ON ESTIMATING A PARAMETER AND ITS SCORE FUNCTION, II

C. A. J. KLAASSEN, A. W. VAN DER VAART AND W. R. VAN ZWET

Department of Mathematics
University of Leiden
Leiden, Netherlands

1. INTRODUCTION

A bound of Cramér-Rao type is provided for an estimator of a real-valued parameter θ in the presence of an abstract nuisance parameter η, such as an unknown distributional shape, on the basis of N i.i.d. observations. The bound consists of the reciprocal of the effective Fisher information in the sample, plus a term involving the integrated mean squared error of an estimator of a multiple of the so-called conditional score function for θ, for the case where θ is known. This implies that an estimator of θ can only perform well over a class of shapes η if it is possible to estimate the conditional score function for θ accurately over this class. For the special case where fully adaptive estimation may be possible, this result was given in a companion paper (Klaassen and van Zwet (1985)).

2. AN INEQUALITY OF CRAMÉR-RAO TYPE

Let $X_1, \ldots, X_N$ be independent and identically distributed (i.i.d.) random variables taking values in some measurable space $(\mathcal{X}, \mathcal{A})$, with a common density $f(\cdot\,; \eta, \theta)$ with respect to a σ-finite measure μ on $(\mathcal{X}, \mathcal{A})$. The parameter of interest θ belongs to an open subset Θ of $\mathbb{R}$ and the nuisance parameter η ranges over an arbitrary set H. For unknown η and θ, it is required to estimate θ and this is done by means of an estimator $T_N = T_N(X_1, \ldots, X_N)$ for some measurable function T_N: $\mathcal{X}^N \to \mathbb{R}$. We are interested in finding a lower bound for the variance of T_N under $f(\cdot\,; \eta, \theta)$. We shall write $P_{\eta\theta}$, $E_{\eta\theta}$ and $\sigma^2_{\eta\theta}$ for probabilities, expectations and variances under this model.

For every fixed $\theta \in \Theta$ and $j = 1, \ldots, N$, let $\psi(X_j; \theta)$ be a sufficient statistic for X_j with respect to $\eta \in H$. According to the factorization theorem this is equivalent to assuming that

$$f(x; \eta, \theta) = g(\psi(x; \theta); \eta, \theta) h(x; \theta) \quad \text{a.e. } [\mu], \tag{2.1}$$

where $g(\cdot\,; \eta, \theta)$ may be chosen to be the density of $\psi(X_1; \theta)$ with respect to a σ-finite measure ν_θ.

We shall assume that $f^{\frac{1}{2}}$ is differentiable in quadratic mean with respect

S. van de Geer and M. Wegkamp (eds.), *Selected Works of Willem van Zwet*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4614-1314-1_8, © Springer Science+Business Media, LLC 2012

to θ with a derivative which is not essentially zero, thus for every (η, θ)

$$\lim_{\theta' \to \theta} \int [(\theta' - \theta)^{-1} \{ f^{\frac{1}{2}}(x; \eta, \theta') - f^{\frac{1}{2}}(x; \eta, \theta) \} - \frac{1}{2} \tau(x; \eta, \theta) f^{\frac{1}{2}}(x; \eta, \theta)]^2 d\mu(x) = 0, \tag{2.2}$$

$$I(\eta, \theta) = E_{\eta\theta} \tau^2(X_1; \eta, \theta) > 0. \tag{2.3}$$

Obviously (2.2) implies that $I(\eta, \theta) < \infty$. Note that $\tau(\cdot\,; \eta, \theta)$ is simply an L_2-version of the classical score function for θ, $\partial \log f(x; \eta, \theta)/\partial\theta$; $I(\eta, \theta)$ is the Fisher information concerning θ which is contained in a single observation X_1 and measures how well θ can be estimated when η is known. However, since η is unknown, one expects the information concerning θ to be smaller. As discussed in Begun, Hall, Huang and Wellner (1983), the information loss results from a reduction of the score function.

