TANGENT SPACES, INFORMATION AND SEMIPARAMETRICS

ROGER KOENKER

Abstract. A brief introduction to information bounds for semiparametric models.

1. Introduction

Semiparametric models require some elaboration of the conventional machinery for evaluation of the efficiency of statistical procedures. These notes, which follow Van der Vaart (2000) closely, are intended to provide a quick and dirty introduction the some of the basic ideas. For further details readers may consult Van der Vaart (2000) or Bickel et al. (1998).

As usual we will assume that we have at hand a random sample, $X_1, \ldots, X_n$ from a distribution P belonging to $\mathcal{P}$, a set of probability measures on a sample space $(\mathcal{X}, \mathcal{A})$. We are interested in estimating a scalar parameter $\psi : \mathcal{P} \mapsto \mathbb{R}$. How well can we expect to be able to do this? Is there an expanded notion of Fisher information and the Cràmer-Rao inequality applicable to such semiparametric settings? The obvious strategy for attacking these questions is to try to reduce the problem back to its parametric formulation, where we know how to proceed.

Estimating $\psi(P)$ in model $\mathcal{P}$ is certainly going to be more difficult than doing so in any restricted submodel $\mathcal{P}_0 \subset \mathcal{P}$. For any (smooth) parametric submodel, say $\mathcal{P}_t = \{P_t : t \in \mathcal{T} \subset \mathbb{R}\} \subset \mathcal{P}$ we can compute Fisher information as usual. So our task would be (simply!) to find the least favorable parametric submodel of this form. This is reminiscent of the Huber least favorable location model problem.

For convenience we will consider only one dimensional families of submodels and assume that our parametric families $t \mapsto P_t$ have densities p_t with respect to some dominating measure μ. Further, we will require that there are measureable functions $g : \mathcal{X} \mapsto \mathbb{R}$, such that,

$$\int \left(\frac{\sqrt{p_t} - \sqrt{p_0}}{t} - \tfrac{1}{2} g \sqrt{p_0} \right)^2 d\mu \to 0. \tag{DQM}$$

We say that the family P_t is differentiable in quadratic mean with score function g.

Obviously there are lots of these parametric models P_t. Our job is to find the worst one, that is the one that makes it hardest to estimate ψ. The collection of the score functions for the submodels is called the tangent set of the model $\mathcal{P}$ at P. Geometrically, as Van der Vaart (2000) notes, $\sqrt{p}$ as P ranges over $\mathcal{P}$ can be viewed as a subset of the unit ball in $\mathcal{L}_2(P)$, and $\frac{1}{2} g \sqrt{p}$ is its tangent set.

Version: March 14, 2016.

Recall [!?] from L8 of 574: Given a parametric model f_t and supposing that $\sqrt{f_t}$ has a derivative at $t = 0$, a.e. μ, then

$$\left.\frac{d\sqrt{f_t}}{dt}\right|_{t=0} = \tfrac{1}{2}\frac{f_0'}{\sqrt{f_0}}$$

so

$$\lim_{t\to 0}\int\left(\frac{\sqrt{f_t}-\sqrt{f_0}}{t}\right)^2 d\mu = \tfrac{1}{4}\int\left(\frac{f_0'}{f_0}\right)^2 f_0 d\mu = \tfrac{1}{4}I(f_0),$$

and thus we can interpret g in this parametric setting as $-f_0'/f_0$, the familiar score function of the MLE, and $I(f_0)$ denotes the usual Fisher information that that describes how much information about the parameter t is contained in a single observation on X in the parametric case. So it shouldn't be too surprising that this trail of breadcrumbs is leading us to a new, more grandiose notion of information for the much larger class of semiparametric models.

The score, g also plays an important role in characterizing the limiting normality of procedures. This is essentially a nonparametric version of the familiar quadratic expansion of the log likelihood.

Lemma 1. *If the path $t \mapsto P_t$ in $\mathcal{P}$ satisfies the prior DQM condition, then $Pg = 0$, $Pg^2 < \infty$ and*

$$\log\prod_{i=1}^n \frac{dP_{1/\sqrt{n}}}{dP}(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i) - \frac{1}{2}Pg^2 + o_P(1)$$

A word on notation: The notation Pg for $\int g dP$ is becoming increasingly common, largely as a result of its advocacy by David Pollard, who attributes it to deFinetti. In effect it replaces the usual $\mathbb{E}_P g$ by Pg, which not only saves ink, but seems altogether more cogent since we can view expectation as a linear operator.

Since we are only interested in $\psi(P)$ we will restrict attention to submodels, $t \mapsto P_t$ such that $\psi(P_t)$ is differentiable, that is we require a linear map $\dot{\psi}_P : \mathcal{L}_2(P) \mapsto \mathbb{R}$ such that for every g in the tangent set and submodel $t \mapsto P_t$ with score g,

$$\frac{\psi(P_t) - \psi(P)}{t} \to \dot{\psi}_P g.$$

By the Riesz representation theorem we can find a $\tilde{\psi}_P$ such that

$$\dot{\psi}_P g = \langle \tilde{\psi}_P, g\rangle \equiv \int \tilde{\psi}_P g dP$$

We can now evaluate the potential performance of procedures for estimating $\psi(P)$. The Fisher information about t in the parametric submodel $t \mapsto P_t$, with score function,

$$g(x) = \left.\frac{\partial \log p_t(x)}{\partial t}\right|_{t=0}$$

is Pg^2, so the optimal asymptotic variance of a procedure $t \mapsto \psi(P_t)$, when evaluated at $t = 0$ is given by the Cramer-Rao inequality as,

$$\left.\frac{(d\psi(P_t)/dt)^2}{Pg^2}\right|_{t=0} = \frac{\langle\tilde{\psi}_P, g\rangle_P^2}{\langle g, g\rangle_P^2}.$$

Finding the supremum over the tangent set yields,

$$\sup_{g\in\mathcal{G}} \frac{\langle\tilde{\psi}_P, g\rangle_P^2}{\langle g, g\rangle_P^2} = P\tilde{\psi}_P^2.$$

Why? That the supremum attains this value can be seen from the Cauchy-Schwarz inequality, $(P\tilde{\psi}_P g)^2 \leq P\tilde{\psi}_P^2 Pg^2$ and some technical requirements to ensure that g is in the closure of the linear space of the tangent set, here denoted by $\mathcal{G}$. These technical issues also arise in determining whether the tangent set can be considered a space, or perhaps just a cone.

Thus, $P\tilde{\psi}_P^2$ the expectation of $\tilde{\psi}_P^2$ under the least favorable P plays the same role as the usual information bound in the parametric setting. Van der Vaart (2000) provides some further justification for this, but I prefer to move on to consider how one might apply all of this to find the least favorable submodel, and thereby achieve semiparametric efficiency.

References

Bickel PJ, Klaassen CA, Ritov Y, Wellner JA. 1998. *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag.

Van der Vaart AW. 2000. *Asymptotic statistics*. Cambridge University Press.