Applications of the van Trees inequality: a Bayesian Cramér-Rao bound

Richard D. Gill and Boris Y. Levit

Mathematical Institute
University Utrecht
Budapestlaan 6
3584 CD Utrecht
Netherlands

email: gill@math.ruu.nl, levit@math.ruu.nl

Abstract. We use a Bayesian version of the Cramér-Rao lower bound due to van Trees to give an elementary proof that the limiting distibution of any regular estimator cannot have a variance less than the classical information bound, under minimal regularity conditions. We also show how minimax convergence rates can be derived in various non- and semi-parametric problems from the van Trees inequality. Finally we develop multivariate versions of the inequality and give applications.

Résumé. Nous utilisons une version Bayesienne de l'inégalité de Cramér-Rao due à van Trees pour établir que la loi limite d'un estimateur regulier ne peut pas avoir une variance moindre que la borne d'information classique, sous les conditions de régularité minimale. Nous démontrons aussi qu'avec l'inégalité de van Trees on peut établir la vitesse de convergence minimax dans plusieurs problèmes non- et sémi-paramétriques. Finalement, nous dévelopons les versions multivarieés de cet inégalité en nous en donnons des applications.

Key words: parameter estimation, non-parametric estimation, semi-parametric models, quaratic risk, lower bounds.

1. Introduction

Basic statistics textbooks like to present the Cramér-Rao lower bound together with an informal description of large sample distributional properties of maximum likelihood estimators as demonstrating some kind of asymptotic optimality of the MLE. This approach is very unconvincing on several counts. Firstly, the Cramér-Rao bound only says something nice about unbiased estimators whereas most estimators in practice are biased, and the arguments for prefering unbiased estimators are rather weak. Also, traditionally the Cramér-Rao bound makes a number of regularity conditions which are difficult to check and the result only compares estimators satisfying these conditions (though see Borovkov, 1984, for a more satisfactory version). Finally, the fact that a limiting variance may not coincide with the variance of a limiting distribution produces another unpleasant gap between the bound and the limit theory (many interesting estimators even have infinite variance).

Of course the beautiful theory of Hájek and Le Cam (local asymptotic minimax theorem, the convolution theorem for regular estimators) solves all these problems in a mathematical sense. However the techniques used are very sophisticated and the notions involved are very delicate.

On the other hand, there is a theory of 'Best Asymptotically Normal' estimators due to Rao (1963) and others, but this is very restrictive in its applications and somewhat neglected nowadays.

Here we show that a simple variation on the Cramér-Rao theme due to van Trees (1968) provides the key-stone of a short and elementary proof that the variance of the limiting distribution of uniformly convergent-in-distribution estimators exceeds or equals the Cramér-Rao information bound. 'Uniformity' of some kind is of course needed to rule out super-efficiency (actually just Hájek regularity will do). The further regularity conditions involved are minimal.

We also give some other applications of van Trees' inequality, demonstrating its power and versatility, in particular as a tool for obtaining optimal convergence rates in non-regular (and non-parametric) problems, and for obtaining global bounds for estimating infinite dimensional parameters. In fact this is just the tip of the iceberg. The bound can also be used to investigate asymptotic admissibility and second order optimality (Levit and Oudshoorn, 1992; Schipper, 1992) and we believe it will find many other applications.

Previous applications of the van Trees inequality have been given by Bobrovsky, Mayer-Wolf and Zakai (1987) and Brown and Gajek (1990). A stronger type of inequality is given by Klaassen (1989); the right hand side of his (5.21) is the lower bound coming from the van Trees inequality.

2. van Trees' inequality.

Let $(\mathcal{X}, \mathcal{F}, \mathrm{P}_\theta : \theta \in \Theta)$ be a dominated family of distributions on some sample space $\mathcal{X}$; denote the dominating measure by μ. Take the parameter space Θ to be a closed interval on the real line. Let $f(x|\theta)$ denote the density of P_θ with respect to μ. Let π be some probability distribution on Θ with a density $\lambda(\theta)$ with respect to Lebesgue measure. Suppose that λ and $f(x|\cdot)$ are both absolutely continuous (μ-almost surely), and that λ converges to zero at the endpoints of the interval Θ. A prime will denote a partial derivative with respect to θ.

Let $\widehat{\theta} = \widehat{\theta}(X)$ denote any estimator of θ, $X \sim \mathrm{P}_\theta$. We write E_θ for expectation with respect to θ. When $\boldsymbol{\theta}$ is drawn from the distribution π, and conditional on $\boldsymbol{\theta} = \theta$, X from P_θ, we write E for expectation with respect to the ensuing joint distribution of X and $\boldsymbol{\theta}$.

Apart from the absolute continuity of f as function of θ, our last assumption is just the usual

$$\mathrm{E}_\theta \left(\log f(X|\theta) \right)' = 0. \tag{1}$$

Define further

$$\mathcal{I}(\theta) = \mathrm{E}_\theta \left(\log f(X|\theta) \right)'^{\,2},$$

$$\mathcal{I}(\lambda) = \mathrm{E} \left(\log \lambda(\boldsymbol{\theta}) \right)'^{\,2},$$

the Fisher information for θ and for a location parameter in λ respectively. We also often write $\mathcal{I}(\pi)$ for the latter quantity. A well known result of Hájek is that (1) follows from

continuity of $\mathcal{I}(\theta)$; in fact (1) holds almost everywhere if $\sqrt{\mathcal{I}(\theta)}$ is just locally integrable in θ, which is enough for our purposes. For full details on both these results, also in the multivariate case, see Borovkov (1984, §16 and §20); see also Borovkov and Sakhanenko, 1980). For our asymptotic bound in the next section we will need that $\mathcal{I}(\theta)$ is continuous in θ.

Now

$$\int (f(x|\theta)\lambda(\theta))' \mathrm{d}\theta = \Big[f(x|\theta)\lambda(\theta) \Big] = 0$$

by the convergence of λ to zero at the endpoints of Θ, while by partial integration and the same fact again

$$\int \theta (f(x|\theta)\lambda(\theta))' \mathrm{d}\theta = \Big[\theta f(x|\theta)\lambda(\theta) \Big] - \int f(x|\theta)\lambda(\theta)\mathrm{d}\theta = -\int f(x|\theta)\pi(\mathrm{d}\theta). \qquad (2)$$

Using both these equalities,

$$\int\int (\widehat{\theta}(x) - \theta)(f(x|\theta)\lambda(\theta))' \mathrm{d}\theta\mu(\mathrm{d}x) = \int\int f(x|\theta)\pi(\mathrm{d}\theta)\mu(\mathrm{d}x) = 1.$$

Cauchy-Schwarz now gives

$$\int\int (\widehat{\theta}(x) - \theta)^2 f(x|\theta)\pi(\mathrm{d}\theta)\mu(\mathrm{d}x) \cdot \int\int (\log(f(x|\theta)\lambda(\theta))'^2 f(x|\theta)\pi(\mathrm{d}\theta)\mu(\mathrm{d}x) \;\geq\; 1.$$

But by our assumption (1) the 'information' part of this expression, i.e., the second term in the product on the left hand side, reduces just to $\int \mathcal{I}(\theta)\pi(\mathrm{d}\theta) + \mathcal{I}(\lambda)$. Dividing out and abbreviating the notation gives the final inequality (van Trees, 1968 ; p. 72)

$$\mathrm{E}(\widehat{\theta}(X) - \boldsymbol{\theta})^2 \;\geq\; \frac{1}{\mathrm{E}\mathcal{I}(\boldsymbol{\theta}) + \mathcal{I}(\lambda)}. \qquad (3)$$

We emphasize that the only assumptions made here were (1) and the regularity conditions in the first paragraph of the Section.