First we define score functions in the η-direction. We shall say that $\beta(\cdot\,; \eta, \theta)$ is an η-score function if there exists a sequence $\eta_k \in H$ such that

$$\lim_{k \to \infty} \int [k \{ f^{\frac{1}{2}}(x; \eta_k, \theta) - f^{\frac{1}{2}}(x; \eta, \theta) \} - \frac{1}{2} \beta(x; \eta, \theta) f^{\frac{1}{2}}(x; \eta, \theta)]^2 d\mu(x) = 0. \tag{2.4}$$

It is easy to see that, in view of (2.1), (2.4) implies that

$$\beta(x; \eta, \theta) = b(\psi(x; \theta); \eta, \theta) \quad \text{a.e. } [P_{\eta\theta}], \tag{2.5}$$

where b satisfies

$$\lim_{k \to \infty} \int [k \{ g^{\frac{1}{2}}(v; \eta_k, \theta) - g^{\frac{1}{2}}(v; \eta, \theta) \} - \frac{1}{2} b(v; \eta, \theta) g^{\frac{1}{2}}(v; \eta, \theta)]^2 d\nu_\theta(v) = 0, \tag{2.6}$$

so that b is an η-score function for the model $\{g(\cdot\,; \eta, \theta)\colon \eta \in H, \theta \in \Theta\}$. Let $B(\eta, \theta)$ denote the set of all η-score functions for the original model — i.e. functions β for which (2.4)–(2.5) hold for an appropriate sequence $\eta_k \in H$ — and let $\tilde{B}(\eta, \theta)$ be the closure in L_2 of the linear span of $B(\eta, \theta)$.

The effective score function τ_E for θ in the presence of the nuisance parameter η, is defined as

$$\tau_E(x; \eta, \theta) = \tau(x; \eta, \theta) - b_E(\psi(x; \theta); \eta, \theta), \tag{2.7}$$

where $b_E(\psi(x; \theta); \eta, \theta)$ is the L_2-projection of τ on $\tilde{B}(\eta, \theta)$, thus

$$\begin{aligned} I_E(\eta, \theta) &= E_{\eta\theta} \{ \tau(X_1; \eta, \theta) - b_E(\psi(X_1; \theta); \eta, \theta) \}^2 \\ &= \min_{\beta \in \tilde{B}(\eta, \theta)} E_{\eta\theta} \{ \tau(X_1; \eta, \theta) - \beta(X_1) \}^2. \end{aligned} \tag{2.8}$$

$I_E(\eta, \theta)$ is the effective Fisher information, which measures how well θ can be estimated when η is unknown (cf. Begun et al. (1983), but note that we do not assume that $B(\eta, \theta)$ itself is a linear space).

Let $C(\eta, \theta)$ denote the set of all square-integrable functions $b(\psi(x; \theta))$ with $E_{\eta\theta}\, b(\psi(X_1; \theta)) = 0$. In the special case where $\tilde{B}(\eta, \theta) = C(\eta, \theta)$, $b_E(v; \eta, \theta)$ equals the conditional expectation $E_{\eta\theta}(\tau(X_1; \eta, \theta) | \psi(X_1; \theta) = v)$ and τ_E and I_E

reduce to

$$\tau_C(x;\eta,\theta) = \tau(x;\eta,\theta) - E_{\eta\theta}(\tau(X_1;\eta,\theta)|\psi(X_1;\theta) = \psi(x;\theta)), \tag{2.9}$$

$$I_C(\eta,\theta) = E_{\eta\theta}\tau_C^2(X_1;\eta,\theta), \tag{2.10}$$

which are called the conditional score function and the conditional Fisher information for θ. In general, however, $\tilde{B}(\eta,\theta)$ may be a proper subset of $C(\eta,\theta)$, and hence

$$I_C(\eta,\theta) \leq I_E(\eta,\theta) \leq I(\eta,\theta) \tag{2.11}$$

as is clear from figure 1. Of course, we still have $\tau_C = \tau_E$ and $I_C = I_E$ if $E_{\eta\theta}(\tau(X_1;\eta,\theta)|\psi(X_1;\theta) = \psi(\cdot;\theta))$ happens to be in $\tilde{B}(\eta,\theta)$.

