Generic Uniqueness of the Population Quasi Maximum Likelihood Parameters*

Whitney K. Newey

November, 1986

Department of Economics
Princeton University
Princeton, NJ 08544

*This research was partially supported by NSF grant 140-6075 and Bell Communications Research. The author is grateful to Roger Klein for useful discussions.

There has been much recent work in econometrics concerning the properties of maximum likelihood estimators in models that are possibly misspecified (e. g. White, 1982). One of the important assumptions that is made in such work is that the value of the parameters that maximizes the population expectation of the log-likelihood is unique. When this parameter value is unique the maximum likelihood estimator can be interpreted as a consistent, asymptotically normal estimator of this parameter value. When this parameter value is not unique there is an identification problem concerning this parameter.

In correctly specified models the uniqueness of the parameter values that maximizes the expected log-likelihood is a straightforward implication of dependence of the data generating value on the value of the parameter. Also, in some models where the likelihood may be specified but some structure is still imposed on the data generating process (e.g. Gourieroux, Monfort, and Trognon, 1984) identifcation results from certain natural conditions. However, in fully misspecified models identification is essentially an unverifiable assumption, involving as it does the unknown data generating process. Nevertheless, it is still often very convenient to impose an assumption of that the population quasi maximum likelihood parameter values are unique (e.g. Newey and McDonald, 1987).

The purpose of this note is to show that if a model satisfies a set of readily verifiable conditions then identification of the quasi maximum likelihood parameter values will hold for most data generating processes, where the technical meaning of the word "most" will be discussed below. This observation can help to justify the identification assumption that is often made. In models where the conditions discussed below apply it will be difficult (although not impossible) to construct examples of data-generating processes where identification does not hold.

In this note the assumptions for the model will be presented

and discussed, the generic identification result presented and discussed, and the proof given.

The type of estimation environment that will be explicitly considered in this note is that of quasi maximum likelihood estimation with i.i.d. observations. Let $f(z|\theta)$, $\theta \in \Theta$ denote a family of probability density functions with respect to a measure μ, where θ is a $q \times 1$ vector of parameters and z is a $p \times 1$ data vector. Let $z_1, \ldots, z_n$ denote a random sample of data drawn from the c.d.f. $G(z)$. The quasi-log likelihood is given by $Q_n(\theta) \equiv \Sigma_{t=1}^{n} \ln f(z_t|\theta)/n$, where the term quasi refers to the situation where $G(z)$ need not be the c.d.f. corresponding to $f(z|\theta)$ for any θ. The quasi-maximum likelihood estimator (QMLE) is defined as

$$(1) \qquad \hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} Q_n(\theta).$$

The asymptotic properties of the QMLE are determined in part by the behavior of the population expectation of the log-likelihood. Define

$$(2) \qquad Q(\theta, G) \equiv \int \ln f(z|\theta) dG(z), \quad \theta(G) \equiv \operatorname{argmax}_{\theta \in \Theta} Q(\theta, G),$$

where the assumptions given below will be sufficient to guarantee that these expressions exist. The set of quasi-true parameters $\theta(G)$ is the population quantity that corresponds to the QMLE.

One of the conditions that is useful to impose in determining the properties of $\theta(G)$ is that $f(z|\theta)$ satisfy a standard set of regularity conditions:

Assumption F: The support Z of $f(z|\theta)$ does not depend on θ and $\ln f(z|\theta)$ is twice differentiable in θ for all $z \in Z$ and $\theta \in \Theta$. Also, $\ln f(z|\theta)$ and its first and second partial derivatives are continuous in (z, θ) on $Z \times \Theta$ for all $\theta \in \Theta$ and there exist functions $\alpha(z)$ and $\gamma(z)$, where $\alpha(z)$ is continuous on Z and a constant $\delta > 0$ such that for all $\theta \in \Theta$,

(3) $$|f(z|\theta)| \leq \gamma(z), \qquad |\ln f(z|\theta)|^{1+\delta} \leq \alpha(z),$$

$$|\partial \ln f(z|\theta)/\partial\theta|^2 \leq \alpha(z), \qquad |\partial^2 \ln f(z|\theta)/\partial\theta\partial\theta'| \leq \alpha(z),$$

$$\int \gamma(z)d\mu < \infty, \qquad \int \alpha(z)^{1+\delta}\gamma(z)d\mu < \infty.$$

Among other things this assumption implies that the equality between the outer product and Hessian versions of the information matrix holds, and that the QMLE will be consistent and asymptotically normal uniformly in θ when the data is actually generated by $f(z|\theta)$ and θ is identified. The assumption concerning continuity in z is not as restrictive as it may first appear, since this continuity is only required to hold on the support Z of z_t. For example, the likelihoods of many limited dependent variable models will satisfy this restriction when z_t includes indicator functions for various possible regimes of the observed dependent variables.

