Chapter 36

LARGE SAMPLE ESTIMATION AND HYPOTHESIS TESTING*

WHITNEY K. NEWEY
Massachusetts Institute of Technology

DANIEL McFADDEN
University of California, Berkeley

Contents

Abstract 2113

1. Introduction 2113

2. Consistency 2120
   2.1. The basic consistency theorem 2121
   2.2. Identification 2124
       2.2.1. The maximum likelihood estimator 2124
       2.2.2. Nonlinear least squares 2125
       2.2.3. Generalized method of moments 2126
       2.2.4. Classical minimum distance 2128
   2.3. Uniform convergence and continuity 2129
   2.4. Consistency of maximum likelihood 2131
   2.5. Consistency of GMM 2132
   2.6. Consistency without compactness 2133
   2.7. Stochastic equicontinuity and uniform convergence 2136
   2.8. Least absolute deviations examples 2138
       2.8.1. Maximum score 2138
       2.8.2. Censored least absolute deviations 2140

3. Asymptotic normality 2141
   3.1. The basic results 2143
   3.2. Asymptotic normality for MLE 2146
   3.3. Asymptotic normality for GMM 2148

*We are grateful to the NSF for financial support and to Y. Ait-Sahalia, J. Porter, J. Powell, J. Robins, P. Ruud, and T. Stoker for helpful comments.

Handbook of Econometrics, Volume IV, Edited by R.F. Engle and D.L. McFadden
© 1994 Elsevier Science B.V. All rights reserved
4. Consistent asymptotic variance estimation 2153
   4.1. The basic results 2155
   4.2. Variance estimation for MLE 2157
   4.3. Asymptotic variance estimation for GMM 2160
5. Asymptotic efficiency 2162
   5.1. Efficiency of maximum likelihood estimation 2162
   5.2. Optimal minimum distance estimation 2164
   5.3. A general efficiency framework 2165
   5.4. Solving for the smallest asymptotic variance 2168
   5.5. Feasible efficient estimation 2171
   5.6. Technicalities 2173
6. Two-step estimators 2175
   6.1. Two-step estimators as joint GMM estimators 2176
   6.2. The effect of first-step estimation on second-step standard errors 2179
   6.3. Consistent asymptotic variance estimation for two-step estimators 2182
7. Asymptotic normality with nonsmooth objective functions 2184
   7.1. The basic results 2185
   7.2. Stochastic equicontinuity for Lipschitz moment functions 2188
   7.3. Asymptotic variance estimation 2189
   7.4. Technicalities 2191
8. Semiparametric two-step estimators 2194
   8.1. Asymptotic normality and consistent variance estimation 2196
   8.2. V-estimators 2200
   8.3. First-step kernel estimation 2203
   8.4. Technicalities 2214
9. Hypothesis testing with GMM estimators 2215
   9.1. The null hypothesis and the constrained GMM estimator 2217
   9.2. The test statistics 2220
   9.3. One-step versions of the trinity 2226
   9.4. Special cases 2228
   9.5. Tests for overidentifying restrictions 2231
   9.6. Specification tests in linear models 2234
   9.7. Specification testing in multinomial models 2236
   9.8. Technicalities 2239
References 2241
Abstract

Asymptotic distribution theory is the primary method used to examine the properties of econometric estimators and tests. We present conditions for obtaining consistency and asymptotic normality of a very general class of estimators (extremum estimators). Consistent asymptotic variance estimators are given to enable approximation of the asymptotic distribution. Asymptotic efficiency is another desirable property then considered. Throughout the chapter, the general results are also specialized to common econometric estimators (e.g. MLE and GMM), and in specific examples we work through the conditions for the various results in detail. The results are also extended to two-step estimators (with finite-dimensional parameter estimation in the first step), estimators derived from nonsmooth objective functions, and semiparametric two-step estimators (with nonparametric estimation of an infinite-dimensional parameter in the first step). Finally, the trinity of test statistics is considered within the quite general setting of GMM estimation, and numerous examples are given.

1. Introduction

Large sample distribution theory is the cornerstone of statistical inference for econometric models. The limiting distribution of a statistic gives approximate distributional results that are often straightforward to derive, even in complicated econometric models. These distributions are useful for approximate inference, including constructing approximate confidence intervals and test statistics. Also, the location and dispersion of the limiting distribution provides criteria for choosing between different estimators. Of course, asymptotic results are sensitive to the accuracy of the large sample approximation, but the approximation has been found to be quite good in many cases and asymptotic distribution results are an important starting point for further improvements, such as the bootstrap. Also, exact distribution theory is often difficult to derive in econometric models, and may not apply to models with unspecified distributions, which are important in econometrics. Because asymptotic theory is so useful for econometric models, it is important to have general results with conditions that can be interpreted and applied to particular estimators as easily as possible. The purpose of this chapter is the presentation of such results.

Consistency and asymptotic normality are the two fundamental large sample properties of estimators considered in this chapter. A consistent estimator $\hat{\theta}$ is one that converges in probability to the true value $\theta_0$, i.e. $\hat{\theta} \overset{P}{\to} \theta_0$, as the sample size $n$ goes to infinity, for all possible true values. This property is sometimes referred to as weak consistency, with strong consistency holding when $\hat{\theta}$ converges almost surely to the true value. Throughout the chapter we focus on weak consistency, although we also show how strong consistency can be proven.
that the estimator is close to the truth when the number of observations is nearly infinite. Thus, an estimator that is not even consistent is usually considered inadequate. Also, consistency is useful because it means that the asymptotic distribution of an estimator is determined by its limiting behavior near the true parameter.