So far we have discussed various aspects of the model. Concerning the estimator T_N, we assume that, whenever $E_{\eta\theta}\, T_N^2 < \infty$ for a certain (η,θ), then

$$\sup_{(\eta',\theta')\in A_\varepsilon} E_{\eta'\theta'} T_N^2 < \infty \tag{2.12}$$

for some $\varepsilon > 0$, where

$$A_\varepsilon = \{(\eta',\theta'): \int |f(x;\eta',\theta') - f(x;\eta,\theta)|d\mu(x) < \varepsilon\} \tag{2.13}$$

consists of parameter values "close" to (η,θ). For simplicity we shall also assume that T_N is an unbiased estimator of θ, i.e. for all (η,θ),

$$E_{\eta\theta}T_N = \theta. \tag{2.14}$$

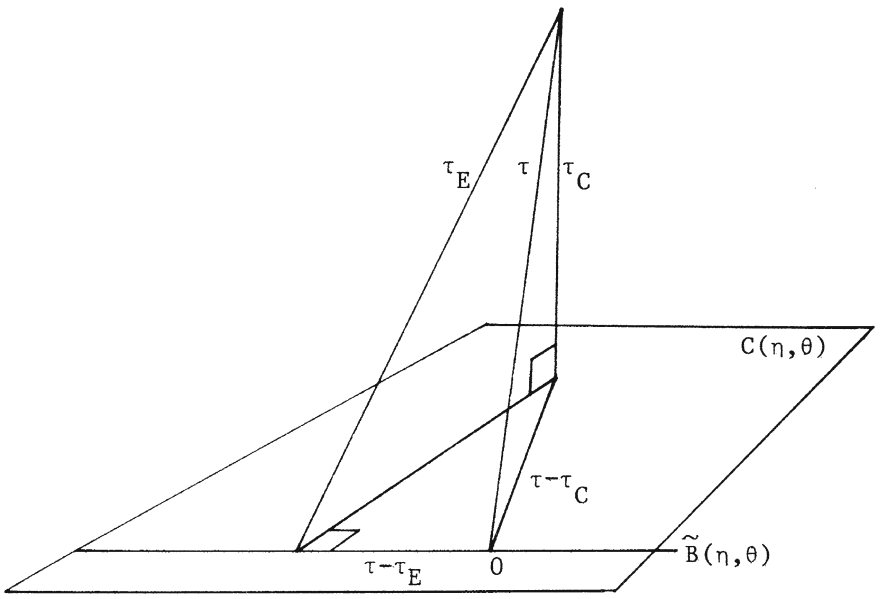

Figure 1.

The role of assumption (2.12)–(2.13) is to ensure that (2.14) implies that

$$E_{\eta\theta}T_N \sum_{j=1}^{N} \tau(X_j;\eta,\theta) = 1, \tag{2.15}$$

$$E_{\eta\theta}T_N \sum_{j=1}^{N} \beta(X_j) = 0 \qquad \text{for all } \beta \in \tilde{B}(\eta,\theta), \tag{2.16}$$

and hence in particular

$$E_{\eta\theta}T_N \sum_{j=1}^{N} \tau_E(X_j;\eta,\theta) = 1 \tag{2.17}$$

in view of (2.7)–(2.8). If we define

$$S_N(\eta,\theta) = \frac{1}{NI_E(\eta,\theta)} \sum_{j=1}^{N} \tau_E(X_j;\eta,\theta) \tag{2.18}$$

then (2.17) asserts that $S_N(\eta,\theta)$ and $T_N - S_N(\eta,\theta)$ are uncorrelated under $P_{\eta\theta}$. As a consequence

$$\sigma^2_{\eta\theta}(T_N) = \sigma^2_{\eta\theta}(S_N(\eta,\theta)) + \sigma^2_{\eta\theta}(T_N - S_N(\eta,\theta)) \tag{2.19}$$

$$= \{N\ I_E(\eta,\theta)\}^{-1} + \sigma^2_{\eta\theta}(T_N - S_N(\eta,\theta)),$$

and hence

$$\sigma^2_{\eta\theta}(T_N) \geq \{N\ I_E(\eta,\theta)\}^{-1} \tag{2.20}$$

which is the form the Cramér-Rao inequality takes for unbiased estimation of θ in the presence of the nuisance parameter η. Note that this is the fixed sample size version of the corresponding asymptotic results in Begun et al. (1983).