It is also useful to impose a regularity condition on the set of distributions that have a density with respect to the dominating measure μ.

Assumption D: For any G with finite support S that is a subset of Z there exists a sequence $\{G_k\}$ converging in distribution to G such that each G_k is absolutely continuous with respect μ and has support contained in a compact set which does not vary with k.

This assumption will be satisfied in most applications. For example, if μ is equal to Lebesgue measure on the real line then for any random variable that takes on a finite number of values there always exists a sequence of continuous distributions with uniformly bounded support that converges in distribution to this random variable.

Another condition that will be imposed is the usual compactness assumption for the parameter space Θ, along with an assumption concerning the smoothness of its boundary:

Assumption T: Θ is compact and for any boundary point θ^* of Θ, Θ can locally be represented as $\{\theta : s(\theta) \leq 0\}$ where $s(\theta)$ is a vector of twice continuously differentiable functions such that $s(\theta^*) = 0$ and $\partial s(\theta^*)/\partial\theta$ has full row rank.

This assumption restricts Θ to be of full dimension locally at each point. This assumption is not as restrictive as it might first appear, because in other cases (e. g. when Θ is the set of values such that the Euclidean norm of θ is equal to some constant), it will often be possible to construct a local parameterization that is of full dimension. It is possible to state a more general assumption that would give the following results, but this assumption is fairly simple and includes many cases of interest, such as the case where Θ is a sphere or a box.

A condition that is obviously essential to existence of a unique maximum for $Q(\theta,G)$ is identification of the parameters θ for the likelihood $f(z|\theta)$. Without such an assumption the likelihood could coincide for two different parameter values (i. e. $f(z|\theta) = f(z|\theta')$ identically in z for $\theta \neq \theta'$) and $\theta(G)$ might not consist of a single point. Also, the local identification assumption of nonsingularity of the information matrix is essential for asymptotic normality of the QMLE in a correctly specified model. The following assumption imposes standard local and global identifiability conditions on $f(z|\theta)$. Let $1(A)$ denote the indicator function for the event A.

Assumption I: For each θ and θ' in Θ, $\int 1[f(z|\theta)\neq f(z|\theta')]f(z|\theta)d\mu > 0$ and $\int[\partial^2 \ln f(z|\theta)/\partial\theta\partial\theta']f(z|\theta)d\mu$ is nonsingular.

Together, Assumptions F, T, and I form a set of regularity and identification conditions that can be checked in a straightforward way for a particular model. The main result of this paper is to show that these assumptions are sufficient to imply that the set of maxima of the expected quasi-log likelihood will consist of a single, locally

identified point for most distributions. To state this result is necessary to state what is meant by local identification for the expected quasi likelihood and to state what is meant by most distributions. The local identfication condition that will be considered is:

Condition L: Either θ^* is an interior point of Θ with $Q_\theta(\theta^*,G) = 0$ and $Q_{\theta\theta}(\theta^*,G)$ negative definite, or θ^* is a boundary point with $Q(\theta^*,G) = s_\theta(\theta^*)'\lambda$ for some vector λ consisting entirely of positive elements and $\Delta' Q_{\theta\theta}(\theta^*,G)\Delta < 0$ for all $\Delta \neq 0$ such that $\Delta' s_\theta(\theta^*) = 0$. Also, $\int[\partial \ln f(z|\theta^*)/\partial\theta][\partial \ln f(z|\theta^*)/\partial\theta]' dG(z)$ is nonsingular.