A more general inequality for estimating an absolutely continuous function ψ of θ is easily obtained in exactly the same way. Replacing θ by $\psi(\theta)$, the equality (2) becomes

$$\begin{aligned}\int \psi(\theta)(f(x|\theta)\lambda(\theta))' \mathrm{d}\theta &= \Big[\psi(\theta) f(x|\theta)\lambda(\theta) \Big] - \int \psi'(\theta) f(x|\theta)\lambda(\theta)\mathrm{d}\theta \\ &= -\int \psi'(\theta) f(x|\theta)\pi(\mathrm{d}\theta).\end{aligned}$$

Replacing now also $\widehat{\theta}(x)$ by $\widehat{\psi}(x)$ in the subsequent development gives

$$\mathrm{E}(\widehat{\psi}(X) - \psi(\boldsymbol{\theta}))^2 \;\geq\; \frac{(\mathrm{E}\psi'(\boldsymbol{\theta}))^2}{\mathrm{E}\mathcal{I}(\boldsymbol{\theta}) + \mathcal{I}(\lambda)}. \qquad (4)$$

Example. *Estimation of θ^α from a sample from the $\mathcal{N}(\theta,1)$ distribution.* Let $X_1,\ldots,X_n$ be a random sample from the $\mathcal{N}(\theta,1)$ distribution, define $\psi(\theta)=\theta^\alpha$, where $\theta\geq 0$ and $0<\alpha<1$. We show that the optimal rate of convergence for estimators of ψ is only $n^{-\alpha/2}$ when θ can be arbitrarily close to zero. We apply (4) with $X_1,\ldots,X_n$ replacing X. The information for θ based on n independent and identically distributed observations is n times the information for one observation. By (4), for an arbitrary estimator ψ_n,

$$\sup_\theta \mathrm{E}_\theta(\psi_n-\psi(\theta))^2 \geq \mathrm{E}(\psi_n-\psi(\boldsymbol{\theta}))^2 \geq \frac{(\alpha\int\theta^{\alpha-1}\lambda(\theta)\mathrm{d}\theta)^2}{n+\mathcal{I}(\lambda)}, \tag{5}$$

where λ is a prior density on a closed bounded interval of $[0,\infty)$, satisfying the conditions stated above. Setting $\lambda(\theta)=a^{-1}\lambda_0(a^{-1}\theta)$ and denoting

$$A=\Big(\int u^{\alpha-1}\lambda_0(u)\mathrm{d}u\Big)^2$$

we obtain from (5)

$$\sup_\theta \mathrm{E}_\theta(\psi_n-\psi(\theta))^2 \geq \frac{Aa^{2(\alpha-1)}\alpha^2}{n+\mathcal{I}(\lambda_0)a^{-2}}.$$

Choosing $a=(\mathcal{I}(\lambda_0)\alpha/((1-\alpha)n))^{1/2}$ so that for large n the prior λ concentrates more and more mass close to 0, we obtain

$$\sup_\theta \mathrm{E}_\theta(\psi_n-\psi(\theta))^2 \geq \frac{(1-\alpha)^{1-\alpha}\alpha^{2+\alpha}A}{n^\alpha\mathcal{I}(\lambda_0)^{1-\alpha}}.$$

This inequality shows that no estimator can have a convergence rate (uniformly) better than $n^{-\alpha/2}$. Fortunately a natural choice for ψ_n, $|\overline{X}_n|^\alpha$ where $\overline{X}_n$ is the sample mean, achieves this bound. This follows by applying the c_r-inequality (see Loève (1963), p. 155):

$$\big||\overline{X}_n|^\alpha-|\theta|^\alpha\big|\leq|\overline{X}_n-\theta|^\alpha,\qquad 0<\alpha<1,$$

from which follows

$$\sup_\theta \mathrm{E}_\theta\big(|\overline{X}_n|^\alpha-|\theta|^\alpha\big)^2\leq \mathrm{E}|\overline{X}_n-\theta|^{2\alpha}=\frac{2^\alpha\Gamma(\alpha+\frac{1}{2})}{\sqrt{\pi}n^\alpha}$$

by straightforward calculation.

Despite our lower bound on the (maximal) mean square error, the estimator ψ_n actually converges in distribution at the faster (usual) rate of $n^{-1/2}$, provided we keep away from $\theta=0$. This case is covered by our asymptotic results in the next section. □

3. An asymptotic Cramér-Rao bound.

Suppose in the previous section, P_θ is replaced by its n-fold product, and X by $\mathbf{X}^{(n)}$, n i.i.d. copies of X. The information for θ then gets multiplied by n. Let π_1 be a fixed distribution on $[-1,1]$ with absolutely continuous density, zero at the end-points. Let θ_0 be a fixed point in the interior of Θ and let $\pi = \pi(H,n)$ be the rescaling of π_1 to the interval $A = [\theta_0 - n^{-\frac{1}{2}}H, \theta_0 + n^{-\frac{1}{2}}H]$ for given $H > 0$. The information for $\pi(H,n)$ is that for π_1 times n/H^2. If $\widehat{\theta}^{(n)}$ is any estimator for θ based on $\mathbf{X}^{(n)}$, the inequality (3) becomes

$$\mathrm{E}(\widehat{\theta}^{(n)} - \boldsymbol{\theta})^2 \;\geq\; \frac{1}{n\mathrm{E}_{\pi(H,n)}\mathcal{I}(\boldsymbol{\theta}) + n\mathcal{I}(\pi_1)/H^2}$$

or

$$\mathrm{E}\left(\sqrt{n}(\widehat{\theta}^{(n)} - \boldsymbol{\theta})\right)^2 \;\geq\; \frac{1}{\mathrm{E}_{\pi(H,n)}\mathcal{I}(\boldsymbol{\theta}) + \mathcal{I}(\pi_1)/H^2}. \tag{6}$$

We assume that $\mathcal{I}(\theta)$ is continuous at θ_0. Then letting first $n \to \infty$ and then $H \to \infty$ the right hand side converges to $1/\mathcal{I}(\theta_0)$, the usual asymptotic information bound.

Now suppose $\sqrt{n}(\widehat{\theta}^{(n)} - \theta)$ converges in distribution, as $n \to \infty$, uniformly in θ; suppose the limiting distribution is also continuous in θ at θ_0. Then $\sqrt{n}(\widehat{\theta}^{(n)} - \theta_n)$ converges in distribution under P_{θ_n}, to a fixed distribution Z say, for all sequences θ_n of the form $\theta_0 + n^{-\frac{1}{2}}h$, $h \in (-H, H)$. (Alternatively one can just make this, weaker, assumption of Hájek regularity). This is the limiting distribution of $\sqrt{n}(\widehat{\theta}^{(n)} - \theta_0)$ under P_{θ_0}, in which we are interested.

Truncate $\widehat{\theta}^{(n)}$ to the interval A. This is the same as truncating $\sqrt{n}(\widehat{\theta}^{(n)} - \theta_n)$ to the interval $[-H - h, H - h]$. The resulting random variable converges in distribution under P_{θ_n} to Z truncated to the same interval, call this $Z_{h,H}$. Since both are bounded, the mean square converges too. Moreover, we have $\mathrm{E}Z_{h,H}^2 \leq \mathrm{E}Z^2$. Applying (6) to the truncated estimator and letting $n \to \infty$, together this gives

$$\mathrm{E}Z^2 \geq \int_{-H}^{H} \mathrm{E}Z_{h,H}^2 \pi_H(\mathrm{d}h) \geq \frac{1}{\mathcal{I}(\theta_0) + \mathcal{I}(\pi_1)/H^2}$$

where π_H is the rescaling of π_1 to the interval $[-H, H]$. Now let $H \to \infty$, and we obtain $\mathrm{E}Z^2 \geq 1/\mathcal{I}(\theta_0)$; in words, the mean square of the limiting distribution of $\sqrt{n}(\widehat{\theta}^{(n)} - \theta_0)$ is at least $1/\mathcal{I}(\theta_0)$. (More general version of this elementary result is known as the Hájek's convolution theorem; see Hájek (1970). For a different approach see Klaassen (1989 ; Theorem 4.1)).

We can improve this bound on the mean square error of the asymptotic distribution to the same bound on the variance of the asymptotic distribution. If the limiting distribution of $\sqrt{n}(\widehat{\theta}^{(n)} - \theta)$ has mean $a(\theta)$, simply apply the above to the new 'estimator' $\widehat{\theta}^{(n)} - n^{-\frac{1}{2}}a(\theta_0)$. Its asymptotic distribution under P_{θ_n} has mean zero but the same variance as before.

Obviously we can also obtain analogous results for estimating a function of θ.