An asymptotically normal estimator $\hat{\theta}$ is one where there is an increasing function $v(n)$ such that the distribution function of $v(n)(\hat{\theta} - \theta_0)$ converges to the Gaussian distribution function with mean zero and variance $V$, i.e. $v(n)(\hat{\theta} - \theta_0) \overset{d}{\to} N(0, V)$. The variance $V$ of the limiting distribution is referred to as the asymptotic variance of $\hat{\theta}$. The estimator is $\sqrt{n}$-consistent if $v(n) = \sqrt{n}$. This chapter focuses on the $\sqrt{n}$-consistent case, so that unless otherwise noted, asymptotic normality will be taken to include $\sqrt{n}$-consistency.

Asymptotic normality and a consistent estimator of the asymptotic variance can be used to construct approximate confidence intervals. In particular, for an estimator $\hat{V}$ of $V$ and for $g_{a/2}$ satisfying $\text{Prob}[\mathcal{N}(0, 1) > g_{a/2}] = \alpha/2$, an asymptotic $1 - \alpha$ confidence interval is

$$\hat{\theta} - g_{a/2}(\hat{V}/n)^{1/2}, \hat{\theta} + g_{a/2}(\hat{V}/n)^{1/2}.$$

If $\hat{V}$ is a consistent estimator of $V$ and $V > 0$, then asymptotic normality of $\hat{\theta}$ will imply that $\text{Prob}(\theta_0 \in \mathcal{I}_{1-\alpha}) \to 1 - \alpha$ as $n \to \infty$. Here asymptotic theory is important for econometric practice, where consistent standard errors can be used for approximate confidence interval construction. Thus, it is useful to know that estimators are asymptotically normal and to know how to form consistent standard errors in applications. In addition, the magnitude of asymptotic variances for different estimators helps choose between estimators in practice. If one estimator has a smaller asymptotic variance, then an asymptotic confidence interval, as above, will be shorter for that estimator in large samples, suggesting preference for its use in applications. A prime example is generalized least squares with estimated disturbance variance matrix, which has smaller asymptotic variance than ordinary least squares, and is often used in practice.

Many estimators share a common structure that is useful in showing consistency and asymptotic normality, and in deriving the asymptotic variance. The benefit of using this structure is that it distills the asymptotic theory to a few essential ingredients. The cost is that applying general results to particular estimators often requires thought and calculation. In our opinion, the benefits outweigh the costs, and so in these notes we focus on general structures, illustrating their application with examples.

One general structure, or framework, is the class of estimators that maximize some objective function that depends on data and sample size, referred to as extremum estimators. An estimator $\hat{\theta}$ is an extremum estimator if there is an

---

2 The proof of this result is an exercise in convergence in distribution and the Slutsky theorem, which states that $Y_n \overset{d}{\to} Y_0$ and $Z_n \overset{p}{\to} c$ implies $Z_n Y_n \overset{d}{\to} c Y_0$. 
Ch. 36: Large Sample Estimation and Hypothesis Testing

objective function $\hat{Q}_n(\theta)$ such that

$$\hat{\theta} \text{ maximizes } \hat{Q}_n(\theta) \text{ subject to } \theta \in \Theta,$$

where $\Theta$ is the set of possible parameter values. In the notation, dependence of $\hat{\theta}$ on $n$ and of $\hat{\theta}$ and $\hat{Q}_n(\theta)$ on the data is suppressed for convenience. This estimator is the maximizer of some objective function that depends on the data, hence the term “extremum estimator”. R.A. Fisher (1921, 1925), Wald (1949), Huber (1967), Jennrich (1969), and Malinvaud (1970) developed consistency and asymptotic normality results for various special cases of extremum estimators, and Amemiya (1973, 1985) formulated the general class of estimators and gave some useful results.

A prime example of an extremum estimator is the maximum likelihood (MLE). Let the data $(z_1, \ldots, z_n)$ be i.i.d. with p.d.f. $f(z|\theta_0)$ equal to some member of a family of p.d.f.'s $f(z|\theta)$. Throughout, we will take the p.d.f. $f(z|\theta)$ to mean a probability function where $z$ is discrete, and to possibly be conditioned on part of the observation $z$. The MLE satisfies eq. (1.1) with

$$\hat{Q}_n(\theta) = n^{-1} \sum_{i=1}^{n} \ln f(z_i|\theta).$$

(1.2)

Here $\hat{Q}_n(\theta)$ is the normalized log-likelihood. Of course, the monotonic transformation of taking the log of the likelihood and normalizing by $n$ will not typically affect the estimator, but it is a convenient normalization in the theory. Asymptotic theory for the MLE was outlined by R.A. Fisher (1921, 1925), and Wald’s (1949) consistency theorem is the prototype result for extremum estimators. Also, Huber (1967) gave weak conditions for consistency and asymptotic normality of the MLE and other extremum estimators that maximize a sample average. 5

A second example is the nonlinear least squares (NLS), where for data $z_i = (y_i, x_i)$ with $E[y|x] = h(x, \theta_0)$, the estimator solves eq. (1.1) with

$$\hat{Q}_n(\theta) = -n^{-1} \sum_{i=1}^{n} [y_i - h(x_i, \theta)]^2.$$

(1.3)

Here maximizing $\hat{Q}_n(\theta)$ is the same as minimizing the sum of squared residuals. The asymptotic normality theorem of Jennrich (1969) is the prototype for many modern results on asymptotic normality of extremum estimators.

