

EC241a

Econometric Theory

Spring 2004

UC Berkeley Department of Economics

GENERALIZED METHOD OF MOMENTS ESTIMATION III:
SEMIPARAMETRIC EFFICIENCY BOUNDS

The basic efficiency argument is the Cramèr–Rao bound for unbiased estimators:

Result 1 (CRAMÈR–RAO BOUND)

Let the probability density function of a random variable X be $f_X(x|\theta)$ for some $\theta_0 \in \Theta$. Let $\hat{\theta}(X)$ be an unbiased estimator for θ_0 . Suppose the derivative $\partial/\partial\theta$ can be passed under the integral sign in $\int f(x|\theta)dx$ and $\int \theta(x)f(x|\theta)dx$, and suppose the Fisher information

$$\mathcal{I}(\theta) = -\mathbb{E} \left[\frac{\partial^2 \ln f}{\partial \theta \partial \theta'}(X|\theta) \right],$$

is finite. Then

$$\text{Var}(\hat{\theta}(X)) \geq \mathcal{I}(\theta_0)^{-1}.$$

Proof: Recall that the square of the covariance of two random variables S and U is less than or equal to the product of the variances (that is the same as saying that the correlation coefficient is less than or equal to one in absolute value:

$$\text{Cov}^2(S, U) \leq V(S) \cdot V(U).$$

Now let us take $S = \hat{\theta}(X)$ and $U = \frac{\partial \ln f}{\partial \theta}(X; \theta)$. First consider the expectation of U , the score function:

$$1 = \int_x f_X(x; \theta) dx.$$

So,

$$0 = \frac{\partial}{\partial \theta} \int_x f_X(x; \theta) dx.$$

Assuming we can change the order of differentiation and integration, we get

$$\begin{aligned} 0 &= \int_x \frac{\partial f_X}{\partial \theta}(x; \theta) dx \\ &= \int_x \frac{\partial \ln f_X}{\partial \theta}(x; \theta) \cdot f_X(x; \theta) dx \\ &= \mathbb{E}\left[\frac{\partial \ln f_X}{\partial \theta}(x; \theta)\right] = \mathbb{E}[U] = 0. \end{aligned}$$

Therefore the covariance of U and S is the expectation of the product of S and U :

$$\begin{aligned} \mathbb{E}[SU] &= \int_x \hat{\theta}(x) \frac{\partial \ln f_X}{\partial \theta}(x; \theta) f_X(x; \theta) dx = \int_x \hat{\theta}(x) \frac{\partial f_X}{\partial \theta}(x; \theta) dx \\ &= \frac{\partial}{\partial \theta} \int_x \hat{\theta}(x) f(x; \theta) dx = \frac{\partial}{\partial \theta} \theta = 1. \end{aligned}$$

So

$$1 \leq V(\hat{\theta}) \cdot V\left(\frac{\partial \ln f}{\partial \theta}(X|\theta)\right),$$

implying

$$V(\hat{\theta}) \geq \left[V\left(\frac{\partial \ln f}{\partial \theta}(X|\theta)\right) \right]^{-1} = \mathbb{E} \left[\left(\frac{\partial \ln f}{\partial \theta}(X|\theta) \right)^2 \right]^{-1}.$$

□.

If X_1, X_2, \dots, X_N are iid with common density $f_X(x|\theta)$, the implied bound is

$$N \cdot \text{Var}(\hat{\theta}(X)) \geq \mathcal{I}(\theta_0)^{-1}.$$

At the same time the limiting distribution of the mle is, under conditions discussed before, equal to

$$\sqrt{N}(\hat{\theta}_{ml} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}).$$

Hence in large samples the mle is approximately unbiased and approximately achieve the Cramèr–Rao bound.

This motivates the following definition of large sample efficiency:

Definition 1 Let X_1, X_2, \dots be iid random variables with common density $f_X(x|\theta)$. A sequence of estimates $\hat{\theta}_N$, a function of X_1, X_2, \dots, X_N such that

$$\sqrt{N}(\hat{\theta}_{ml} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta)^{-1}),$$

whatever the true value $\theta \in \Theta$ is, is said to be asymptotically efficient.

So by definition we declare maximum likelihood estimators to be efficient. This is not completely innocuous. We can actually do better in large samples than maximum likelihood estimators. In particular, there exist superefficient estimators that do well in a particular part of the sample space but not in others. For example, suppose X_1, X_2, \dots are iid normal with mean θ and variance 1. The mle is \bar{x} with normalized variance equal to 1. Now consider the estimator

$$\hat{\theta} = \begin{cases} \bar{x} & \text{if } n^{1/4}|\bar{x}| \geq 1 \\ 0 & \text{if } n^{1/4}|\bar{x}| < 1. \end{cases}$$

If θ differs from zero the large sample distribution of $\hat{\theta}$ is the same as that of the mle \bar{x} . If $\theta = 0$, the normalized variance is zero. Hence at zero $\hat{\theta}$ does better, and if $\theta \neq 0$ this estimator does equally well as the mle. Nevertheless, this is not a particularly attractive estimator as clearly around zero it is going to have relatively unattractive properties, especially in small samples. For example, small changes in the observations can lead to relatively large changes in the estimator, when $n^{1/4}|\bar{x}|$ is close to 1.

