A CONSTRUCTIVE PROOF OF THE OPTIMALITY OF THE ML ESTIMATOR

by

Franco Peracchi

Department of Economics

University of California
Los Angeles

June 1988
Working Paper # 484

Abstract

This paper presents a simple constructive proof of the strong optimality of the ML estimator of a correctly specified parametric model. The method of proof, based on the Projection Theorem for Hilbert spaces, reveals the simple geometric nature of the problem, and illustrates a general way of constructing estimators with pre-specified statistical properties.

JEL No. 211
Keywords: M-estimators, ML estimators, asymptotic efficiency.

1. Introduction

Econometric practice is largely based on the method of maximum likelihood (ML). This method requires specifying a parametric model for the observations, and leads to estimators that, under certain regularity conditions, are consistent and asymptotically efficient when the assumed parametric model is correctly specified. Efficiency is typically established in an indirect way, namely by showing that in large samples the ML estimator attains the Cramér-Rao lower bound for the variance of a consistent estimator of a correctly specified model.

In this paper we present a simple constructive proof of the optimality of the ML estimator of a correctly specified parametric model [1]. This proof, based on the Projection Theorem for Hilbert spaces, is interesting because it reveals the simple geometric nature of the problem, and illustrates a general method for constructing estimators with pre-specified statistical properties. This method can for example be applied to construct estimators that satisfy, in addition to the classical consistency and efficiency properties, other requirements such as robustness to small departures from the assumed statistical model.

The next Section collects a few results that relate the asymptotic properties of an estimator to the properties of its influence function [Hampel (1974)]. These results are then used in Section 3 to establish the strong optimality of the ML estimator. Section 4 contains some conclusions.

2. The statistical model and the class of M-estimators

Let $z_1,\ldots,z_N$ be a sequence of independently and identically distributed (i.i.d.) random vectors, with values in a known subset $\mathcal{Z}$ of $\mathbb{R}^m$, and common d.f. F_0. The d.f. F_0 is typically unknown, but prior information may be available to justify the assumption that F_0 belongs to some set $\mathcal{F}$ of d.f.'s over $\mathcal{Z}$. In this paper $\mathcal{F}$ is taken to be a family of d.f.'s indexed by a p-dimensional parameter θ, i.e. $\mathcal{F} = \{F_\theta: \theta \in \Theta\}$. The parametric model $\mathcal{F}$ is assumed to be correctly specified and identifiable, in the sense that there is a unique $\theta_0 \in \Theta$ such that $F_0 = F_{\theta_0}$. We also assume that $\mathcal{F}$ is regular in some neighborhood Θ_0 of θ_0, that is, for all $\theta \in \Theta_0$ the d.f. F_θ possesses a smooth density function $f(\cdot,\theta)$, a likelihood score function $s(\cdot,\theta) = (\partial/\partial\theta) \ln f(\cdot,\theta)$, and a finite positive definite (p.d.) Fisher information matrix $J(\theta)$.

An M-estimator $\hat{\theta}_N$ of the unknown parameter θ_0 is a root of an implicit equation of the form

$$N^{-1} \Sigma_{n=1}^{N} \eta_N(z_n,\theta) = 0, \tag{1}$$

where $\eta_N(z,\theta)$, called the score function associated with $\hat{\theta}_N$, is some measurable function mapping $\mathcal{Z} \times \mathbb{R}^p$ into $\mathbb{R}^p$[2]. If (1) has multiple roots, we assume that some selection rule has been specified to make $\hat{\theta}_N$ unique.

Consistency and asymptotic normality of M-estimators can be established under very general conditions [see e.g. Huber (1967)]. Moreover, as we shall see, estimators that are not in this class are often asymptotically equivalent to M-estimators.

