Godambe (1985) initiated a line of attack for such EF's that are potentially applicable to various stochastic

processes where independence of the observations may not hold. A good deal of extensions of his seminal work has taken place during the past ten years. If we have a good feeling of the conditional distributions of the observations given the past ones, then Godambe's scores can be obtained in an appropriate manner, and his suggested avenue leads to a finite sample optimality property, interpreted in terms of the smallness of the mean square error of the estimators. However, sans the knowledge of these underlying distributions, we may not be able to decide a proper choice of the Godambe scores, and this may vitiate his small sample optimality properies when the assumed scores do not correspond to the likelihood based ones, so such semiparametric procedures are likely to be quite nonrobust to possible departures of the assumed conditional distributions from the true ones. Therefore, as in Huber (1964), we should consider some scores which remain robust in such situations. In a genuine semiparametric model, we usually allow a finite dimensional estimable parameter space retaining the infinite dimensionality of the nuisance parameter space; for example, the d.f.'s are unknown and arbitrary, but the linearity of regression prevails. As such, compared to pure nonparametrics, such semiparametrics may yield more efficient estimators when the postulated model is correct, but is naturally more nonrobust to plausible departures from such assumed models. One other advantage of semiparametrics is that the finite dimensionality of the estimable parameter space usually permits the adoption of adaptive procedures which are asymptotically optimal with respect to the postulated model. We refer again to the semiparametric linear models for which adaptive R- or M-estimators which are asymptotically efficient [viz., Hušková and Sen (1985,1986)]. In a relatively more general setup of statistical functionals, under a similar semiparametric modeling, such adaptive EF's have also been discussed in Sen (1996a). From this perspective it is clear that modeling part is a vital task in formulating the estimation space, and EF's are to be considered in the light of dimensionality and structure of this setup. In this context, robustness and (asymptotic) efficiency considerations are of utmost importance. In most biomedical, clinical and environmental studies, generally this modeling is far more complex, and conventional parametric GLM's may not be that appropriate even following suitable transformations. Therefore, there is a need to focus on the appropriateness of suitable semiparametric and nonparametric models, and model flexibility often favors the latter choice. We refer to Sen (1996c) for some discussion of EF's in GLM's in biostatistical applications.

6 Acknowledgements

The author is grateful to Professor V. P. Godambe and the referees for their useful comments on the manuscript.

References

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser.B* 74, 187-220.

Godambe, V. P. (1985). The foundation of finite sample estimation in stochastic processes. *Biometrika 72,* 419-428.

Gutenbrunner, C., and Jurečková, J. (1992). Regression rank scores and regression quantiles. *Ann. Statist.* 20, 305-330.

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

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* 19, 293-325.

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

Hušková, M. and Sen, P. K. (1985). On sequentially adaptive asymptotically efficient rank statistics. *Sequen. Anal. 4,* 125-151.

Hušková, M. and Sen, P. K. (1986). Sequentially adaptive signed rank statistics. *Sequen. Anal. 5,* 237-251.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist. 43,* 1449-1458.

Jurečková, J. and Sen, P. K. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations.* Wiley, New York.

Koenker, R., and Bassett, G. (1978). Regression quantiles. *Econometrica 46,* 33-50.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models,* 2nd ed. Chapman and Hall, London.

Nolan, D., and Pollard, D. (1987). U-processes: rates of convergence. *Ann. Statist. 15,* 780-789.

Ren, J.-J., and Sen, P. K. (1991). Hadamard differentiability of extended statistical functionals. *J. Multivar. Anal. 39,* 30-43.

Ren, J.-J., and Sen, P. K. (1994). Asymptotic normality of regression M-estimators: Hadamard differentiability approaches. In *Asmptotic Statistics* (eds. Mandl, P. and Hušková, M.), Physica-Verlag, Vienna, pp. 131-147.

Ren, J.-J., and Sen, P. K. (1995). Hadamard differentiability on $D[0,1]^p$. *J. Multivar. Anal. 45,* 14-28.

Sen, P. K. (1964). On some properties of rank weighted means. *J. Ind. Soc. Agri. Statist. 16,* 51-61.

Sen, P. K. (1981). *Sequential Nonparametrics : Invariance Principles and Statistical Inference.* Wiley, New York.

Sen, P. K. (1993). Perspectives in multivariate nonparametrics: Conditional functionals and ANOCOVA models. *Sankhyā, Ser.A 55,* 516-532.

Sen, P. K. (1994). Regression quantiles in nonparametric regression. *J. Nonparamet. Statist. 3,* 237-253.

Sen, P. K. (1995). Robust and nonparametric methods in linear models with mixed effects. *Tetra Mount. Math. Public. 7,* 331-342.

Sen, P. K. (1996a). Statistical functionals, Hadamard differentiability and martingales. In *A Festschrift for J. Medhi* (eds. Borthakur, A. C., and Chaudhury, H.), New Age Press, Delhi, pp.29-47.

Sen, P. K. (1996b). Regression rank scores estimation in ANOCOVA. *Ann. Statist. 24,* 1586-1601.

Sen, P. K. (1996c). An appraisal of generalized linear models in biostatistical applications. *J. Appl. Statist. Sc. 5,* 61-78.

Shorack, G. R., and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist. 18,* 309-348.