In the next section we will derive genuine multidimensional versions of the van Trees inequalities (3) and (4). For the moment notice that one can obtain a multidimensional

version of the just obtained asymptotic bound by considering certain one-dimensional submodels, satisfying (by assumption) the regularity conditions given above. Suppose θ is now a (column) vector and the matrix $\mathcal{I}(\theta)$ exists and is non-singular. Write $\mathcal{I}_0 = \mathcal{I}(\theta_0)$. Apply the preceding to estimation of the linear combination $c^\top\theta$ in the one-dimensional submodel $\theta = \theta_0 + \eta(c^\top\mathcal{I}_0^{-1}c)^{-1}\mathcal{I}_0^{-1}c$ where η is a real parameter. The information for $\eta = c^\top(\theta - \theta_0)$ at $\eta = 0$ is $(c^\top\mathcal{I}_0^{-1}c)^{-1}$. The resulting bound $c^\top\mathcal{I}_0^{-1}c$, holding for all c, on the variance of the asymptotic distribution of $c^\top\sqrt{n}(\widehat{\theta}^{(n)} - \theta_0)$, implies the bound $\mathcal{I}_0^{-1}$ on the covariance matrix of the asymptotic distribution of $\sqrt{n}(\widehat{\theta}^{(n)} - \theta_0)$.

4. Multivariate extensions.

Curiously there are many ways to extend the van Trees inequality to higher dimensional parameters (cf. van Trees (1968, p. 84) ; Bobrovsky, Mayer-Wolf and Zakai (1987)). Here we present a very general version involving arbitrary choices of certain matrix weight functions. The inequality allows several interesting special cases.

We consider immediately the case of estimating a possibly vector valued function ψ of a vector parameter θ, respectively p and s dimensional. Suppose also from the start that we have n independent and identically distributed observations X_i from a common distribution P_θ with density $f(x,\theta)$ with respect to some measure μ (all on an arbitrary measure space $\mathcal{X}$). Suppose $\theta \in \Theta \subseteq \mathbb{R}^s$. Write X for a generic observation X_i.

We choose next a prior density (with respect to Lebesgue measure) $\lambda(\theta)$, a symmetric $p \times p$ matrix function $B(\theta)$, and a $p \times s$ matrix function $C(\theta)$. We need a number of regularity conditions on f, λ, B and C.

We will say that a real function $g(\theta)$, $\theta \in \Theta$, is *nice* if, for each j, it is absolutely continuous in θ_j for almost all values of the other components of θ and its partial derivatives $\partial g/\partial\theta_j$ are measurable in θ. We will treat θ and ψ as column vectors; partial derivatives with respect to the components of θ are set out in rows so $\partial\psi/\partial\theta$ is a $p \times s$ matrix. The symbol $\top$ denotes transpose of a matrix. The operation 'diag' of a square matrix replaces its off-diagonal elements by zeros.

Assumptions:

1) f is nice in θ for almost all x and its partial derivatives with respect to θ are measurable in x, θ.
2) The Fisher information matrix

$$\mathcal{I}(\theta) = \mathrm{E}_\theta\left(\left(\frac{\partial \log f(X,\theta)}{\partial\theta}\right)^\top \frac{\partial \log f(X,\theta)}{\partial\theta} \right)$$

   exists and $\mathrm{diag}(\mathcal{I}(\theta))^{1/2}$ is locally integrable in θ.
3) The components of ψ and C are nice.
4) B is positive definite; suppose $B(\theta) = A(\theta)^\top A(\theta)$ for a $p \times p$ matrix $A(\theta)$.
5) λ is nice; Θ is compact with boundary which is piecewise C^1-smooth; λ is positive on the interior of Θ and zero on its boundary.

Assumptions 1) and 2) imply (Borovkov, 1984, Borovkov and Sakhanenko, 1980) that the expected score vector is zero for almost all θ:

$$\mathrm{E}_\theta \frac{\partial \log f(X,\theta)}{\partial \theta} = 0.$$

Theorem 1. (multivariate van Trees inequality). *Under conditions* 1)–5)*, for any estimator* ψ_n

$$\begin{aligned} R(\psi_n,\lambda) &= \int_\Theta \mathrm{E}_\theta(\psi_n-\psi(\theta))^\top B(\theta)^{-1}(\psi_n-\psi(\theta))\lambda(\theta)\mathrm{d}\theta \\ &\geq \frac{\Big(\int_\Theta \mathrm{trace}\big(C(\theta)(\partial\psi/\partial\theta)^\top\big)\lambda(\theta)\mathrm{d}\theta\Big)^2}{n\int_\Theta \mathrm{trace}\big(B(\theta)^\top C(\theta)\mathcal{I}(\theta)C(\theta)^\top\big)\lambda(\theta)\mathrm{d}\theta + \widetilde{\mathcal{I}}(\lambda)} \end{aligned} \tag{7}$$

where

$$\widetilde{\mathcal{I}}(\lambda) = \int_\Theta \Bigg(\sum_{i,j,k,l} B_{ij}(\theta)\frac{\partial}{\partial\theta_k}\big(C_{ik}(\theta)\lambda(\theta)\big)\frac{\partial}{\partial\theta_l}\big(C_{jl}(\theta)\lambda(\theta)\big)\Bigg)\frac{1}{\lambda(\theta)}\mathrm{d}\theta. \tag{8}$$

Proof. For random p-vectors X, Y, we define the scalar product

$$\langle X,Y\rangle = \mathrm{E}X^\top Y.$$

This makes the space of square-integrable random vectors a Hilbert space and we have the Cauchy-Schwarz inequality

$$\langle X,Y\rangle^2 \leq \langle X,X\rangle\langle Y,Y\rangle.$$

Since $B = A^\top A$ for non-singular A we can also apply this to $U = (A^{-1})^\top X$ and $V = AY$ getting

$$(\mathrm{E}X^\top Y)^2 = (\mathrm{E}U^\top V)^2 \leq (\mathrm{E}U^\top U)(\mathrm{E}V^\top V) = (\mathrm{E}X^\top B^{-1}X)(\mathrm{E}Y^\top BY). \tag{9}$$

Essentially without loss of generality we prove the required result in the case $n = 1$ (after that simply apply it to the joint density of the n observations, by which $\mathcal{I}(\theta)$ becomes $n\mathcal{I}(\theta)$ thanks to the fact that the expected score is zero).

As before the proof goes by Cauchy-Schwarz, in the form (9), and integration by parts. Take for X and Y

$$X = \psi_n - \psi(\theta),$$

$$Y_i = \sum_{j=1}^{s}\frac{\partial}{\partial\theta_j}\big(C_{ij}(\theta)f(\theta)\lambda(\theta)\big)\frac{1}{f(\theta)\lambda(\theta)},$$

where $f(\theta) = f(\cdot,\theta)$ is the likelihood for θ. In (9) we find

$$\mathrm{E}X^\top B^{-1}X = R(\psi_n,\lambda),$$

and, using integration by parts,

$$\begin{aligned}
\mathrm{E}X^\top Y &= \int_{\mathcal{X}}\int_\Theta \sum_i (\psi_n(x)-\psi(\theta))_i \sum_{k=1}^s \frac{\partial}{\partial\theta_k}\big(C_{ik}(\theta)f(x,\theta)\lambda(\theta)\big)\,\mathrm{d}\theta\,\mu(\mathrm{d}x)\\
&= \int_\Theta\int_{\mathcal{X}} \sum_{ik} \frac{\partial\psi_i(\theta)}{\partial\theta_k} C_{ik}(\theta)f(x,\theta)\lambda(\theta)\mu(\mathrm{d}x)\mathrm{d}\theta\\
&= \int_\Theta \mathrm{trace}\Big(C(\theta)\big(\frac{\partial\psi}{\partial\theta}\big)^\top\Big)\lambda(\theta)\mathrm{d}\theta.
\end{aligned}$$