The approach to formalizing the statement that uniqueness of $\theta(G)$ holds for most distributions is to consider a class of distributions and a metric for this class of distributions and to show that uniqueness holds on an open and dense subset of this class of distributions. It is desirable to restrict the set of distributions to be absolutely continuous with respect to μ. For example, this restriction would gaurantee that variables that are continuously distributed under $f(z|\theta)$ are also continuously distributed under $G(z)$. Also, to guarantee that the function $Q(\theta,G)$ and its derivatives are well behaved (exist and are continuous in G) it is useful to impose conditions such that the expectation of the dominating function $\alpha(z)$ is bounded. Let B be any constant such that $B > \int \alpha(z)^{1+\delta}\gamma(z)d\mu$. The class of distributions that will be considered is

$$\mathcal{G}(B) = \{G : G \text{ is absolutely continuous with respect to } \mu \text{ and } \int \alpha(z)^{1+\delta} dG(z) < B\}.$$

The metric on $\mathcal{G}(B)$ that will be considered is the variational distance

$$d(G,G') = \sup_A |\int_A dG - \int_A dG'|.$$

The main result of this note can now be stated as:

Theorem 1: If Assumptions F, D, T, and I are satisfied then the subset $\mathcal{G}(B)$ such that $\theta(G)$ consists of a single point θ^* where Condition L is satisfied is open and dense in $\mathcal{G}(B)$

It is natural to ask to what extent this result depends on the choice of metric on the set of distributions, which corresponds to a choice of a family of neighborhoods for any particular distribution. It is possible to change the family of neighborhoods somewhat and retain this result. For example, if the metric is chosen to correspond to convergence in distribution and $\mathcal{G}(B)$ is further restricted to the class of distributions of z_t that are tight then this result will still hold. What is not known is whether or not it is possible to enlarge or shrink the family of neighborhoods and retain this result.

It is straightforward to extend this result to other types of misspecified estimation environments. For example, consider an estimator $\tilde{\theta} = \text{argmax}_{\theta\in\Theta}\Sigma_{t=1}^{n}\rho(z_t,\theta)/n$ for some $\rho(z,\theta)$ function that need not be a quasi log likelihood. This estimator is of the general type considered by Huber (1967). It is possible to show that under conditions regularity conditions analagous to those of Assumption F and a global and local identification condition under correct specification (analagous to assumption I) that the population parameter value which is estimated by $\tilde{\theta}$, namely the maximum of $\int\rho(z,\theta)dG(z)$, will be unique and locally identified for most distributions.

Proof of Theorem 1: Openness will first be shown. Consider some distribution G^* for which $\theta(G^*) = \{\theta^*\}$ and Condition L is satisfied at θ^*. Let G_k be a sequence of distributions converging to G^*. For any distribution G in $\mathcal{G}(B)$ let g denote its density relative to μ. Then we have

$$(4)\quad |Q(\theta,G_k)-Q(\theta,G^*)| \le \int |\ln f(z|\theta)|\,|g_k(z)-g^*(z)|\,d\mu$$

$$\le \int \alpha(z)\,|g_k(z)-g^*(z)|\,d\mu$$

$$\le \int 1(\alpha(z)>K)\alpha(z)[g_k(z)+g^*(z)]\gamma(z)\,d\mu + K2d(G_k,G^*),$$

$$\le \frac{1}{K}\int 1(\alpha(z)>K)(\alpha(z)/K)^{1+\delta}[g_k(z)+g^*(z)]\gamma(z)\,d\mu + K2d(G_k,G^*),$$

$$\le K^{-\delta}2B + K2d(G_k,G^*).$$

Since $K^{-\delta}2B$ can be made arbitrarily small by choosing K large enough and $d(G_k,G^*)$ goes to zero it follows that $Q(\theta,G_k)$ converges to $Q(\theta,G^*)$, where this convergence is uniform in θ since the last line of this equation does not depend on θ. Also, it follows by an identical argument that the first and second derivatives of $Q(\theta,G_k)$ converge uniformly in θ to the first and second derivatives of $Q(\theta,G^*)$, respectively, so that $Q(\theta,G)$ and its first and second partial derivatives are jointly continuous at (θ^*,G^*), where $\theta(G^*) = \{\theta^*\}$. Convergence of $\theta(G_k)$ to $\theta(G^*) = \{\theta^*\}$ follows by the theorem of the maximum. For θ^* an interior point it follows from Condition L satisfied at (θ^*,G^*) and uniform convergence that $Q(\theta,G_k)$ will be strictly concave in a neighborhood of θ^* for large enough k, implying that $\theta(G_k)$ must be a singleton. Consider the case where θ^* is a boundary point of Θ. By Assumption T it is possible to, without loss of generality, consider a local reparameterization such that $\theta^* = 0$ and in a neighborhood $\mathcal{N}$ of θ the set Θ is parameterized by $\{\theta : \theta_1 \le 0\}$ for some partition $\theta = (\theta_1',\theta_2')'$. Condition L for this local reparameterization implies that all the elements of $\partial Q(\theta^*,G^*)/\partial\theta_1$ are negative, which must also hold for $Q_\theta(\theta,G_k)/\partial\theta_1$ on a neighborhood of θ^* for large enough k, implying that all elements of $\theta(G_k)$ must have $\theta_1 = 0$ for large enough k, so that $\theta(G_k)$ a singleton for large enough k follows by applying the interior case argument to θ_2 and $Q(0,\theta_2,G_k)$.