3“Extremum” rather than “maximum” appears here because minimizers are also special cases, with objective function equal to the negative of the minimand.

4More precisely, $f(z|\theta)$ is the density (Radon–Nikodym derivative) of the probability measure for $z$ with respect to some measure that may assign measure 1 to some singleton’s, allowing for discrete variables, and for $z = (y, x)$ may be the product of some measure for $y$ with the marginal distribution of $x$, allowing $f(z|\theta)$ to be a conditional density given $x$.

5Estimators that maximize a sample average, i.e. where $\hat{Q}_n(\theta) = n^{-1} \sum_{i=1}^{n} q(z_i, \theta)$, are often referred to as $m$-estimators, where the “$m$” means “maximum-likelihood-like”.
A third example is the generalized method of moments (GMM). Suppose that there is a "moment function" vector \( g(z, \theta) \) such that the population moments satisfy \( E[g(z, \theta_0)] = 0 \). A GMM estimator is one that minimizes a squared Euclidean distance of sample moments from their population counterpart of zero. Let \( \hat{W} \) be a positive semi-definite matrix, so that \((m'\hat{W}m)^{1/2}\) is a measure of the distance of \( m \) from zero. A GMM estimator is one that solves eq. (1.1) with

\[
\hat{Q}_n(\theta) = - \left[ n^{-1} \sum_{i=1}^{n} g(z_i, \theta) \right]' \hat{W} \left[ n^{-1} \sum_{i=1}^{n} g(z_i, \theta) \right].
\] (1.4)

This class includes linear instrumental variables estimators, where \( g(z, \theta) = x' (y - Y' \theta) \), \( x \) is a vector of instrumental variables, \( y \) is a left-hand-side dependent variable, and \( Y \) are right-hand-side variables. In this case the population moment condition \( E[g(z, \theta_0)] = 0 \) is the same as the product of instrumental variables \( x \) and the disturbance \( y - Y' \theta_0 \) having mean zero. By varying \( \hat{W} \) one can construct a variety of instrumental variables estimators, including two-stage least squares for \( \hat{W} = (n^{-1} \sum_{i=1}^{n} x_i x_i')^{-1} \).\(^6\) The GMM class also includes nonlinear instrumental variables estimators, where \( g(z, \theta) = x \cdot \rho(z, \theta) \) for a residual \( \rho(z, \theta) \), satisfying \( E[x \cdot \rho(z, \theta_0)] = 0 \). Nonlinear instrumental variable estimators were developed and analyzed by Sargan (1959) and Amemiya (1974). Also, the GMM class was formulated and general results on asymptotic properties given in Burguete et al. (1982) and Hansen (1982).

The GMM class is general enough to also include MLE and NLS when those estimators are viewed as solutions to their first-order conditions. In this case the derivatives of \( \ln f(x|\theta) \) or \(-[y - h(x, \theta)]^2\) become the moment functions, and there are exactly as many moment functions as parameters. Thinking of GMM as including MLE, NLS, and many other estimators is quite useful for analyzing their asymptotic distribution, but not for showing consistency, as further discussed below.

A fourth example is classical minimum distance estimation (CMD). Suppose that there is a vector of estimators \( \hat{\pi} \rightarrow \pi_0 \) and a vector of functions \( h(\theta) \) with \( \pi_0 = h(\theta_0) \). The idea is that \( \pi \) consists of "reduced form" parameters, \( \theta \) consists of "structural" parameters, and \( h(\theta) \) gives the mapping from structure to reduced form. An estimator of \( \theta \) can be constructed by solving eq. (1.1) with

\[
\hat{Q}_n(\theta) = - [\hat{\pi} - h(\theta)]' \hat{W} [\hat{\pi} - h(\theta)],
\] (1.5)

where \( \hat{W} \) is a positive semi-definite matrix. This class of estimators includes classical minimum chi-square methods for discrete data, as well as estimators for simultaneous equations models in Rothenberg (1973) and panel data in Chamberlain (1982). Its asymptotic properties were developed by Chiang (1956) and Ferguson (1958).

A different framework that is sometimes useful is minimum distance estimation,
a class of estimators that solve eq. (1.1) for \( \hat{Q}_n(\theta) = -\hat{g}_n(\theta)\hat{W}\hat{g}_n(\theta) \), where \( \hat{g}_n(\theta) \) is a vector of the data and parameters such that \( \hat{g}_n(\theta_0) \overset{P}{\to} 0 \) and \( \hat{W} \) is positive semi-definite. Both GMM and CMD are special cases of minimum distance, with \( \hat{g}_n(\theta) = n^{-1}\sum_{i=1}^{n}g(z_i, \theta) \) for GMM and \( \hat{g}_n(\theta) = \pi h(\theta) \) for CMD.\(^7\) This framework is useful for analyzing asymptotic normality of GMM and CMD, because (once) differentiability of \( \hat{g}_n(\theta) \) is a sufficient smoothness condition, while twice differentiability is often assumed for the objective function of an extremum estimator [see, e.g. Amemiya (1985)]. Indeed, as discussed in Section 3, asymptotic normality of an extremum estimator with a twice differentiable objective function \( Q_\theta(\theta) \) is actually a special case of asymptotic normality of a minimum distance estimator, with \( \hat{g}_n(\theta) = V_\theta\hat{Q}_n(\theta) \) and \( \hat{W} \) equal to an identity matrix, where \( V_\theta \) denotes the partial derivative. The idea here is that when analyzing asymptotic normality, an extremum estimator can be viewed as a solution to the first-order conditions \( V_\theta\hat{Q}_n(\theta) = 0 \), and in this form is a minimum distance estimator.