Finally

$$\begin{aligned}
\mathrm{E}Y^\top BY &= \mathrm{E}\sum_{ijkl} B_{ij}\frac{C_{ik}\lambda\frac{\partial f}{\partial\theta_k}+f\frac{\partial(C_{ik}\lambda)}{\partial\theta_k}}{f\lambda}\frac{C_{jl}\lambda\frac{\partial f}{\partial\theta_l}+f\frac{\partial(C_{jl}\lambda)}{\partial\theta_l}}{f\lambda}\\
&= \mathrm{E}\sum_{ijkl} B_{ij}\left(C_{ik}\frac{\partial}{\partial\theta_k}\log f+\frac{1}{\lambda}\frac{\partial}{\partial\theta_k}(C_{ik}\lambda)\right)\left(C_{jl}\frac{\partial}{\partial\theta_l}\log f+\frac{1}{\lambda}\frac{\partial}{\partial\theta_l}(C_{jl}\lambda)\right)\\
&= \int_\Theta \mathrm{trace}\big(B(\theta)^\top C(\theta)\mathcal{I}(\theta)C(\theta)^\top\big)\lambda(\theta)\mathrm{d}\theta+\widetilde{\mathcal{I}}(\lambda)
\end{aligned}$$

using 'expected score is zero' to get rid of the cross products. □

Various natural choices of B and C lead to versions of (7) useful for different purposes. The simplest choice, when $p=s$ so that ψ has the same dimension as θ, is to take both matrix functions equal to the identity. Using $|\cdot|$ to stand for the Euclidean norm of vectors and writing $\operatorname{div}\psi=\sum_i \partial\psi_i/\partial\theta_i=\operatorname{trace}\partial\psi/\partial\theta$, we obtain

$$\int_\Theta \mathrm{E}_\theta|\psi_n-\psi(\theta)|^2\lambda(\theta)\mathrm{d}\theta \;\ge\; \frac{\Big(\int \operatorname{div}\psi(\theta)\lambda(\theta)\mathrm{d}\theta\Big)^2}{n\int\operatorname{trace}\mathcal{I}(\theta)\lambda(\theta)\mathrm{d}\theta+\operatorname{trace}\mathcal{I}(\lambda)} \tag{10}$$

where $\mathcal{I}(\lambda)$ is the matrix information for λ. We call (10) an *L_2-norm type inequality*. The one-dimensional case with ψ the identity is our first version of the van Trees inequality.

In some respects more natural, and also available when $p\neq s$, is still to take B the identity but to choose

$$C(\theta)=\frac{\partial\psi}{\partial\theta}\mathcal{I}(\theta)^{-1},$$

$\mathcal{I}(\theta)$ supposed invertible. Define also

$$\mathcal{J}_\psi(\theta)=\frac{\partial\psi}{\partial\theta}\mathcal{I}(\theta)^{-1}\frac{\partial\psi}{\partial\theta}^\top,$$

the inverse of which one could call the information for ψ. We now find from (7)

$$\int_\Theta \mathrm{E}_\theta|\psi_n-\psi(\theta)|^2\lambda(\theta)\mathrm{d}\theta \geq \frac{\left(\int \mathrm{trace}\mathcal{J}_\psi(\theta)\lambda(\theta)\mathrm{d}\theta\right)^2}{n\int \mathrm{trace}\mathcal{J}_\psi(\theta)\lambda(\theta)\mathrm{d}\theta+\widetilde{\mathcal{I}}(\lambda)} \geq \frac{1}{n}\int \mathrm{trace}\mathcal{J}_\psi(\theta)\lambda(\theta)\mathrm{d}\theta-\frac{1}{n^2}\widetilde{\mathcal{I}}(\lambda) \tag{11}$$

where

$$\widetilde{\mathcal{I}}(\lambda)=\int\sum_{ikl}\frac{1}{\lambda(\theta)}\left(\frac{\partial}{\partial\theta_k}C_{ik}(\theta)\lambda(\theta)\right)\left(\frac{\partial}{\partial\theta_l}C_{il}(\theta)\lambda(\theta)\right)\mathrm{d}\theta. \tag{12}$$

For $s=1$ this contains the so called Borovkov-Sakhanenko inequality; see Borovkov and Sakhanenko (1980), Theorem 3. We call (11) the *natural multivariate van Trees inequality*.

Apart from the identity an alternative natural choice for B is the inverse information for ψ itself, $B(\theta)=\mathcal{J}_\psi(\theta)$. Generalising the choice of C for our natural inequality (11), i.e.,

$$C(\theta)=B(\theta)^{-1}\frac{\partial\psi}{\partial\theta}\mathcal{I}(\theta)^{-1},$$

we find the *normalized risk inequality*

$$\int \mathrm{E}_\theta\Big((\psi_n-\psi(\theta))^\top\mathcal{J}_\psi(\theta)^{-1}(\psi_n-\psi(\theta)\Big)\lambda(\theta)\mathrm{d}\theta \geq \frac{p^2}{np+\widetilde{\mathcal{I}}(\lambda)} \geq \frac{p}{n}-\frac{\widetilde{\mathcal{I}}(\lambda)}{n^2} \tag{13}$$

where $\widetilde{\mathcal{I}}(\lambda)$ is as in (12).

Finally, the *weighted quadratic risk inequality* is obtained when $B(\theta)=\mathrm{diag}(\mathcal{J}_\psi(\theta))$ and $C(\theta)=B(\theta)^{-1}(\partial\psi/\partial\theta)\mathcal{I}(\theta)^{-1}$, resulting in

$$\int \mathrm{E}_\theta\sum_{i=1}^p\frac{\big(\psi_{ni}-\psi_i(\theta)\big)^2}{(\mathcal{J}_\psi)_{ii}}\lambda(\theta)\mathrm{d}\theta \geq \frac{p^2}{np+\widetilde{\mathcal{I}}(\lambda)} \geq \frac{p}{n}-\frac{\widetilde{\mathcal{I}}(\lambda)}{n^2}, \tag{14}$$

where $\widetilde{\mathcal{I}}(\lambda)$ is given by (12) for the present choice of B and C.

Inequality (10) is sufficent for first and second order, local investigations; see Section 6 and Schipper (1992). Inequalities (11)–(14) can be used for first and second order global considerations; see, e.g., Levit and Oudshoorn (1992).

5. Semiparametric models: rates of convergence

First we consider some non- or semi-parametric problems where the optimal rate of convergence is not the square root of n but something smaller. In each case we use the univariate van Trees inequality to derive an asymptotic lower bound to the maximum mean square error of an estimator over a small neighbourhood of the parameter. The proof works by guessing a most difficult parametric submodel for estimating the functional of interest. The procedure sometimes suggests ad hoc estimators which achieve the lower bound, thereby demonstrating its optimality (at least as far as the rate is concerned). The proof can also be adapted to give a lower bound to the variance of the asymptotic distribution of a uniformly convergent-in-distribution estimator, just as in the regular parametric case studied in the previous example.

The examples studied here have been considered by Bakker (1988), Weits (1992), and Groeneboom and Wellner (1992), using other methods. They compare the constants appearing in their lower bounds to those obtainable by the estimators under consideration.

The first example, 'completely censored data', occurs in biostatistical applications (carcinogenicity experiments) in which one is interested in the distribution of the time of a certain event. At another random 'observation time' one can see whether or not the event has already occurred (e.g., animal sacrifice to determine whether or not a tumour is already present).

So, let $X_1 \ldots, X_n$ be i.i.d. with unknown d.f. F and let $Y_1, \ldots, Y_n$ be independent of the X_i's and i.i.d. with density g. The data consists of the pairs $Z_i = (\Delta_{i1}, Y_i)$, $i = 1, \ldots, n$, where

$$\Delta_{i1} = 1\{X_i \leq Y_i\};$$

define also

$$\Delta_{i2} = 1 - \Delta_{i1} = 1\{X_i > Y_i\}.$$

The target functional of F to be estimated is

$$\psi(F) = F(x_0),$$

x_0 fixed. We consider g also fixed; whether it is known or not is irrevelant. Let $T_n = T_n(Z_1, \ldots, Z_n)$ denote an estimator of $F(x_0)$. We make the following two

Assumptions:

1) $F \in \mathcal{F} = \Bigg\{ F : \ 0 < F(x_0) < 1$ and $\exists\ \gamma_1(F),\ \gamma_2(F),\ \gamma_3(F) > 0$ such that

$$\gamma_1(F) < \frac{F(x_1) - F(x_2)}{x_1 - x_2} < \gamma_2(F)$$

$$\text{for all } x_0 - \gamma_3(F) < x_1 < x_2 < x_0 + \gamma_3(F) \Bigg\}.$$

2) $g(x_0) > 0$, g is continuous at x_0.

On $\mathcal{F}$ we introduce a topology $\mathcal{T}$ generated by the neighbourhoods

$$V_{\varepsilon\delta}(F) = \left\{G \in \mathcal{F} : \ \|G - F\|_{(x_0-\varepsilon, x_0+\varepsilon)} < \delta\right\}$$

where

$$\|F\|_C = \sup_{x\in C} |F(x)| + \sup_{x_1,x_2\in C} \frac{|F(x_1) - F(x_2)|}{|x_1 - x_2|}$$

Theorem 2. *Let $F_0 \in \mathcal{F}$ and $V = V_{\varepsilon\delta}(F_0)$ be fixed. Then there exists $C(V) > 0$ and $n_0 < \infty$ such that:*

$$r_n(V) = \inf_{T_n} \sup_{F\in V} \mathrm{E}_F (T_n - F(x_0))^2 \geq \frac{c(V)}{n^{2/3}}, \quad n \geq n_0$$

where the infimum is taken over all possible estimators T_n.