However, (2.19) contains essentially more information than inequality (2.20). It implies that $\sigma^2_{\eta\theta}(T_N)$ can only come close to the Cramér-Rao bound (2.20) if $(T_N - \theta)$ is close to $S_N(\eta,\theta)$ under $P_{\eta\theta}$. It follows that if T_N performs well as an estimator of θ for all $\eta \in H$, then $(T_N - \theta)$ must resemble $S_N(\eta,\theta)$ under $P_{\eta\theta}$ for all $\eta \in H$ and $\theta \in \Theta$. It would seem therefore that $(T_N - \theta)$ contains information about the unknown function $\tau_E(\cdot;\eta,\theta)$. Let us try to extract this information.

For $S_N(\eta,\theta)$ as defined in (2.18) we have

$$E_{\eta\theta}(S_N(\eta,\theta)|\psi(X_j;\theta) \quad \text{for } j \neq i; X_i = x) \tag{2.21}$$

$$- E_{\eta\theta}(S_N(\eta\theta)|\psi(X_j;\theta) \quad \text{for } j \neq i; \psi(X_i;\theta) = \psi(x;\theta))$$

$$= \frac{1}{N\ I_E(\eta,\theta)}[\tau_E(x;\eta\theta) - E_{\eta,\theta}(\tau_E(X_1;\eta,\theta)|\psi(X_1;\theta) = \psi(x;\theta))]$$

$$= \frac{1}{N\ I_E(\eta,\theta)}\tau_C(x;\eta,\theta)$$

in view of (2.7) and (2.9). If $(T_N - \theta)$ resembles $S_N(\eta,\theta)$ under $P_{\eta\theta}$, we may hope that

$$J_N(x;\theta) = \sum_{i=1}^{N}\{E_{\eta\theta}(T_N|\psi(X_j;\theta) \quad \text{for } j \neq i; X_i = x) \tag{2.22}$$
$$- E_{\eta\theta}(T_N|\psi(X_j;\theta) \quad \text{for } j \neq i; \psi(X_i;\theta) = \psi(x;\theta))\}$$

can serve as an estimator of

$$J(x;\eta,\theta) = \frac{1}{I_E(\eta,\theta)}\tau_C(x;\eta,\theta). \tag{2.23}$$

Note that since for each j, $\psi(X_j;\theta)$ is sufficient for X_j with respect to $\eta \in H$ for fixed $\theta \in \Theta$, J_N is indeed independent of η. For known θ, it is therefore a legitimate estimator of J.

We shall prove the following result.

Theorem 2.1. *Suppose that assumptions* (2.1)–(2.3) *and* (2.12)–(2.14) *are satisfied for every* (η,θ). *Then, for every* (η,θ),

$$\sigma^2_{\eta\theta}(T_N) \geq \frac{1}{N\ I_E(\eta\theta)} + \frac{1}{N}E_{\eta\theta}\int\{J_N(x;\theta) - J(x;\eta,\theta)\}^2 f(x;\eta,\theta)d\mu(x). \tag{2.24}$$

The theorem asserts that the Cramér-Rao bound (2.20) may be improved by adding N^{-1} times the integrated mean squared error (MSE) of the estimator J_N of the function J, which is an unknown multiple of the conditional score function τ_C. For practical purposes it is unsatisfactory that the right-hand side of (2.24) depends on the choice of T_N. However, one may obviously rephrase the theorem to assert only the existence of an estimator J_N such that (2.24) holds. The message of the theorem is then clear: the accuracy with which one can estimate θ for unknown η is delimited by the effective Fisher information on the one hand and by how well one can estimate $J(\cdot;\eta,\theta)$ for known θ on the other. Clearly the latter depends heavily on the class H. If $J(\cdot;\eta,\theta)$ runs through a large class of functions as η ranges over H, then the integrated MSE of any estimator of J may be quite large, especially for particularly irregular choices of J. If J is restricted to a smaller class of nicely behaved functions as $\eta \in H$, then the integrated MSE may be much smaller. Finally, if η is known so that H consists of a single element, then $J(\cdot;\eta,\theta)$ can serve as an estimator of itself and (2.24) reduces to the Cramér-Rao inequality (2.20).