For consistency, it can be a bad idea to treat an extremum estimator as a solution to first-order conditions rather than a global maximum of an objective function, because the first-order condition can have multiple roots even when the objective function has a unique maximum. Thus, the first-order conditions may not identify the parameters, even when there is a unique maximum to the objective function. Also, it is often easier to specify primitive conditions for a unique maximum than for a unique root of the first-order conditions. A classic example is the MLE for the Cauchy location-scale model, where \( z \) is a scalar, \( \mu \) is a location parameter, \( \sigma \) a scale parameter, and \( f(z|\theta) = C\sigma^{-1}(1 + [(z - \mu)/\sigma]^2)^{-1} \) for a constant \( C \). It is well known that, even in large samples, there are many roots to the first-order conditions for the location parameter \( \mu \), although there is a global maximum to the likelihood function; see Example 1 below. Econometric examples tend to be somewhat less extreme, but can still have multiple roots. An example is the censored least absolute deviations estimator of Powell (1984). This estimator solves eq. (1.1) for \( \hat{Q}_n(\theta) = -n^{-1}\sum_{i=1}^{n}|y_i - \max\{0, x_i'\theta\}| \), where \( y_i = \max\{0, x_i'\theta_0 + e_i\} \), and \( e_i \) has conditional median zero. A global maximum of this function over any compact set containing the true parameter will be consistent, under certain conditions, but the gradient has extraneous roots at any point where \( x_i'\theta < 0 \) for all \( i \) (e.g. which can occur if \( x_i \) is bounded).

The importance for consistency of an extremum estimator being a global maximum has practical implications. Many iterative maximization procedures (e.g. Newton–Raphson) may converge only to a local maximum, but consistency results only apply to the global maximum. Thus, it is often important to search for a global maximum. One approach to this problem is to try different starting values for iterative procedures, and pick the estimator that maximizes the objective from among the converged values. As long as the extremum estimator is consistent and the true parameter is an element of the interior of the parameter set \( \Theta \), an extremum estimator will be

\(^7\)For GMM, the law of large numbers implies \( \hat{g}_n(\theta_0) \overset{P}{\to} 0 \).
a root of the first-order conditions asymptotically, and hence will be included among the local maxima. Also, this procedure can avoid extraneous boundary maxima, e.g. those that can occur in maximum likelihood estimation of mixture models.

Figure 1 shows a schematic, illustrating the relationships between the various types of estimators introduced so far. The name or mnemonic for each type of estimator (e.g. MLE for maximum likelihood) is given, along with objective function being maximized, except for GMM and CMD where the form of \( g_n(\theta) \) is given. The solid arrows indicate inclusion in a class of estimators. For example, MLE is included in the class of extremum estimators and GMM is a minimum distance estimator. The broken arrows indicate inclusion in the class when the estimator is viewed as a solution to first-order conditions. In particular, the first-order conditions for an extremum estimator are \( \nabla_\theta \hat{Q}_n(\hat{\theta}) = 0 \), making it a minimum distance estimator with \( \hat{g}_n(\theta) = \nabla_\theta \hat{Q}_n(\theta) \) and \( \hat{W} = I \). Similarly, the first-order conditions for MLE make it a GMM estimator with \( g(z, \theta) = \nabla_\theta \ln f(z|\theta) \) and those for NLS a GMM estimator with \( g(z, \theta) = -2[y - h(x, \theta)]\nabla_\theta h(x, \theta) \). As discussed above, these broken arrows are useful for analyzing the asymptotic distribution, but not for consistency. Also, as further discussed in Section 7, the broken arrows are not very useful when the objective function \( \hat{Q}_n(\theta) \) is not smooth.

The broad outline of the chapter is to treat consistency, asymptotic normality, consistent asymptotic variance estimation, and asymptotic efficiency in that order. The general results will be organized hierarchically across sections, with the asymptotic normality results assuming consistency and the asymptotic efficiency results assuming asymptotic normality. In each section, some illustrative, self-contained examples will be given. Two-step estimators will be discussed in a separate section, partly as an illustration of how the general frameworks discussed here can be applied and partly because of their intrinsic importance in econometric applications. Two later sections deal with more advanced topics. Section 7 considers asymptotic normality when the objective function \( \hat{Q}_n(\theta) \) is not smooth. Section 8 develops some asymptotic theory when \( \hat{\theta} \) depends on a nonparametric estimator (e.g. a kernel regression, see Chapter 39).

This chapter is designed to provide an introduction to asymptotic theory for nonlinear models, as well as a guide to recent developments. For this purpose,
Sections 2–6 have been organized in such a way that the more basic material is collected in the first part of each section. In particular, Sections 2.1–2.5, 3.1–3.4, 4.1–4.3, 5.1, and 5.2, might be used as text for part of a second-year graduate econometrics course, possibly also including some examples from the other parts of this chapter.

The results for extremum and minimum distance estimators are general enough to cover data that is a stationary stochastic process, but the regularity conditions for GMM, MLE, and the more specific examples are restricted to i.i.d. data. Modeling data as i.i.d. is satisfactory in many cross-section and panel data applications. Chapter 37 gives results for dependent observations.