Because of the independence assumption, the order of the observations in the sample does not matter. One can therefore replace functions of the observations by statistical functionals, i.e. functionals defined over a set of d.f.'s. In particular, if the score function $\eta_N(z,\theta)$ does not depend on N, equation (1) and the given selection rule implicitly define a functional $\hat{\theta}$: $F_N \longrightarrow \hat{\theta}_N = \hat{\theta}(F_N)$, where F_N denotes the empirical d.f. of the observations. More generally, if the score function $\eta_N(z,\theta)$ has an a.s. limit $\eta(z,\theta)$, then we can associate with $\eta(z,\theta)$ the functional $\hat{\theta}$: $F \longrightarrow \hat{\theta}(F)$ implicitly defined by

$$\int_{\mathfrak{Z}} \eta(z,\hat{\theta}(F))\, dF(z) = 0. \qquad (3)$$

The domain of this functional is taken to be a convex set containing the true d.f. F_0, the assumed model $\mathcal{F}$ and all empirical d.f.'s over $\mathfrak{Z}$. We assume that the score function $\eta(z,\theta)$ is piecewise continuously differentiable with respect to θ, and that the matrix $P(\hat{\theta},F) = (\partial/\partial\theta)\ [\ -\ E_F\ \eta(z,\hat{\theta}(F))]$ is finite and p.d. for all d.f.'s F in a neighborhood of F_0 [3].

In the case of a statistical functional it is more reasonable to adopt a definition of consistency that differs slightly from the usual definition of weak consistency [see e.g. Cox and Hinkley (1974)].

DEFINITION 1: $\hat{\theta}$ is called Fisher-consistent for θ if $\hat{\theta}(F_\theta) = \theta$.

Thus, an M-estimator defined by (3) is Fisher consistent for θ_0 whenever the assumed model $\mathcal{F}$ is correctly specified and the equation $E_0\ \eta(z,\theta) = 0$ has a unique root at $\theta = \theta_0$. For ML estimators of regular parametric models, the latter condition corresponds to the standard assumption that the likelihood function has a unique global maximum at θ_0. If the assumed model is misspecified, $\hat{\theta}(F_0)$ defines the population parameter estimated by $\hat{\theta}$.

A statistical functional $\hat{\theta}$ need not be linear. However, when it is linear, that is, $\hat{\theta}(F) = E_F \psi(z)$ for some function ψ, its asymptotic properties are easy to establish. This suggests deriving the asymptotic properties of a non-linear functional by means of a suitable linearization.

Under appropriate regularity conditions [4], a statistical functional $\hat{\theta}$ possesses the (von Mises) expansion

$$\hat{\theta}(F_N) - \hat{\theta}(F_0) = N^{-1} \sum_{n=1}^{N} IF(z_n, \hat{\theta}, F_0) + o_P(N^{-1/2}) \tag{4}$$

where $IF(\cdot, \hat{\theta}, F_0)$ is called the influence function of $\hat{\theta}$ [Hampel (1974)]. A more explicit definition of this function is the following:

DEFINITION 2: Let F be an arbitrary d.f., let $\Delta_{(z)}$ be a d.f. wit mass concentrated at the point z and let $F_{\varepsilon,z} = (1 - \varepsilon) F + \varepsilon \Delta_{(z)}$. Then the IF of $\hat{\theta}$ at F is defined by

$$IF(z, \hat{\theta}, F) = \lim_{\varepsilon \longrightarrow 0+} [\hat{\theta}(F_{\varepsilon,z}) - \hat{\theta}(F)]/\varepsilon,$$

provided that the limit exists.

Thus, the IF is just the collection of Gateaux differentials of the functional $\hat{\theta}$ in the direction of point mass distributions [5].

By the Implicit Function Theorem applied to (3), if the M-estimator $\hat{\theta}$ is Fisher consistent, its IF at the d.f. F_θ is equal to

$$IF(z, \hat{\theta}, F_\theta) = P(\hat{\theta}, F)^{-1} \eta(z, \theta). \tag{5}$$

Thus, the IF of $\hat{\theta}$ at F_{θ} is just a non-singular linear transformation of the score function that defines $\hat{\theta}$. For example, if $\hat{\theta}$ is the ML estimator, then $P(\hat{\theta},F_{\theta}) = J(\theta)$ and so $IF(z,\hat{\theta},F_{\theta}) = J(\theta)^{-1}\ s(z,\theta)$. Moreover,

$$P(\hat{\theta},F_{\theta}) = E_{\theta}\ \eta(z,\theta)\ s(z,\theta)', \tag{6}$$

which generalizes the familiar equality between the information matrix of a correctly specified model and the expectation of the outer product of the likelihood score.