To show denseness, consider G such that $\Theta(G)$ does not consist of a singelton and consider $\varepsilon > 0$. Let θ^* denote an element of $\Theta(G)$ and let $g_\delta(z) = (1-\delta)g(z) + \delta f(z|\theta^*)$ for $0 \le \delta \le 1$ and G_δ denote the corresponding c.d.f.. Note that

$$(5) \quad d(G,G_\delta) = \delta \int |g(z)-f(z|\theta^*)|d\mu \le 2\delta,$$

so that $d(G,G_\delta) < \varepsilon/2$ for $\delta < \varepsilon/4$. Furthermore, by Assumption I and the information inequality (e.g. Rao, 1973, eq. 1e.6.6), $Q(\theta,G_1) < Q(\theta^*,G_1)$ for all $\theta \neq \theta^*$, so that for $\delta > 0$ and $\theta \neq \theta^*$,

$$(6) \quad Q(\theta,G_\delta) = (1-\delta)Q(\theta,G) + \delta Q(\theta,G_1)$$

$$< (1-\delta)Q(\theta^*,G) + \delta Q(\theta^*,G_1) = Q(\theta^*,G_\delta),$$

so that $\Theta(G_\delta) = \{\theta^*\}$. If θ^* is an interior point then the first and second order conditions of Condition L follow immediately from the first and second order necessary conditions for θ^* to be a maximum of $Q(\theta,G)$, which are $Q_\theta(\theta^*,G) = 0$ and $Q_{\theta\theta}(\theta^*,G)$ negative semi definite, and from the nonsingularity of the information matrix that is posited in Assumption I.

It remains to consider the case where θ^* is a boundary point. Consider a local reparameterization of θ at θ^* as used in the proof of openness. From the first order necessary conditions it can be assumed without loss of generality that for some partition $\theta_1 = (\theta_{11}', \theta_{12}')$

$$(7) \quad Q_{\theta_{11}}(0,G) \ll 0, \quad Q_{\theta_{12}}(0,G) = 0, \quad Q_{\theta_2}(0,G) = 0,$$

where $\ll$ denotes "less than for each element". Partition $\theta = (\theta_{11}', \theta_b')$. The second-order necessary (sufficient) condition for a local maximum is

(8) $\theta_b' Q_{\theta_b \theta_b}(0,G)\theta_b \leq (<)\ 0$ for $\|\theta_b\| = 1$, $\theta_{12} \leq 0$,

as can readily be verified by a second-order mean value expansion.

To begin the construction of the required distribution that is close to G, let G^* denote the distribution function for $f(z|\theta^*)$ and let $s(z,\theta) = \partial \ln f(z|\theta)/\partial\theta$ denote the score vector. By Assumptions F and I, $Q_\theta(\theta^*,G^*) = \int s(z,\theta^*)dG^* = 0$ and $\int s(z,\theta)s(z,\theta)'dG^* = Q_{\theta\theta}(\theta^*,G^*)$ is nonsingular. It follows from Chamberlain (1987, Lemma 3) that there exists $\{z_1,\ldots,z_m\} \subseteq Z$ and p_i, $(i=1,\ldots,m)$, such that $p_i \geq 0$, $\Sigma_{i=1}^m p_i = 1$ and, for $S = [s(z_1,\theta^*),\ldots,s(z_m,\theta^*)]'$, and P be the diaganol matrix $P = \mathrm{diag}[p_1,\ldots,p_m]$,

(9) $p'S = 0$, $S'PS$ is nonsingular.