Proof. Let ψ satisfy:

$$\psi(x) = 0, \quad |x| \geq 1$$

$$\psi(0) = 1$$

$$\|\psi\|_{(-1,1)} \leq 1.$$

Choose $B > 0$ fixed and let, for given $h > 0$, $\{F_c(\cdot)\}$ denote the family

$$F_c(x) = F_0(x) + c\psi\left(\frac{x - x_0}{h}\right), \quad |c| < Bh. \tag{15}$$

B can be chosen so that for small enough h,

$$\{F_c : |c| < Bh\} \subseteq V \subseteq \mathcal{F}; \tag{16}$$

and of course we have

$$F_c(x_0) = F_0(x_0) + c.$$

For use in the proof of Theorem 3 below, note that if moreover $B < \gamma_1(F_0)$ then $\gamma_1(F_c)$ is uniformly bounded away from zero (in c and h, for sufficiently small h).

Let ν be counting measure on $\{0, 1\}$ and let $f_c(\delta_1, \delta_2, y)$ be the density of $(\Delta_{i1,}, \Delta_{i2}, Y_i)$ with respect to $\nu \times \nu \times$ Lebesgue measure;

$$f_c(\delta_1, \delta_2, y) = F_c(y)^{\delta_1} \left(1 - F_c(y)\right)^{\delta_2} g(y)$$

for $\delta_2 = 1 - \delta_1 \in \{0, 1\}$; zero otherwise. The score l'_c is given by

$$l'_c = \frac{\partial \log f_c}{\partial c} = \frac{\delta_1 \psi\left(\frac{y-x_0}{h}\right)}{F_c(y)} - \frac{\delta_2 \psi\left(\frac{y-x_0}{h}\right)}{1 - F_c(y)}$$

and the conditional information given Y_i is

$$\begin{aligned}
\mathrm{E}(l_c'^2 \,|\, Y_i = y) &= \frac{\psi^2(\frac{y-x_0}{h})}{F_c(y)} + \frac{\psi^2(\frac{y-x_0}{h})}{1-F_c(y)} \\
&= \frac{\psi^2(\frac{y-x_0}{h})}{F_c(y)(1-F_c(y))} \\
&= \frac{\psi^2(\frac{y-x_0}{h})}{F_0(y)(1-F_0(y))}(1+o(1)) \quad \text{as} \quad h \to 0, \quad \text{uniformly in} \quad y.
\end{aligned}$$

The unconditional information is therefore

$$\begin{aligned}
\mathcal{I}(c) = \mathrm{E}l_c'^2 &= (1+o(1)) \int_{x_0-h}^{x_0+h} \frac{\psi^2\left(\frac{y-x_0}{h}\right)}{F_0(y)\,(1-F_0(y))} g(y)\,\mathrm{d}y \\
&= \frac{g(x_0)}{F_0(x_0)\,(1-F_0(x_0))} \cdot \int_{-1}^{1} \psi^2(t)\,\mathrm{d}t \cdot h \cdot (1+o(1)) \\
&= Ah\,(1+o(1))\,,
\end{aligned}$$

say, using assumptions 1) and 2).

Let λ_0 be a prior density on $(-1,1)$ satisfying the conditions of the van Trees inequality and let

$$\begin{aligned}
\mathcal{I}_0 &= \int_{-1}^{1} \frac{\lambda_0'(c)^2}{\lambda_0(c)}\,\mathrm{d}c, \\
\lambda(c) &= \frac{1}{Bh}\lambda_0\left(\frac{c}{Bh}\right).
\end{aligned}$$

Then

$$\begin{aligned}
r_n(V) &\geq \inf_{T_n} \sup_{|c|<Bh} \mathrm{E}_{F_c}\left(T_n - F_c(x_0)\right)^2 \\
&= \inf_{T_n} \sup_{|c|<Bh} \mathrm{E}_{F_c}\left((T_n - F_0(x_0)) - c\right)^2 \\
&\geq \inf_{T_n'} \int_{-Bh}^{Bh} \mathrm{E}_{F_c}(T_n' - c)^2 \lambda(c)\,\mathrm{d}c \\
&\geq \left(n \int_{-Bh}^{Bh} \mathcal{I}(c)\lambda(c)\,\mathrm{d}c + \int_{-Bh}^{Bh} \frac{\lambda'(c)^2}{\lambda(c)}\,\mathrm{d}c \right)^{-1} \\
&\qquad\qquad \text{by van Trees' inequality} \\
&= \left(nAh\,(1+o(1)) + I_0/(B^2h^2)\right)^{-1} \\
&= n^{-2/3}\left(A + \mathcal{I}_0/B^2\right)^{-1}(1+o(1))
\end{aligned}$$

if we take $h = n^{-1/3}$. □

The rate $n^{-1/3}$ is known to be achieved in this problem by the non-parametric maximum likelihood estimator (NPMLE) of F; see Groeneboom and Wellner (1992). They

make slightly stronger assumptions than ours, but on the other hand derive the limiting distribution of the NPMLE. The lower bound is given in Groeneboom (1987) using a different approach. The problem is very similar to that of estimating a density, known to be monotone, for which the rate $n^{-1/3}$ also applies; see Groeneboom (1985).

A simple ad hoc estimator of $F(x_0)$ achieving this rate is given by the following modification of a histogram estimator; in fact a histogram with appropriate bin-width is also an optimal rate density estimator under weak smoothness conditions.

Let

$$\begin{aligned}\xi_i &= 1\{Y_i \in (x_0 - h, x_0 + h)\},\\ \eta_i &= 1\{Y_i \in (x_0 - h, x_0 + h), X_i \le Y_i\},\\ F_n(x_0) &= \frac{\sum_{i=1}^n \eta_i}{\sum_{i=1}^n \xi_i}\end{aligned}$$

where $h = h_n > 0$,

$$h_n \to 0 \qquad \text{and} \qquad nh_n^{2+\delta} \to \infty$$

as $n \to \infty$ for some $\delta > 0$. We show that under our assumptions 1) and 2), uniformly on any neighbourhood $V_{\varepsilon\delta}(F_0)$ with $F_0 \in \mathcal{F}$,

$$\mathrm{E}_F(F_n(x_0) - F(x_0))^2 \le \Big(\frac{F(x_0)(1 - F(x_0))}{2g(x_0)hn} + (\gamma_2(F)h)^2\Big)(1 + o(1))$$

as $n \to \infty$. From this, choosing $h \asymp n^{-1/3}$ results in the optimal rate

$$\mathrm{E}_F(F_n(x_0) - F(x_0)) \asymp n^{-2/3}.$$

To demonstrate this, let

$$\begin{aligned}\alpha &= \mathrm{E}_F\xi_1 = \mathrm{E}_F\xi_1^2\\ \beta &= \mathrm{E}_F\eta_1 = \mathrm{E}_F\eta_1^2.\end{aligned}$$

One obtains then

$$\begin{aligned}\alpha &= \int_{x_0-h}^{x_0+h} g(y)\mathrm{d}y = 2g(x_0)h(1 + o(1)),\\ \beta &= \int_{x_0-h}^{x_0+h} F(y)g(y)\mathrm{d}y = (F(x_0) + \gamma)\alpha,\end{aligned}$$

say, where for sufficiently large n (and hence small enough h) we have $\alpha > 0$ and

$$|\gamma| \le \gamma_2(F)h.$$

All these arguments carry through uniformly on the given neighbourhood $V_{\varepsilon\delta}$. Denoting further

$$\begin{aligned}\overline{\xi}_n &= \frac{1}{\alpha n}\sum_{i=1}^n(\xi_i - \alpha)\\ \overline{\eta}_n &= \frac{1}{\alpha n}\sum_{i=1}^n(\eta_i - \beta)\end{aligned}$$

a simple calculation gives

$$F_n(x_0) - F(x_0) = \frac{\overline{\eta}_n - F(x_0)\overline{\xi}_n + \gamma}{1 + \overline{\xi}_n}.$$