How does this lead to efficiency arguments for models that are not fully specified, such as generalized method of moment estimators. The main type of argument, discussed in among others Begun, Hall, Huang, and Wellner (1983), Bickel, Klaassen, Ritov, and Wellner (1990) and Newey (1991), goes along the following lines. Suppose X_1, X_2, \dots are iid with unknown density function $f(x)$. The density depends on some parameters of interest θ but these parameters do not completely describe the distribution. That is, there is more unknown about the distribution than just these parameters. We specify that by making the density function a function of the unknown parameter θ and an unknown function h :

$$X \sim f(x|\theta, h(\cdot)).$$

For example, if we are interested in the probability that a random variable with density $f(x)$ is positive, we could define

$$\theta = \mathbb{E}\left[1\{X > 0\}\right],$$

and

$$h(y) = \frac{f(y) \cdot 1\{y > 0\}}{\int_0^{\infty} f(z) dz} + \frac{f(y) \cdot 1\{y \leq 0\}}{\int_{-\infty}^0 f(z) dz}.$$

In that case

$$f(x) = f(x|\theta, h(\cdot)) = 1\{x > 0\} \cdot \theta \cdot h(x) + 1\{x \leq 0\} \cdot (1 - \theta) \cdot h(x).$$

Now suppose we pretend we know $h(\cdot)$. In that case we have a fully parametric model and we can calculate the Cramér-Rao bound. Of course we do not expect we can find an estimator that achieves that bound because we do not actually know $h(\cdot)$. However, if we were able to find an estimator that achieves this bound without requiring knowledge of $h(\cdot)$ then this estimator is certainly efficient: there cannot be a better estimator even if we add information in the form of knowledge of $h(\cdot)$.

Now suppose we pretend we know $h(\cdot)$ up to a finite dimensional parameter vector γ . Then we again have a fully parametric model

$$f(x|\theta, \gamma) = f(x|\theta, h(\gamma)).$$

Hence we can calculate the Cramér-Rao bound for θ . Partitioning the information matrix for $(\theta', \gamma)'$ and its inverse in

$$\mathcal{I}(\theta, \gamma) = \begin{pmatrix} \mathcal{I}_{\theta\theta'} & \mathcal{I}_{\theta\gamma'} \\ \mathcal{I}_{\gamma\theta'} & \mathcal{I}_{\gamma\gamma'} \end{pmatrix} \quad \text{and} \quad \mathcal{I}(\theta, \gamma)^{-1} = \begin{pmatrix} \mathcal{I}^{\theta\theta'} & \mathcal{I}^{\theta\gamma'} \\ \mathcal{I}^{\gamma\theta'} & \mathcal{I}^{\gamma\gamma'} \end{pmatrix},$$

the Cramér-Rao bound implies that

$$\text{Var}(\hat{\theta}) \geq \mathcal{I}^{\theta\theta'} = \left(\mathcal{I}_{\theta\theta'} - \mathcal{I}_{\theta\gamma'} \mathcal{I}_{\gamma\gamma'}^{-1} \mathcal{I}_{\gamma\theta'}\right)^{-1}. \quad (1)$$

This relation has to be true for any parametrization of the unknown function $h(\cdot)$. What we are looking for is the worst possible parametrization, that is, the parametrization that

maximizes the right hand side of (1). The lowest possible variance for any estimator for θ that does not use knowledge of $h(\cdot)$ has to be at least as high as the lowest variance we can get if we know more, that is, the Cramér–Rao bound for any parametric submodel. So the semiparametric efficiency bound is the largest lower bound we can get for any parametric submodel. It is of course difficult to see if we have found the largest lower bound. Often, however, an candidate efficient estimator exists. If we find a parametric model with variance equal to the variance of this estimator, we know that this is an efficient estimator, and we know the efficiency bound.

To see how this works in the above example, suppose we actually know $h(x)$. Then we have a fully parametric model with

$$f(x) = 1\{x > 0\} \cdot \theta \cdot h(x) + 1\{x \leq 0\} \cdot (1 - \theta) \cdot h(x).$$

The logarithm of the density function is

$$f(x) = 1\{x > 0\} \cdot \ln \theta + 1\{x > 0\} \cdot \ln h(x) + 1\{x \leq 0\} \cdot \ln(1 - \theta) + 1 \cdot \{x \leq 0\} h(x).$$

The score function is

$$\mathcal{S}(x|\theta) = \frac{1\{x > 0\}}{\theta} - \frac{1\{x \leq 0\}}{1 - \theta} = \frac{1\{x > 0\} - \theta}{\theta(1 - \theta)}.$$

The variance of the score, equal to the information, is

$$\mathbb{E}\mathcal{S}(x|\theta)^2 = \frac{1}{\theta(1 - \theta)},$$

and the variance bound is $\theta(1 - \theta)$. Now consider the estimator

$$\hat{\theta} = \sum_{i=1}^N 1\{x > 0\}/N.$$

This has mean θ and variance $\theta(1 - \theta)/N$. Hence $\hat{\theta}$ is efficient and $\theta(1 - \theta)$ is the semi-parametric efficiency bound.

ADDITIONAL REFERENCES

BEGUN, J., W. HALL, W. HUANG, AND J. WELLNER, (1983) “Information and Asymptotic Efficiency in parametric–nonparametric Models”, *Annals of Statistics*, Vol. 11, 432–452.

BICKEL, P., C. KLAASSEN, Y. RITOV, AND J. WELLNER, *Efficient and Adaptive Inference in Semiparametric Models*, Johns Hopkins University Press.

NEWBY, W., (1990), “Semiparametric Efficiency Bounds”, *Journal of Applied Econometrics*, Vol. 5, 99–135.