When the matrix $E_0\ [IF(z,\hat{\theta},F_0)\ IF(z,\hat{\theta},F_0)']$ is finite and p.d., the term $N^{-1} \sum_{n=1}^{N} IF(z_n,\hat{\theta},F_0)$ is an average of i.i.d. random vectors with zero mean and finite p.d. variance. It then follows from (4) that

$$N^{1/2}\ [\hat{\theta}(F_N) - \hat{\theta}(F_0)] \xrightarrow{d} \mathcal{N}(0,AV(\hat{\theta},F_0)), \tag{7}$$

where $AV(\hat{\theta},F_0) = E_0\ [IF(z,\hat{\theta},F_0)\ IF(z,\hat{\theta},F_0)']$. If $\hat{\theta}$ is an M-estimator $AV(\hat{\theta},F_0) = P_0^{-1}\ Q_0\ P_0'^{-1}$, where $P_0 = P(\hat{\theta},F_0)$ and $Q_0 = E_0\ [\eta(z,\hat{\theta}(F_0))\ \eta(z,\hat{\theta}(F_0))']$. In particular, if $\hat{\theta}$ is the ML estimator based on the correctly specified parametric model $\mathcal{F}$, then $P_0 = Q_0 = J(\theta_0)$ and one obtains the familiar result that $AV(\hat{\theta},F_0) = J(\theta_0)^{-1}$.

It is clear from (5) and (7) that any estimator $\hat{\theta}$ which possesses an expansion of the form (4) has the same IF and hence the same asymptotic distribution as an M-estimator based on the score function $\eta(z,\theta) = IF(z,\hat{\theta},F_{\theta})$. Thus, if interest focuses only on asymptotic properties, there is little loss of generality in considering only M-estimators.

3. Optimality of the ML estimator

It is well known that, if the assumed model is correctly specified and certain regularity conditions hold, then any consistent estimator of θ_0 has an AVM that exceeds the AVM of the ML estimator by a positive semi-definite matrix. This strong optimality property, which justifies the use of ML estimators, is typically established by showing that in large samples the ML estimator attains the Cramer-Rao lower bound for the variance of a consistent estimator of a correctly specified model. We shall now provide an alternative, constructive proof of this fundamental result. This proof is interesting because it reveals the simple geometric nature of the problem, and illustrates a general method for constructing estimators with pre-specified statistical properties.

We restrict attention to 'regular' M-estimators, namely the ones that are Fisher consistent at the assumed model, possess an IF, and have an asymptotically normal distribution with a finite, p.d. AVM. Efficiency of estimation is defined in terms of an asymptotic mean square error (MSE) criterion of the form $\mathrm{MSE}(\hat{\theta},F_\theta) = \mathrm{trace}\ [Q\ \mathrm{AV}(\hat{\theta},F_\theta)]$, where $Q = [q_{ij}]$ is some p.d. matrix. We shall consider the problem of finding an optimal M-estimator $\tilde{\theta}$, namely one that is Fisher-consistent and has minimum asymptotic MSE at the assumed model.

We first translate the problem of finding an optimal M-estimator into the one of finding an optimal score function in a certain class. In order to do so, we need to ensure that there is a one-to-one correspondence between the set of M-estimators and the set of score functions. We therefore impose a 'normalization' conditions on the score function $\eta(\cdot,\theta)$ that defines $\hat{\theta}$, namely $\eta(\cdot,\theta)$ must be such that $(\partial/\partial\theta')\ E_F\ [-\ \eta(z,\theta)] = I_p$. Under this condition, the IF of a Fisher-consistent estimator $\hat{\theta}$ at F_θ is equal to $\eta(\cdot,\theta)$ and its AVM is equal to $E_\theta\ [\eta(z,\theta)\ \eta(z,\theta)']$. When the assumed model is correctly specified, an optimal estimator of the unknown parameter θ_0 can therefore be based on the score function that solves the following problem

$$\text{Min}_{\ \eta(\cdot,\theta)\in H}\ E_\theta\ \eta(z,\theta)'\ Q\ \eta(z,\theta) \tag{8}$$

$$\text{s.t.} \qquad E_\theta\ \eta(z,\theta) = 0 \tag{9}$$

$$E_\theta\ \eta(z,\theta)\ s(z,\theta)' = I_p \tag{10}$$

for all θ in a neighborhood of θ_0. The set H is the linear vector space of measurable functions, mapping $\mathfrak{X}$ into $\mathbb{R}^p$, that are square integrable with respect to F_θ for all θ in a neighborhood of θ_0. Constraint (9) ensures Fisher-consistency. Given (6), constraint (10) corresponds to the normalization condition discussed earlier. Given (6), (7) and (10), the objective functional (8) is the asymptotic MSE of the M-estimator based on the score function $\eta(\cdot,\theta)$.