Let $\rho = \rho_n = (nh^2)^{-1/4}$ and define the event $A = A_n = \{|\overline{\xi}_n| < \rho_n\}$. Thus $\rho_n \to 0$, and by Hoeffding's inequality (see, e.g., Pollard, 1984, p. 192)

$$\mathrm{P}(A_n^c) \le 2\exp\left(-\frac{n\alpha^2\rho^2}{2}\right) = O\left((nh^2)^{-(1+\delta)/\delta}\right) = o\big((nh)^{-1}\big)$$

for $n \to \infty$, by our assumptions on h_n and ρ_n. Notice also that

$$\mathrm{E}_F|(\overline{\eta}_n - F(x_0)\overline{\xi}_n)\gamma|1_{A_n^c} \le 2\alpha^{-1}|\gamma|\mathrm{P}(A_n^c) = o\big((nh)^{-1}\big).$$

Therefore for $n \to \infty$

$$\begin{aligned}
E_F(F_n(x_0) - F(x_0))^2 &= \mathrm{E}_F(F_n(x_0) - F(x_0))^2 1_{A_n} + o((nh)^{-1}) \\
&= (\mathrm{E}_F(\overline{\eta}_n - F(x_0)\overline{\xi}_n)^2 1_{A_n} + \gamma^2)(1 + O(\rho_n)) + o((nh)^{-1}) \\
&\le (\mathrm{E}_F(\overline{\eta}_n - F(x_0)\overline{\xi}_n)^2 + \gamma^2)(1 + o(1)) + o((nh)^{-1}) \\
&\le (n^{-1}\alpha^{-2}\mathrm{E}_F(\eta_1 - F(x_0)\xi_1)^2 + \gamma^2)(1 + o(1)) + o((nh^{-1})) \\
&= (n^{-1}\alpha^{-2}\mathrm{E}_F(\eta_1^2 - 2F(x_0)\eta_1\xi_1 + F(x_0)^2\xi_1^2) + \gamma^2)(1 + o(1)) + o((nh)^{-1}) \\
&= (n^{-1}\alpha^{-1}(F(x_0)(1 - F(x_0)) + (1 - 2F(x_0))\gamma) + \gamma^2))(1 + o(1)) + o((nh^{-1})) \\
&= \Big(\frac{F(x_0)(1 - F(x_0))}{2g(x_0)nh} + \gamma^2\Big)(1 + o(1))
\end{aligned}$$

which is the required result.

This estimator does have two disadvantages compared to the NPMLE: one must make an arbitrary choice of the constant in the bin-width; and the estimator of $F(x_0)$ is not a monotone function of x_0 .

Our next example is a modification of the first one: now there are two 'observation times' instead of one. Most of the information for estimating $F(x_0)$ comes from those observations for which the two observation times are close to each other and on each side of x_0, and when the event of interest occurs between these two times. This leads to a different (higher) optimal rate of convergence provided the joint density of the observation times is positive at (x_0, x_0).

Let $X_1, .., X_n$ be i.i.d. with d.f. F as in Example 1; let $\psi(F)$, $\mathcal{F}$ and $\mathcal{T}$ be as in that example. Let (Y_{i1}, Y_{i2}), $i = 1, \ldots, n$, be i.i.d. pairs independent of the X_i's with joint density g and with $Y_{i1} < Y_{i2}$. The data consists of $Z_1, \ldots, Z_n$ with

$$Z_i = (\Delta_{i1}, \Delta_{i2}, Y_{i1}, Y_{i2}),$$

where

$$\Delta_{i1} = 1\{X_i \le Y_{i1}\},$$
$$\Delta_{i2} = 1\{Y_{i1} < X_i \le Y_{i2}\},$$

define also

$$\Delta_{i3} = 1\{Y_{i2} < X_i\}.$$

Assumption 2) is replaced by:

2′) $g(y_1, y_2)$ is positive and continuous at (x_0, x_0); its marginals g_1 and g_2 are also positive and continous at x_0.

Theorem 3. *For all $V = V_{\varepsilon\delta}(F_0)$ there exists $c(V) > 0$ and n_0 such that*

$$r_n(V) = \inf_{T_n} \sup_{F \in V} \mathrm{E}_F (T_n - F(x_0))^2 \ge \frac{c(V)}{(n \log n)^{2/3}}, \quad n \ge n_0.$$

Proof. We define the family $\{F_c(\cdot) : |c| < Bh\}$ as in (15)–(16). B is taken so that the positive constants $\gamma_1(F_c)$, $\gamma_2(F_c)$ and $\gamma_3(F_c)$ can be chosen uniformly in c for sufficiently small h, as we mentioned in the proof of theorem 2. The joint density of $(\Delta_{i1}, \Delta_{i2}, \Delta_{i3}, Y_{i1}, Y_{i2})$ is now

$$f_c = F_c(y_1)^{\delta_1} \left(F_c(y_2) - F_c(y_1)\right)^{\delta_2} \left(1 - F_c(y_2)\right)^{\delta_3} g(y_1, y_2)$$

and the score is

$$l_c' = \frac{\delta_1 \psi\left(\frac{y_1 - x_0}{h}\right)}{F_c(y_1)} + \frac{\delta_2 \left(\psi\left(\frac{y_2 - x_0}{h}\right) - \psi\left(\frac{y_1 - x_0}{h}\right)\right)}{F_c(y_2) - F_c(y_1)} - \frac{\delta_3 \psi\left(\frac{y_2 - x_0}{h}\right)}{1 - F_c(y_2)}.$$

The conditional information for c given $(Y_{i1}, Y_{i2}) = (y_1, y_2)$ is

$$\frac{\psi^2\left(\frac{y_1 - x_0}{h}\right)}{F_c(y_1)} + \frac{\left(\psi\left(\frac{y_2 - x_0}{h}\right) - \psi\left(\frac{y_1 - x_0}{h}\right)\right)^2}{F_c(y_2) - F_c(y_1)} + \frac{\psi^2\left(\frac{y_2 - x_0}{h}\right)}{1 - F_c(y_2)}.$$

The unconditional information $\mathcal{I}(c)$ can be correspondingly split into three terms, $I(c) = I_1(c) + I_2(c) + I_3(c)$.

By the same calculations as in Theorem 2,

$$I_1(c) + I_3(c) = \left(\frac{g_1(x_0)}{F_0(x_0)} + \frac{g_2(x_0)}{1 - F_0(x_0)}\right) \cdot \int_{-1}^{1} \psi^2(t)\,\mathrm{d}t \cdot h \cdot (1 + o(1))$$

Now $I_2(c)$ is an integral over the set

$$\left\{(y_1, y_2) : \quad y_1 < y_2, \quad \min\bigl(|y_1 - x_0|, |y_2 - x_0|\bigr) \le h\right\}$$

We further split the integral $I_2(c)$ as

$$I_2(c) = J_1 + J_2 + J_3 + J_4 + J_5,$$

according to

$$\left(\int_{-\infty}^{x_0-\delta} \mathrm{d}y_1 + \int_{x_0-\delta}^{x_0-h} \mathrm{d}y_1\right) \int_{x_0-h}^{x_0+h} \mathrm{d}y_2 + \\ + \int_{x_0-h}^{x_0+h} \mathrm{d}y_1 \left(\int_{y_1}^{x_0+h} \mathrm{d}y_2 + \int_{x_0+h}^{x_0+\delta} \mathrm{d}y_2 + \int_{x_0+\delta}^{\infty} \mathrm{d}y_2\right).$$

Here, $\delta \to 0$ as $h \to 0$, in such a way that $\delta|\log h| \to \infty$ (so $h \to 0$ much faster than δ); the interesting terms will come from J_2 and J_4.

Figure 1. $I_2(c) = J_1 + J_2 + J_3 + J_4 + J_5$.