In a sense, the result of the theorem is not surprising. Estimation of a parameter θ for an unknown distributional shape is based typically on a preliminary estimate of an unknown score function followed by a good estimate of θ for the distributional shape corresponding to the estimated score function.

For such estimators a result like (2.24) is to be expected. The interesting aspect of the theorem, however, is that it is not assumed that the estimator T_N is based on a preliminary estimate of a score function, but that an estimate of J for known θ is derived from T_N. In effect we are saying that a successful estimation procedure for θ must involve — either explicitly or implicitly — the estimation of J and that because of this, the accuracy of estimating J enters into the lower bound for the variance of the estimator of θ.

Although the theorem is purely a finite sample result, it obviously has asymptotic implications. An asymptotic analogue would imply that effective estimation of θ, i.e.

$$\{N\, I_E(\eta,\theta)\}^{\frac{1}{2}}(T_N - \theta) \xrightarrow{D} N(0,1), \tag{2.25}$$

is possible only if the function J can be estimated consistently with respect to integrated MSE for known θ. In this context it is unsatisfactory that J involves the conditional score function τ_C rather than the effective score function τ_E and, indeed, Klaassen (1987) has shown that a somewhat stronger version of (2.25) does entail consistent estimation of $\tau_E/I_E(\eta,\theta)$.

Of course this discrepancy disappears if $\tau_C = \tau_E$, i.e. if the function $E_{\eta\theta}(\tau(X_1;\eta,\theta)|\psi(X_1;\theta) = \psi(x;\theta))$ is an element of $\tilde{B}(\eta,\theta)$. This situation is rather common and examples, including non-i.i.d. models, are given by van der Vaart (1986), who also explicitly constructs an effective estimator of θ based on a preliminary consistent estimator of τ_C for such models.

An even more special case occurs if $\tau_C = \tau$, so that $I_E = I$ and $J = \tau/I$. Now (2.24) provides a finite sample analogue of the statement that fully adaptive estimation of θ is possible only if τ/I can be estimated consistently. This situation was discussed in the companion paper Klaassen and van Zwet (1985).

3. PROOF OF THE THEOREM

The proof resembles that of theorem 1.1 in Klaassen and van Zwet (1985). Let

$$f_N(x) = \prod_{j=1}^{N} f(x_j;\eta,\theta)$$

denote the density of $X = (X_1,\ldots,X_N)$ with respect to the N-fold product measure μ^N taken at the point $x = (x_1,\ldots,x_N)$ and write

$$\frac{f_N^{\frac{1}{2}}(x;\eta,\theta') - f_N^{\frac{1}{2}}(x;\eta,\theta)}{(\theta' - \theta)} = \frac{1}{2}\rho_N(x;\eta,\theta) + \Delta_N(x;\eta,\theta,\theta'), \tag{3.1}$$

with

$$\rho_N(x;\eta,\theta) = f_N^{\frac{1}{2}}(x;\eta,\theta)\sum_{i=1}^{N}\tau(x_i;\eta,\theta). \tag{3.2}$$

Since N is fixed, a standard argument shows that (2.2) implies

$$\lim_{\theta'\to\theta}\int \Delta_N^2(x;\eta,\theta,\theta')d\mu_N(x) = 0. \tag{3.3}$$

In view of (2.14) we have

$$1 = \int T_N(x)\frac{f_N^{\frac{1}{2}}(x;\eta,\theta') - f_N^{\frac{1}{2}}(x;\eta,\theta)}{(\theta'-\theta)}\{f_N^{\frac{1}{2}}(x;\eta,\theta') + f_N^{\frac{1}{2}}(x;\eta,\theta)\}d\mu_N(x) \tag{3.4}$$

$$= \int T_N(x)\{\frac{1}{2}\rho_N(x;\eta,\theta) + \Delta_N(x;\eta,\theta,\theta')\}\{f_N^{\frac{1}{2}}(x;\eta,\theta') + f_N^{\frac{1}{2}}(x;\eta,\theta)\}d\mu_N(x).$$