The space H can be converted to a Hilbert space by defining the inner product $(x|y) = E_\theta\ x(z)'\ Q\ y(z)$, for any $x,y\in H$. The induced norm of a vector $x\in H$ is given by $\|x\| = (x|x)^{1/2} = [E_\theta\ x(z)'\ Q\ x(z)]^{1/2}$. Problem (8)-(10) can therefore be rewritten as the following minimum norm problem

$$\begin{aligned} &\text{Min}_{\ \eta\in H}\ \|\eta\|^2 \\ &\text{s.t.} \quad (e_i|\eta) = 0, \qquad i = 1,\ldots,p, \\ &\qquad\quad (s_{ij}|\eta) = q_{ij}, \qquad i,j = 1,\ldots,p, \end{aligned}$$

where the function e_i maps $z\in\mathfrak{X}$ into a p-vector with a one in the i-th position and zeros everywhere else, and the function s_{ij} maps $z \in \mathfrak{X}$ into a p-vector with the i-th component of the likelihood score $s(z,\theta)$ in the j-th position and zeros everywhere else. The dependence of the function $\eta(\cdot,\theta)$ on θ has been omitted for simplicity. The constraints $(e_i|\eta) = 0$, $i = 1,\dots,p$, and $(s_{ij}|\eta) = 0$, $i,j = 1,\dots,p$, define a closed subspace M of H. Since the matrix $[(s_{ij}|s_{ji})] = J(\theta) \otimes Q$ is p.d. and $(e_i|s_{ij}) = 0$, the functions $\{e_i\}$ and $\{s_{ij}\}$ are linearly independent vectors in H. Thus the constraints (9) and (10) define a hyperplane, i.e. a translation of M, of codimension $p(p + 1)$. It then follows from the Projection Theorem for Hilbert spaces [see e.g Luenberger (1969), Section 3.10, Thm. 2] that problem (8)-(10) has a unique solution $\tilde{\eta}(\cdot,\theta)$. Moreover, the optimal score function $\tilde{\eta}(\cdot,\theta)$ is orthogonal to M and can be characterized as a linear combination of the functions $\{e_i\}$ and $\{s_{ij}\}$

$$\eta(\cdot,\theta) = \sum_{i=1}^{p} \beta_i e_i + \sum_{i=1}^{p} \sum_{j=1}^{p} \gamma_{ij} s_{ij} = \beta + \Gamma\, s(\cdot,\theta),$$

where the p-vector $\beta = (\beta_1,\dots,\beta_p)$ and the pxp matrix $\Gamma = [\gamma_{ij}]$ are the solution to the normal equations

$$\begin{bmatrix} I_p & 0' \\ 0 & J(\theta) \otimes Q \end{bmatrix} \begin{bmatrix} \beta \\ \text{vec}\,\Gamma \end{bmatrix} = \begin{bmatrix} 0 \\ \text{vec}\,Q \end{bmatrix}.$$

Clearly, $\beta = 0$ and $\Gamma = J(\theta)^{-1}$. The optimal solution to (8)-(10) is therefore given by $\tilde{\eta}(\cdot,\theta) = J(\theta)^{-1} s(\cdot,\theta)$ [6]. Since $J(\theta)^{-1}$ exists for all θ in a neighborhood Θ_0 of θ_0, the function $\tilde{\eta}(\cdot,\theta)\colon \mathfrak{X} \longrightarrow \mathbb{R}^p$ can be extended to a function $\tilde{\eta}\colon \mathfrak{X} \times \Theta_0 \longrightarrow \mathbb{R}^p$. The resulting estimator is obviously the ML estimator. Finally, since the optimal score function does not depend on the choice of the matrix Q, the ML estimator is optimal in a strong sense, that is, with respect to the partial ordering of p.d. matrices.