From the assumptions on F, g and $\{F_c\}$ (using the fact that $\gamma_1(F_c)$ is bounded away

from zero uniformly in c and h, for small enough h)

$$
\begin{aligned}
J_1 + J_5 &\le \frac{\text{const}}{\delta - h} \int_{x_0-h}^{x_0+h} (g_1(t) + g_2(t))\, \mathrm{d}t = \mathcal{O}\left(\frac{h}{\delta - h}\right) \\
J_3 &\le \text{const} \int_{x_0-h}^{x_0+h} \int_{y_1}^{x_0+h} \frac{\left(\frac{y_1-y_2}{h}\right)^2}{y_2 - y_1} g(y_1, y_2) \mathrm{d}y_2 \mathrm{d}y_1 = \mathcal{O}(h) \\
J_2 + J_4 &= \int_{x_0-h}^{x_0+h} \int_{x_0-\delta}^{x_0-h} \frac{\psi^2\left(\frac{y_2-x_0}{h}\right)}{F_c(y_2) - F_c(y_1)} g(y_1, y_2)\, \mathrm{d}y_1 \mathrm{d}y_2 \\
&\qquad + \int_{x_0-h}^{x_0+h} \int_{x_0+h}^{x_0+\delta} \frac{\psi^2\left(\frac{y_1-x_0}{h}\right)}{F_c(y_2) - F_c(y_1)} g(y_1, y_2)\, \mathrm{d}y_2 \mathrm{d}y_1 \\
&\le \text{const.}\, (1 + o(1))\, g(x_0, x_0) \left(\int_{x_0-h}^{x_0+h} \int_{x_0-\delta}^{x_0-h} \frac{\psi^2\left(\frac{y_2-x_0}{h}\right)}{y_2 - y_1}\, \mathrm{d}y_1 \mathrm{d}y_2 \right. \\
&\qquad \left. + \int_{x_0-h}^{x_0+h} \int_{x_0+h}^{x_0+\delta} \frac{\psi^2\left(\frac{y_1-x_0}{h}\right)}{y_2 - y_1}\, \mathrm{d}y_2 \mathrm{d}y_1 \right) \\
&= \text{const.} g(x_0, x_0)\, (1 + o(1)) \int_{-1}^{1} \psi^2(t)\, h \log\left(\frac{th + \delta}{th + h}\right) \mathrm{d}t \\
&= \text{const.} g(x_0, x_0)\, h\, |\log h| \int_{-1}^{1} \psi^2(t)\, \mathrm{d}t\, (1 + o(1)) \\
&= A_1\, h\, |\log h\,| (1 + o(1)), \qquad \text{say.}
\end{aligned}
$$

Now van Trees' inequality, just as in Theorem 2, gives

$$
\begin{aligned}
r_n(V) &\ge \left(n A_1 h |\log h| (1 + o(1)) + \frac{\mathcal{I}_0}{B^2 h^2} \right)^{-1} \\
&= \left(\frac{A_1}{3} + \frac{\mathcal{I}_0}{B^2} \right) (n \log n)^{-2/3} (1 + o(1))
\end{aligned}
$$

if we take $h = (n \log n)^{-\frac{1}{3}}$. $\square$

This optimal rate has been found by Bakker (1988) and is shown in Groeneboom and Wellner (1992) also to be achieved by the NPMLE. A reasonably simple ad hoc estimator achieving this rate has been constructed by L. Birgé, in fact a kind of weighted histogram estimator. The NPMLE itself can only be computed by an iterative procedure, unlike in the first example where a simple calculation is available.

6. Semiparametric models: asymptotic global bounds

We now use the multivariate van Trees inequality to derive asymptotic bounds for an infinite dimensional estimator in a semiparametric model, i.e., we consider the estimator in its entirity instead of at a single point. For simplicity we consider the simple and familiar example of estimating an unknown distribution function based on a random sample from the distribution itself. We show that the ordinary empirical distribution function achieves a local asymptotic minimax lower bound for the *integrated* mean square error of an arbitrary estimator. Such results have been established earlier by Levit (1978) and Millar (1979), using the theory of Local Asymptotic Normality. The approach here, based on the elementary inequalities obtained above, can also be applied to more elaborate examples. The difficulty will be rather to show that certain estimators achieve the bounds than to obtain the bounds themselves.

Truncation arguments as in Section 3 can be used to turn the results into bounds on the integrated risk of an asymptotic distribution of an estimator sequence, under uniformity and continuity conditions (or under Hájek regularity).

Let $X_1, \ldots, X_n$ be independent vector observations in $\mathbb{R}^s$ with unknown distribution function F; write X for a generic observation. Let F_n denote an arbitrary estimator of F. We study the integrated mean square error

$$R_n(F_n, F) = n\mathrm{E}_F \int_{\mathbb{R}^s} \big(F_n(x) - F(x)\big)^2 \mu(\mathrm{d}x)$$

where μ is a fixed finite measure on $\mathbb{R}^s$.

For x and y in $\mathbb{R}^s$ we denote by $x \wedge y$ the vector containing the coordinatewise minima of the components of x and y, while $\chi(x)$ is the indicator function of the closed positive orthant in $\mathbb{R}^s$:

$$\begin{aligned} (x \wedge y)_i &= x_i \wedge y_i, \\ \chi(x) &= 1\{x_i \geq 0 \ \forall i\} \end{aligned}.$$

Let $\widehat{F}_n(x) = \frac{1}{n}\sum \chi(x - X_i)$ be the empirical distribution function. We have

$$R_n(\widehat{F}_n, F) = \int F(x)(1 - F(x))\mu(\mathrm{d}x) = R_0(F), \tag{17}$$

say.

We do not necessarily assume that F is completely unkown. However the class of possible F must be rich enough if $R_0(F)$ is to figure as a lower bound for $R_n(F_n, F)$. In order to specify exactly what is a rich enough class, introduce a complete orthonormal system of functions $\{\phi_i(x)\}$ in $L^2_\mu(\mathbb{R}^s)$. Define functions g_i by

$$g_i(y) = \int \chi(x - y)\phi_i(x)\mu(\mathrm{d}x). \tag{18}$$

Let $\mathcal{F}$ be a set of distribution functions on $\mathbb{R}^s$ endowed with the topology $\mathcal{T}$ induced by total variation distance. The following assumption is now essential (Levit, 1978, Millar, 1979): we say that $\mathcal{F}$ is *rich enough* if for each $F_0 \in \mathcal{F}$ and any $k = 1, 2, \ldots$ there exists

a sequence of functions $g_{ij}(y)$, $i = 1, \ldots, k$, $j = 1, \ldots$, and neighbourhoods C_j of zero in $\mathbb{R}^k$, such that

$$\lim_{j\to\infty} \int \big(g_{ij}(y) - g_i(y)\big)^2 \mathrm{d}F_0(y) = 0, \quad i = 1, \ldots, k, \tag{19}$$

and for each $j = 1, 2, \ldots$ the family of distributions $\mathcal{F}_j = \{F_c : c \in C_j\}$ on $\mathbb{R}^s$ defined by

$$\frac{\mathrm{d}F_c}{\mathrm{d}F_0}(y) = \exp\big(\sum_{i=1}^{k} c_i g_{ij}(y) - b(c)\big) \tag{20}$$

belongs to $\mathcal{F}$.

Now we can state our main result:

Theorem 4. *If $\mathcal{F}$ is rich enough, then*

$$\lim_{V \downarrow \{F_0\}} \varliminf_{n\to\infty} \inf_{F_n} \sup_{F\in V} R_n(F_n, F) = R_0(F_0) \tag{21}$$

where the infimum is taken over all possible estimators F_n and $V \downarrow \{F_0\}$ denotes the limit in the net of shrinking neighbourhoods (with respect to the variation distance) of F_0.

Proof. Since $R_n(\widehat{F}_n, F) = R_0(F)$ and this quantity is continuous in the variation distance topology, $R_0(F_0)$ is certainly an upper bound to the left hand side of (21). We must show that the later also cannot be less than $R_0(F_0)$. Suppose $R_0(F_0) > 0$ since otherwise the result is trivial. Now, let us define

$$\phi_i(F) = \int F(x)\phi_i(x)\mu(\mathrm{d}x) = \int g_i(y)\mathrm{d}F(y),$$

and for a given estimator F_n we let

$$\phi_{ni} = \int F_n(x)\phi_i(x)\mu(\mathrm{d}x), \quad \text{and} \quad \phi_n = (\phi_{n1}, \phi_{n2}, \ldots).$$

Note that by Parseval's identity

$$R_n(F_n, F) = \mathrm{E}_F \; n\sum_{i=1}^{\infty} \big(\phi_{ni} - \phi_i(F)\big)^2.$$

The reader may also verify that for given F_0 and k, j, the family of distributions F_c, $c \in C_j \subseteq \mathbb{R}^k$ is continuous in c with respect to the topology $\mathcal{T}$.