This chapter assumes some familiarity with elementary concepts from analysis (e.g. compact sets, continuous functions, etc.) and with probability theory. More detailed familiarity with convergence concepts, laws of large numbers, and central limit theorems is assumed, e.g. as in Chapter 3 of Amemiya (1985), although some particularly important or potentially unfamiliar results will be cited in footnotes. The most technical explanations, including measurability concerns, will be reserved to footnotes.

Three basic examples will be used to illustrate the general results of this chapter.

Example 1.1 (Cauchy location-scale)

In this example \( z \) is a scalar random variable, \( \theta = (\mu, \sigma)' \) is a two-dimensional vector, and \( z \) is continuously distributed with p.d.f. \( f(z|\theta_0) \), where \( f(z|\theta) = C \cdot \sigma^{-1} \{1 + [(z - \mu)/\sigma]^2\}^{-1} \) and \( C \) is a constant. In this example \( \mu \) is a location parameter and \( \sigma \) a scale parameter. This example is interesting because the MLE will be consistent, in spite of the first-order conditions having many roots and the nonexistence of moments of \( z \) (e.g. so the sample mean is not a consistent estimator of \( \theta_0 \)).

Example 1.2 (Probit)

Probit is an MLE example where \( z = (y, x') \) for a binary variable \( y, y \in \{0, 1\} \), and a \( q \times 1 \) vector of regressors \( x \), and the conditional probability of \( y \) given \( x \) is \( f(z|\theta_0) \) for \( f(z|\theta) = \Phi(x'|\theta)\{1 - \Phi(x'|\theta)\}^{1-\gamma} \). Here \( f(z|\theta_0) \) is a p.d.f. with respect to integration that sums over the two different values of \( y \) and integrates over the distribution of \( x \), i.e. where the integral of any function \( a(y, x) \) is \( \int a(y, x) \, dz = E[a(1, x)] + E[a(0, x)] \). This example illustrates how regressors can be allowed for, and is a model that is often applied.

Example 1.3 (Hansen–Singleton)

This is a GMM (nonlinear instrumental variables) example, where \( g(z, \theta) = x \cdot \rho(z, \theta) \) for \( \rho(z, \theta) = \beta \cdot w \cdot y^\gamma - 1 \). The functional form here is from Hansen and Singleton (1982), where \( \beta \) is a rate of time preference, \( \gamma \) a risk aversion parameter, \( w \) an asset return, \( y \) a consumption ratio for adjacent time periods, and \( x \) consists of variables.
in the information set, of an agent maximizing expected constant relative risk aversion utility. This example is interesting because it illustrates the difficulty of specifying primitive identification conditions for GMM and the type of moment existence assumptions that are often useful.

2. Consistency

To motivate the precise conditions for consistency it is helpful to sketch the ideas on which the result is based. The basic idea is that if \( \hat{Q}_n(\theta) \) converges in probability to \( Q_0(\theta) \) for every \( \theta \), and \( Q_0(\theta) \) is maximized at the true parameter \( \theta_0 \), then the limit of the maximum \( \hat{\theta} \) should be the maximum \( \theta_0 \) of the limit, under conditions for interchanging the maximization and limiting operations. For example, consider the MLE. The law of large numbers suggests \( \hat{Q}_n(\theta) \overset{P}{\to} Q_0(\theta) = E[\ln f(z|\theta)] \). By the well known information inequality, \( Q_0(\theta) \) has a unique maximum at the true parameter when \( \theta_0 \) is identified, as further discussed below. Then under technical conditions for the limit of the maximum to be the maximum of the limit, \( \hat{\theta} \) should converge in probability to \( \theta_0 \). Sufficient conditions for the maximum of the limit to be the limit of the maximum are that the convergence in probability is uniform and that the parameter set is compact.\(^8\)

These ideas are illustrated in Figure 2. Let \( \varepsilon \) be a small positive number. If \( \hat{Q}_n(\theta) \) lies in the "sleeve" \([Q_0(\theta) - \varepsilon, Q_0(\theta) + \varepsilon]\), for all \( \theta \), then \( \hat{\theta} \) must lie in \([\theta_l, \theta_u]\), i.e. must be "close" to the value \( \theta_0 \) that maximizes \( Q_0(\theta) \). The estimator should then be consistent as long as \( \theta_0 \) is the true parameter value.

It is essential for consistency that the limit \( Q_0(\theta) \) have a unique maximum at the true parameter value. If there are multiple maxima, then this argument will only

---

\(^8\)These ideas are also related to the result that the probability limit of a continuous function is the function of the probability limit. The maximum is a continuous function of \( \{Q(\theta)\} \) where the maximum is unique, in the metric of uniform convergence on a compact set. Thus, if the probability limit, in this metric, of \( \hat{Q}(\theta) \) is \( Q(\theta) \), and the maximum of \( Q(\theta) \) is unique, then the probability limit of \( \hat{\theta} \) is the maximum of the limit \( Q(\theta) \).
lead to the estimator being close to one of the maxima, which does not give consistency (because one of the maxima will not be the true value of the parameter). The condition that $Q_0(\theta)$ have a unique maximum at the true parameter is related to identification.

The discussion so far only allows for a compact parameter set. In theory compactness requires that one know bounds on the true parameter value, although this constraint is often ignored in practice. It is possible to drop this assumption if the function $\hat{Q}_n(\theta)$ cannot rise “too much” as $\theta$ becomes unbounded, as further discussed below.

Uniform convergence and continuity of the limiting function are also important. Uniform convergence corresponds to the feature of the graph that $\hat{Q}_n(\theta)$ was in the “sleeve” for all values of $\theta \in \Theta$. Conditions for uniform convergence are given below.