It follows from this equation that there does not exist a vector λ with nonegative components and a nonzero vector r such that $Sr = \lambda$, since $p'\lambda = p'Sr = 0$ implies $\lambda = 0$, which implies $S'PSr = S'P\lambda = 0$, a violation of nonsingularity. It then follows by Theorem ?? of Rockafellar that there exists a vector $q = (q_1,\ldots,q_m)'$ of nonnegative numbers such that

(10) $q'S_1 \ll 0$, $q'S_2 = 0$.

where $S = [S_1, S_2]$ is partitioned conformably with θ. Note that q can be chosen so that $\Sigma_{i=1}^m q_i = 1$. Let $\bar{G}$ denote the distribution with $\mathrm{Prob}(z=z_i) = q_i$. Consider the family of distributions

(11) $G_{\delta,\delta'} = (1-\delta)G + \delta[(1-\delta')G^* + \delta'\bar{G}]$.

Note that $d(G,G_{\delta,\delta'}) < 4\delta$. Hold δ fixed so that this distance is as small as desired, and for notational convenience supress the δ subscript on $G_{\delta,\delta'}$. As above $\Theta(G_0) = \{0\}$, while by the theorem of the maximum $\lim_{\delta' \to 0} \Theta(G_{\delta'}) = \{0\}$. Also, there is a $\bar{\delta}'$ and a neighborhood $\mathcal{N}$ of 0 on which the eigenvalues of $Q_{\theta\theta}(\theta,(1-\delta')G^*+\delta'\bar{G})$

are bounded negative uniformly in $\delta' \leq \bar{\delta}'$. Therefore, the second order sufficient conditions from equation (8) for $\theta = 0$ to be a unique local maximum of $Q(\theta, G_{\delta'})$ are satisfied uniformly in $\delta' \leq \bar{\delta}'$. Furthermore, the first and second partial derivatives of $Q(\theta, G_{\delta'})$ are continuous uniformly in δ', and by equation (10) $\partial Q(0, G_{\delta'})/\partial\theta_{12} \leq 0$, $\partial Q(0, G_{\delta'})/\partial\theta_2 = 0$, and the components of $\partial Q(0, G_{\delta'})/\partial\theta_{11}$ are bounded negative uniformly in δ'. It follows that 0 is the unique local maximum of $(\theta, G_{\delta'})$ for θ in a neighborhood of 0 that does not depend on δ' for δ' small enough, and is thus the unique global maximum for δ' small enough. Condition L is satisfied at 0 by equations (7) and (9), $\partial Q(0, G^*)/\partial\theta = 0$, and by the sufficient version of (8) with $\theta_{12} = 0$.

By Assumption D there is a sequence $\{G_k\}$ of distributions that are absolutely continuous with respect to μ and have support contained within some fixed compact set $C \subseteq Z$ such that this sequence converges in distribution to $\bar{G}$. Also, by continuity in z and θ of $\ln f(z|\theta)$ and its first and second partial derivatives and compactness of $C \times \Theta$, it follows that for any sequence $\theta_k \to \theta$, $Q(\theta_k, G_k) \to Q(\theta, \bar{G})$, and that the analagous property hold for the first and second partial derivatives of $Q(\theta, G_k)$. By construction, the analagous results hold for $Q(\theta, G_{\delta,\delta',k})$ and $Q(\theta, G_{\delta,\delta'})$ and their respective first and second partial derivatives, where $G_{\delta,\delta',k} = (1-\delta)G + \delta[(1-\delta')G^* + \delta' G_k]$. Note also that by $\alpha(z)$ continuous $\alpha(z)$ is bounded on C, sot that $\int \alpha(z)^{1+\delta} dG_k$ is bounded in k. Thus, $G_{\delta,\delta',k}$ will be an element of $\mathcal{G}(B)$ for all k if δ' is small enough. The conclusion then follows from $d(G, G_{\delta,\delta',k}) \leq 4\delta$ and from the same argument as used to prove openness.

REFERENCES

Chamberlain, G., 1987, "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions, "*Journal of Econometrics*, forthcoming. 34, 305-334

Gourieroux, G., A. Monfort, and A. Trognon, 1984, "Psuedo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681-700.

Huber, P. J., 1967, "The Behavior of Maximum Likelihood Estimators Under Nonstandard Conditions," In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA: University of California.

McDonald, J. B., and W. Newey, 1987, "Partially Adaptive Estimation of Regression Models via the Generalized T Distribution," mimeo, Brigham Young University.

Rao, C. R., 1973, *Linear Statistical Inference and Its Applications*, New York: John Wiley and Sons

R.T.
Rockafellar, 1970, *Convex Analysis*. Princeton, N. J.: Princeton University Press.

White, H., 1982, "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1-25.