Now let V be any neighbourhood of $F = F_0$, and for any given k, j let C_j be a neighbourhood of $0 \in \mathbb{R}^k$ such that $F_c \in V$ for all $c \in C_j$. Choose a prior density $\lambda(c)$ with C, the support of λ, contained in C_j, together satisfying the assumptions of the

multivariate van Trees inequality. Then by the L_2 norm version (10) of the van Trees inequality

$$\begin{aligned} r(V) &= \underline{\lim}_n \inf_{F_n} \sup_{F\in V} R_n(F_n, F) \geq \underline{\lim}_n \inf_{F_n} \sup_{c\in C_j} R_n(F_n, F_c) \\ &\geq \underline{\lim}_n \inf_{F_n} \int_C R_n(F_n, F_c)\lambda(c)\mathrm{d}c \\ &\geq \underline{\lim}_n \inf_{F_n} \int_C E_{F_c} n \sum_{i=1}^k \big(\phi_{ni} - \phi_i(F_c)\big)^2 \lambda(c)\mathrm{d}c \\ &\geq \underline{\lim}_n \frac{n\left(\int_C \sum_{i=1}^k \frac{\partial\phi_i(F_c)}{\partial c_i}\lambda(c)\mathrm{d}c\right)^2}{n\int_C \sum_{i=1}^k \mathrm{Var}_{F_c} g_{ij}(X)\lambda(c)\mathrm{d}c + \widetilde{\mathcal{I}}(\lambda)} \\ &= \frac{\left(\int_C \sum_{i=1}^k \frac{\partial\phi_i(F_c)}{\partial c_i}\lambda(c)\mathrm{d}c\right)^2}{\int_C \sum_{i=1}^k \mathrm{Var}_{F_c} g_{ij}(X)\lambda(c)\mathrm{d}c}, \end{aligned} \tag{22}$$

provided the denominator of the final term is non-zero. We have used here the fact from exponential family theory that the $b(c)$ in (20) satisfy the relations

$$\frac{\partial b}{\partial c_i}(c) = \mathrm{E}_{F_c} g_{ij}(X), \quad i = 1, \ldots, k.$$

Observe next that the functions $g_i(y)$ in (18) are uniformly bounded (by $(\mu(\mathbb{R}^s))^{1/2}$), and

$$\begin{aligned} \frac{\partial\phi_i(F_c)}{\partial c_i} &= \int g_i(y)\big(g_{ij}(y) - \mathrm{E}_{F_c} g_{ij}(X)\big)\mathrm{d}F_c \\ &= \mathrm{Cov}_{F_c}\big(g_i(X), g_{ij}(X)\big) \end{aligned}$$

and hence according to (19)–(20)

$$\begin{aligned} \lim_{j\to\infty}\lim_{c\to 0} \frac{\partial\phi_i(F_c)}{\partial c_i} &= \mathrm{Var}_{F_0} g_i(X), \\ \lim_{j\to\infty}\lim_{c\to 0} \mathrm{Var}_{F_c} g_{ij}(X) &= \mathrm{Var}_{F_0} g_i(X). \end{aligned}$$

Thus letting $C \downarrow \{0\}$ in (22) and then $j \to \infty$ we find

$$r(V) \;\geq\; \sum_{i=1}^k \mathrm{Var}_{F_0} g_i(X)$$

and, since k is arbitrary, using again Parseval's identity

$$\begin{aligned} r(V) \geq \sum_{i=1}^{\infty} \mathrm{Var}_{F_0} g_i(X) &= \int \sum_{i=1}^{\infty} g_i(y)^2 \mathrm{d}F_0(y) - \sum_{i=1}^{\infty} \Big(\int g_i(y) \mathrm{d}F_0(y) \Big)^2 \\ &= \int \int \chi^2(x-y) \mathrm{d}\mu(x) \mathrm{d}F_0(y) - \sum_{i=1}^{\infty} \phi_i(F_0)^2 \\ &= \int F_0(x) \mathrm{d}\mu(x) - \int F_0(x)^2 \mathrm{d}\mu(x) = R_0(F_0), \end{aligned}$$

concluding the proof of the theorem. □

The approach we have just given works, without any changes being needed at all, for well-behaved semiparametric models. Suppose for instance we still want to estimate a distribution function, now not the distribution of the data but an infinite dimensional parameter of its distribution. Considering integrated mean square error, the problem is converted into estimating a sequence of real functionals of F and then by truncation into estimating just a finite number of real functionals. We next consider the finite dimensional submodel whose score functions are exactly the optimal influence functions for these functionals (the projections into the tangent space of the gradients of the functionals). As in our calculations above, the denominator in the final line of (22) is the same, when squared, as the numerator, giving as lower bound the sum of the lower bounds for each functional separately. This lower bound corresponds to the integrated mean square error of the optimal limiting distribution found in the convolution or local asymptotic minimax theorems. Indeed, denoting the projected or canonical gradients by $g_i(x)$, $i = 1, \ldots, k$, the general theory says that provided $\mathcal{F}$ is rich enough $\sum_{i=1}^{k} \mathrm{Var}_{F_0} g_i(X)$ is the required lower bound for estimating the finite number of functionals with respect to sum of squared error losses.

One could apply this to the usual random censorship model. Weits (1991) shows that n times the mean square error of the Kaplan-Meier estimator is equal to the mean square error of its limiting distribution, up to an error of order $1/n$, provided one stays away from the right tail of the distribution of the data; this limiting distribution is known to be optimal in the sense just mentioned. Therefore the estimator has asymptotically optimal integrated mean square error with respect to measures μ with support strictly inside the support of the observations. It would be nice to extend this to a result on the whole line under appropriate integrability conditions. We believe Weits' striking results on second order properties of the Kaplan-Meier estimator could possibly have been obtained more easily by exploiting the van Trees inequality.

7. References.

D. Bakker (1988), *Nonparametric maximum likelihood estimation of the distribution function of interval censored observations*, Master's Thesis, Dept. Math., Univ. Amsterdam.

B. Z. Bobrovsky, E. Mayer-Wolf and M. Zakai (1987), Some classes of global Cramér-Rao bounds, *Ann. Statist.* **15**, 1421–1438.

A. A. Borovkov (1984), *Mathematical Statistics. Parameter Estimation*, Nauka, Moscow.

A. A. Borovkov and A. U. Sakhanenko (1980), On estimates of the expected quadratic risk, *Probab. Math. Statist.* **1**, 185–195.

L. D. Brown and L. Gajek (1990), Information inequalities for the Bayes risk, *Ann. Statist.* **18**, 1578–1594.

P. Groeneboom (1985), Estimating a monotone density. In *Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II* (L.M. Le Cam and R.A.Olshen, eds.), 539-555. Wadsworth, Hayward.

P. Groeneboom (1987), *Asymptotics for interval censored observations*, report 87–18, Dept. Math., Univ. Amsterdam.

P. Groeneboom and J. A. Wellner (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, DMV Seminar Vol. 19, Birkhäuser Verlag.

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

C. A. J. Klaassen (1989), The asymptotic spread of estimators, *J. Statist. Planning and Inference* **23**, 267–285.

B. Y. Levit (1978), Infinite-dimensional informational inequalities, *Theor. Probab. Appl.* **23**, 371–377.

B. Y. Levit and C. G. M. Oudshoorn (1993), Second order admissible variance estimation. In *Recent Developments in Estimation Theory and Related Topics (Suppl. Issue No. 3 to Statistics & Decisions)* (E.J.Dudewicz, D. Plachky and P.K. Sen, eds.) , 17–29. Oldenbourg, München.

M. Loève (1963), *Probability Theory*, 3rd ed., van Nostrand, New York.

P.W. Millar (1979), Asymptotic minimax theorems for the sample distribution function, *Z. Wahrsch. verw. geb.* **48**, 233–252.

D. Pollard (1984), *Convergence of Stochastic Processes*, Springer, New York.

C. R. Rao (1963), Criteria of estimation in large samples, *Sankhya A* **18**, 371–378.

M. Schipper (1992), *Optimal rates of the second order improvements on the sample distribution in L_2 norm*, Master's thesis, Mathematical Institute, University Utrecht.

H. L. van Trees (1968), *Detection, Estimation and Modulation Theory, Part 1.* Wiley, New York.

E. A. G. Weits (1991), The second order optimality of a smoothed Kaplan-Meier estimator, *Scand. J. Statist.* **20**, 111-132.

E. A. G. Weits (1992), *Results on the minimax risk for non-regular estimation in two models for i.i.d. observations*, Report 92–04, Dept. Math., Wageningen Agricultural University.