The rest of this section develops this descriptive discussion into precise results on consistency of extremum estimators. Section 2.1 presents the basic consistency theorem. Sections 2.2–2.5 give simple but general sufficient conditions for consistency, including results for MLE and GMM. More advanced and/or technical material is contained in Sections 2.6–2.8.

2.1. The basic consistency theorem

To state a theorem it is necessary to define precisely uniform convergence in probability, as follows:

**Uniform convergence in probability**: $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$ means $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$.

The following is the fundamental consistency result for extremum estimators, and is similar to Lemma 3 of Amemiya (1973).

**Theorem 2.1**

If there is a function $Q_0(\theta)$ such that (i) $Q_0(\theta)$ is uniquely maximized at $\theta_0$; (ii) $\Theta$ is compact; (iii) $Q_0(\theta)$ is continuous; (iv) $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$, then $\hat{\theta} \xrightarrow{P} \theta_0$.

**Proof**

For any $\varepsilon > 0$ we have with probability approaching one (w.p.a.1) (a) $\hat{Q}_n(\hat{\theta}) > \hat{Q}_n(\theta_0) - \varepsilon/3$ by eq. (1.1); (b) $Q_0(\hat{\theta}) > \hat{Q}_n(\hat{\theta}) - \varepsilon/3$ by (iv); (c) $\hat{Q}_n(\theta_0) > Q_0(\theta_0) - \varepsilon/3$ by (iv). The probability statements in this proof are only well defined if each of $\hat{\theta}, \hat{Q}_n(\hat{\theta}),$ and $\hat{Q}_n(\theta_0)$ are measurable. The measurability issue can be bypassed by defining consistency and uniform convergence in terms of outer measure. The outer measure of a (possibly nonmeasurable) event $\mathcal{E}$ is the infimum of $E[Y]$ over all random variables $Y$ with $Y \geq 1(\mathcal{E})$, where $1(\mathcal{E})$ is the indicator function for the event $\mathcal{E}$. 

9
Therefore, w.p.a.1,

\[
Q_\Theta(\hat{\theta}) > \hat{Q}_n(\hat{\theta}) - \varepsilon/3 > \hat{Q}_n(\theta_0) - 2\varepsilon/3 > Q_\Theta(\theta_0) - \varepsilon.
\]

Thus, for any \(\varepsilon > 0\), \(Q_\Theta(\hat{\theta}) > Q_\Theta(\theta_0) - \varepsilon\) w.p.a.1. Let \(\mathcal{N}\) be any open subset of \(\Theta\) containing \(\theta_0\). By \(\Theta \cap \mathcal{N}^c\) compact, (i), and (iii), \(\sup_{\theta \in \Theta \cap \mathcal{N}^c} \hat{Q}_\Theta(\theta) = Q_\Theta(\theta^*) < Q_\Theta(\theta_0)\) for some \(\theta^* \in \Theta \cap \mathcal{N}^c\). Thus, choosing \(\varepsilon = Q_\Theta(\theta_0) - \sup_{\theta \in \Theta \cap \mathcal{N}^c} \hat{Q}_\Theta(\theta)\), it follows that w.p.a.1 \(Q_\Theta(\hat{\theta}) > \sup_{\theta \in \Theta \cap \mathcal{N}^c} \hat{Q}_\Theta(\theta)\), and hence \(\hat{\theta} \in \mathcal{N}\). Q.E.D.

The conditions of this theorem are slightly stronger than necessary. It is not necessary to assume that \(\hat{\theta}\) actually maximizes the objective function. This assumption can be replaced by the hypothesis that \(\hat{Q}_n(\hat{\theta}) \geq \sup_{\theta \in \Theta} \hat{Q}_n(\theta) + o_p(1)\). This replacement has no effect on the proof, in particular on part (a), so that the conclusion remains true. These modifications are useful for analyzing some estimators in econometrics, such as the maximum score estimator of Manski (1975) and the simulated moment estimators of Pakes (1986) and McFadden (1989). These modifications are not given in the statement of the consistency result in order to keep that result simple, but will be used later.

Some of the other conditions can also be weakened. Assumption (iii) can be changed to upper semi-continuity of \(Q_\Theta(\theta)\) and (iv) to \(\hat{Q}_n(\theta_0) \overset{p}{\to} Q_\Theta(\theta_0)\) and for all \(\varepsilon > 0\), \(\hat{Q}_n(\hat{\theta}) < Q_\Theta(\theta) + \varepsilon\) for all \(\theta \in \Theta\) with probability approaching one. Under these weaker conditions the conclusion still is satisfied, with exactly the same proof.

Theorem 2.1 is a weak consistency result, i.e. it shows \(\hat{\theta} \overset{p}{\to} \theta_0\). A corresponding strong consistency result, i.e. \(\hat{\theta} \overset{a.s.}{\to} \theta_0\), can be obtained by assuming that \(\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_\Theta(\theta)| \overset{a.s.}{\to} 0\) holds in place of uniform convergence in probability. The proof is exactly the same as that above, except that “a.s. for large enough \(n\)” replaces “with probability approaching one”. This and other results are stated here for convergence in probability because it suffices for the asymptotic distribution theory.

This result is quite general, applying to any topological space. Hence, it allows for \(\theta\) to be infinite-dimensional, i.e. for \(\theta\) to be a function, as would be of interest for nonparametric estimation of (say) a density or regression function. However, the compactness of the parameter space is difficult to check or implausible in many cases where \(\theta\) is infinite-dimensional.