If $E_{\eta\theta}T_N^2 = \infty$, there is nothing to prove. Suppose therefore that $E_{\eta\theta}T_N^2 < \infty$. Since (2.2) ensures that

$$\lim_{\theta'\to\theta}\int |f(x;\eta,\theta') - f(x;\eta,\theta)|d\mu(x) = 0,$$

(2.12) and (2.13) yield

$$\limsup_{\theta'\to\theta} E_{\eta\theta'}T_N^2 < \infty. \tag{3.5}$$

Together, (3.3), (3.5) and the Cauchy-Schwarz inequality show that

$$\lim_{\theta'\to\theta}\int T_N(x)\Delta_N(x;\eta,\theta,\theta')\{f_N^{\frac{1}{2}}(x;\eta,\theta') + f_N^{\frac{1}{2}}(x;\eta,\theta)\}d\mu_N(x) = 0, \tag{3.6}$$

$$\left|\int T_N(x)\rho_N(x;\eta,\theta)\{f_N^{\frac{1}{2}}(x;\eta,\theta') - f_N^{\frac{1}{2}}(x;\eta,\theta)\}d\mu_N(x)\right|$$

$$\leq \{C^2\int\{f_N^{\frac{1}{2}}(x;\eta,\theta') - f_N^{\frac{1}{2}}(x;\eta,\theta)\}2d\mu_N(x)\int\rho_N^2(x;\eta,\theta)d\mu_N(x)\}^{\frac{1}{2}} \tag{3.7}$$

$$+\{\int T_N^2(x)\{f_N^{\frac{1}{2}}(x;\eta,\theta') - f_N^{\frac{1}{2}}(x;\eta,\theta)\}^2 d\mu_N(x)\int_{\{|T_N|>C\}}\rho_N^2(x;\eta,\theta)d\mu_N(x)\}^{\frac{1}{2}}$$

for every $C > 0$. As θ' tends to θ, the first term on the right tends to zero for every C in view of (2.2). Since $E_{\eta\theta}T_N^2 < \infty$ and $E_{\eta\theta}\tau^2(X_1;\eta,\theta) < \infty$, the second term converges to zero as $C \to \infty$. It follows that the left-hand side of (3.7) converges to zero, and together with (3.4) and (3.6) this proves (2.15). A similar argument produces (2.16) and (2.19) follows.

It remains to show that

$$\sigma_{\eta\theta}^2(T_N - S_N(\eta,\theta)) \geq \frac{1}{N}E_{\eta\theta}\int\{J_N(x;\theta) - J(x;\eta,\theta)\}^2 f(x;\eta,\theta)d\mu(x). \tag{3.8}$$

To see this, we copy the argument leading from (2.9) to (2.11) in Klaassen and

van Zwet (1985), even though $S_N(\eta, \theta)$ is defined differently in that paper. We find

$$\sigma^2_{\eta\theta}(T_N - S_N(\eta,\theta)) \tag{3.9}$$

$$\geq N^{-1} E_{\eta\theta} \int \{\sum_{i=1}^{N} [E_{\eta\theta}(T_N - S_N(\eta,\theta)|\psi(X_j;\theta) \text{ for } j \neq i; X_i = x)$$

$$- E_{\eta\theta}(T_N - S_N(\eta,\theta)|\psi(X_j;\theta) \text{ for } j \neq i; \psi(X_i;\theta) = \psi(x;\theta))]\}^2 f(x;\eta,\theta) d\mu(x).$$

In view of (2.21)–(2.23), (3.9) is identical to (3.8) and the proof complete.

ACKNOWLEDGMENTS

Research was supported in part by the Office of Naval Research Contract N00014-80-C-0163.

BIBLIOGRAPHY

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11**, 432–452.

Klaassen, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, to appear.

Klaassen, C. A. J. and van Zwet, W. R. (1985). On estimating a parameter and its score function. *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, L. Le Cam and R. A. Olshen, editors, Vol. II, 827–839.

van der Vaart, A. W. (1986). Estimating a real parameter in a class of semi-parametric models. Submitted to *Ann. Statist.*