4. Conclusions

In this paper we have presented a simple constructive proof of the strong optimality of the ML estimator. Unfortunately, the ML estimator may loose its optimality properties for even small departures from the assumed parametric model. This problem is particularly severe when the assumed model is Gaussian. One advantage of our approach is that it can easily be generalized to incorporate, in addition to the classical requirements of consistency and efficiency, robustness constraints such as an upper bound on the sup-norm of the IF [see e.g. Hampel et al. (1986) and Peracchi (1987)]. Additional robustness constraints might also be introduced. For example, the sup-norm of the gradient of the IF, called the local-shift sensitivity [Mallows (1975)], provides a quantitative measure of the effects of rounding or grouping of the observations. The asymptotic variance of an estimator can itself be represented as a functional, and its Gateaux derivative in the direction of a point mass distribution, called the change-of-variance function [Hampel, Rousseeuw and Ronchetti (1981)], provides a quantitative measure of the stability of the asymptotic variance of an estimator under a small perturbation of the underlying d.f. of the observations. Problem (8)-(10) could then be further generalized to incorporate upper bounds on the local shift sensitivity and/or the change-of-variance function. The general form of the optimal solution in these cases is yet unknown.

References

Burguete, J., Gallant, A.R., and Souza, G. (1982), "On unification of asymptotic theory of non-linear econometrics", *Econometric Reviews*, 1, 151-190.

Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics*, Chapman and Hall, London.

Fernholz, L.T. (1983), *Von Mises Calculus for Statistical Functionals*, Lecture Notes in Statistics No. 19, Springer, New York.

Hampel, F.R. (1974), "The influence curve and its role in robust estimation", *Journal of the American Statistical Association*, 69, 383-393.

Hampel, F.R., Rousseeuw, P.J. and Ronchetti, E. (1981), "The change-of-variance curve and optimal redescending M-estimators", *Journal of the American Statistical Association*, 76, 643-648.

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

Hansen, L.P. (1982), "Large sample properties of generalized method of moments estimators", *Econometrica*, 50, 1029-1054.

Luenberger, D. (1969), *Optimization by Vector Space Methods*, Wiley, New York.

Mallows, C.L. (1975), "On some topics in robustness", Technical Memorandum, Bell Telephone Laboratories, Murray Hill, N.J.

Peracchi, F. (1987), Bounded Influence Methods in Econometrics with an Application to the Censored Regression Model, Ph.D. Dissertation, Princeton University.

Prakasa Rao, B.L.S. (1987), *Asymptotic Theory of Statistical Inference*, Wiley, New York.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Footnotes

[1] An alternative proof, based on Lagrangean methods, is given in Peracchi (1987).

[2] The class of M-estimators is very large and includes most common econometric estimators, such as ML, pseudo-ML, least squares, least absolute deviations and generalized method of moments estimators [see e.g. Burguete, Gallant and Souza (1982) and Hansen (1982)]. For example, if $\eta_N(z,\theta) = s(z,\theta)$ then $\hat{\theta}_N$ is the ML estimator based on model $\mathcal{F}$.

[3] The following notation is used. E_0 denotes expectations taken with respect to F_0, the true d.f. of a single observation. Expectations with respect to some d.f. in the parametric family $\{F_\theta\}$ are denotes by E_θ, and expectations taken with respect to any other d.f. F by E_F.

[4] A sufficient condition for (4) is some kind of differentiability of the functional $\hat{\theta}$, such as Frechét differentiability [see Serfling (1980) and Huber (1981)], or the weaker Hadamard differentiability [see Fernholz (1983) and Prakasa Rao (1987)].

[5] The IF plays a key role in the literature on robustness, where it is interpreted as a measure of the asymptotic bias of $\hat{\theta}$, as an estimator of $\hat{\theta}(F)$, arising from an arbitrarily small contamination of the d.f. F by a point mass.

[6] Alternatively, the optimal solution can be characterized as the saddle-point of a Lagrangean function [Peracchi (1987)]. In this case the inverse of the information matrix can be shown to be equal to the matrix of Lagrange multipliers associated with the constraint (10). On the other hand, since $E_\theta\ s(z,\theta) = 0$, the Fisher consistency constraint is not binding and so the associated Lagrange multiplier vector is equal to zero.