To use this result to show consistency of a particular estimator it must be possible to check the conditions. For this purpose it is important to have primitive conditions, where the word “primitive” here is used synonymously with the phrase “easy to interpret”. The compactness condition is primitive but the others are not, so that it is important to discuss more primitive conditions, as will be done in the following subsections.

---

10 Upper semi-continuity means that for any \(\theta \in \Theta\) and \(\varepsilon > 0\) there is an open subset \(\mathcal{N}\) of \(\Theta\) containing \(\theta\) such that \(Q_\Theta(\theta') < Q_\Theta(\theta) + \varepsilon\) for all \(\theta' \in \mathcal{N}\).
Condition (i) is the identification condition discussed above, (ii) the boundedness condition on the parameter set, and (iii) and (iv) the continuity and uniform convergence conditions. These can be loosely grouped into “substantive” and “regularity” conditions. The identification condition (i) is substantive. There are well known examples where this condition fails, e.g. linear instrumental variables estimation with fewer instruments than parameters. Thus, it is particularly important to be able to specify primitive hypotheses for \( Q_0(\theta) \) to have a unique maximum. The compactness condition (ii) is also substantive, with \( \theta_0 \in \Theta \) requiring that bounds on the parameters be known. However, in applications the compactness restriction is often ignored. This practice is justified for estimators where compactness can be dropped without affecting consistency of estimators. Some of these estimators are discussed in Section 2.6.

Uniform convergence and continuity are the hypotheses that are often referred to as “the standard regularity conditions” for consistency. They will typically be satisfied when moments of certain functions exist and there is some continuity in \( \hat{\theta}_n(\theta) \) or in the distribution of the data. Moment existence assumptions are needed to use the law of large numbers to show convergence of \( \hat{\theta}_n(\theta) \) to its limit \( \theta_0(\theta) \). Continuity of the limit \( Q_0(\theta) \) is quite a weak condition. It can even be true when \( \hat{\theta}_n(\theta) \) is not continuous, because continuity of the distribution of the data can “smooth out” the discontinuities in the sample objective function. Primitive regularity conditions for uniform convergence and continuity are given in Section 2.3. Also, Section 2.7 relates uniform convergence to stochastic equicontinuity, a property that is necessary and sufficient for uniform convergence, and gives more sufficient conditions for uniform convergence.

To formulate primitive conditions for consistency of an extremum estimator, it is necessary to first find \( Q_0(\theta) \). Usually it is straightforward to calculate \( Q_0(\theta) \) as the probability limit of \( \hat{Q}_n(\theta) \) for any \( \theta \), a necessary condition for (iii) to be satisfied. This calculation can be accomplished by applying the law of large numbers, or hypotheses about convergence of certain components. For example, the law of large numbers implies that for MLE the limit of \( \hat{Q}_n(\theta) \) is \( Q_0(\theta) = E[\ln f(z | \theta)] \) and for NLS \( Q_0(\theta) = -E[\{y - h(x, \theta)\}^2] \). Note the role played here by the normalization of the log-likelihood and sum of squared residuals, that leads to the objective function converging to a nonzero limit. Similar calculations give the limit for GMM and CMD, as further discussed below. Once this limit has been found, the consistency will follow from the conditions of Theorem 2.1.

One device that may allow for consistency under weaker conditions is to treat \( \hat{\theta} \) as a maximum of \( \hat{Q}_n(\theta) - \hat{Q}_n(\theta_0) \) rather than just \( \hat{Q}_n(\theta) \). This is a magnitude normalization that sometimes makes it possible to weaken hypotheses on existence of moments. In the censored least absolute deviations example, where \( \hat{Q}_n(\theta) = -n^{-1} \sum_{t=1}^{n} |y_t - \max\{0, x'_t \theta\}| \), an assumption on existence of the expectation of \( y \) is useful for applying a law of large numbers to show convergence of \( \hat{Q}_n(\theta) \). In contrast \( \hat{Q}_n(\theta) - \hat{Q}_n(\theta_0) = -n^{-1} \sum_{t=1}^{n} [|y_t - \max\{0, x'_t \theta\}| - |y_t - \max\{0, x'_t \theta_0\}|] \) is a bounded function of \( y_t \), so that no such assumption is needed.
2.2. Identification

The identification condition for consistency of an extremum estimator is that the limit of the objective function has a unique maximum at the truth.\textsuperscript{11} This condition is related to identification in the usual sense, which is that the distribution of the data at the true parameter is different than that at any other possible parameter value. To be precise, identification is a necessary condition for the limiting objective function to have a unique maximum, but it is not in general sufficient.\textsuperscript{12} This section focuses on identification conditions for MLE, NLS, GMM, and CMD, in order to illustrate the kinds of results that are available.

2.2.1. The maximum likelihood estimator

An important feature of maximum likelihood is that identification is also sufficient for a unique maximum. Let $Y_1 \neq Y_2$ for random variables mean $\text{Prob}(\{Y_1 \neq Y_2\}) > 0$.

Lemma 2.2 (Information inequality)

If $\theta_0$ is identified [$\theta \neq \theta_0$ and $\theta \in \Theta$ implies $f(z|\theta) \neq f(z|\theta_0)$] and $\mathbb{E}[\ln f(z|\theta)] < \infty$ for all $\theta$ then $Q_0(\theta) = \mathbb{E}[\ln f(z|\theta)]$ has a unique maximum at $\theta_0$.

Proof

By the strict version of Jensen’s inequality, for any nonconstant, positive random variable $Y$, $-\ln(\mathbb{E}[Y]) < \mathbb{E}[\ -\ln(Y)]$.\textsuperscript{13} Then for $a = f(z|\theta)/f(z|\theta_0)$ and $\theta \neq \theta_0$, $Q_0(\theta_0) - Q_0(\theta) = \mathbb{E}[\ -\ln[f(z|\theta)/f(z|\theta_0)]] > \ -\ln(\mathbb{E}[f(z|\theta)/f(z|\theta_0)]) = \ -\ln(\mathbb{E}[f(z|\theta)dz]) = 0$. Q.E.D.

The term “information inequality” refers to an interpretation of $Q_0(\theta)$ as an information measure. This result means that MLE has the very nice feature that uniqueness of the maximum of the limiting objective function occurs under the very weakest possible condition of identification of $\theta_0$.

Conditions for identification in particular models are specific to those models. It

\textsuperscript{11}If the set of maximands $\mathcal{M}$ of the objective function has more than one element, then this set does not distinguish between the true parameter and other values. In this case further restrictions are needed for identification. These restrictions are sometimes referred to as normalizations. Alternatively, one could work with convergence in probability to a set $\mathcal{M}$, but imposing normalization restrictions is more practical, and is needed for asymptotic normality.

\textsuperscript{12}If $\theta_0$ is not identified, then there will be some $\tilde{\theta} \neq \theta_0$ such that the distribution of the data is the same when $\tilde{\theta}$ is the true parameter value as when $\theta_0$ is the true parameter value. Therefore, $Q_0(\tilde{\theta})$ will also be limiting objective function when $\tilde{\theta}$ is the true parameter, and hence the requirement that $Q_0(\tilde{\theta})$ be maximized at the true parameter implies that $Q_0(\tilde{\theta})$ has at least two maxima, $\theta_0$ and $\tilde{\theta}$.

\textsuperscript{13}The strict version of Jensen’s inequality states that if $a(y)$ is a strictly concave function [e.g. $a(y) = \ln(y)$] and $Y$ is a nonconstant random variable, then $a(\mathbb{E}[Y]) > \mathbb{E}[a(Y)]$. 

is often possible to specify them in a way that is easy to interpret (i.e. in a "primitive" way), as in the Cauchy example.

**Example 1.1 continued**

It will follow from Lemma 2.2 that $E[\ln f(z|\theta)]$ has a unique maximum at the true parameter. Existence of $E[\ln f(z|\theta)]$ for all $\theta$ follows from $|\ln f(z|\theta)| \leq C_1 + \ln(1 + \sigma^{-2}|z - \mu|^2) \leq C_1 + \ln(C_2 + C_3|z|^2)$ for positive constants $C_1$, $C_2$, and $C_3$, and existence of $E[\ln f(z|\theta)]$. Identification follows from $f(z|\theta)$ being one-to-one in the quadratic function $(1 + [(z - \mu)/\sigma]^2)$, the fact that quadratic functions intersect at no more than two points, and the fact that the probability of any two points is zero, so that $\text{Prob}(\{z:f(z|\theta) \neq f(z|\theta_0)\}) = 1 > 0$. Thus, by the information inequality, $E[\ln f(z|\theta)]$ has a unique maximum at $\theta_0$. This example illustrates that it can be quite easy to show that the expected log-likelihood has a unique maximum, even when the first-order conditions for the MLE do not have unique roots.

**Example 1.2 continued**

Throughout the probit example, the identification and regularity conditions will be combined in the assumption that the second-moment matrix $E[xx']$ exists and is nonsingular. This assumption implies identification. To see why, note that nonsingularity of $E[xx']$ implies that it is positive definite. Let $\theta \neq \theta_0$, so that $E[\{x'(\theta - \theta_0)\}^2] = (\theta - \theta_0)'E[xx'](\theta - \theta_0) > 0$, implying that $x'(\theta - \theta_0) \neq 0$, and hence $x'\theta \neq x'\theta_0$, where as before "not equals" means "not equal on a set of positive probability". Both $\Phi(v)$ and $\Phi(-v)$ are strictly monotonic, so that $x'\theta \neq x'\theta_0$ implies both $\Phi(x'\theta) \neq \Phi(x'\theta_0)$ and $1 - \Phi(x'\theta) \neq 1 - \Phi(x'\theta_0)$, and hence that $f(z|\theta) = \Phi(x'\theta)[1 - \Phi(x'\theta)] \neq f(z|\theta_0)$.

Existence of $E[xx']$ also implies that $E[|\ln f(z|\theta)|] < \infty$. It is well known that the derivative $d \ln \Phi(v)/dv = \lambda(v) = \phi(v)/\Phi(v)$ [for $\phi(v) = \nabla_v \Phi(v)$], is convex and asymptotes to $-v$ as $v \to -\infty$ and to zero as $v \to \infty$. Therefore, a mean-value expansion around $\theta = 0$ gives

$$|\ln \Phi(x'\tilde{\theta})| = |\ln \Phi(0) + \lambda(x'\tilde{\theta})x'| \leq |\ln \Phi(0)| + \lambda(x'\tilde{\theta})|x'|$$

$$\leq |\ln \Phi(0)| + C(1 + |x'|)|x'| \leq |\ln \Phi(0)| + C(1 + |x| || \theta || || x || || \theta ||)$$

Since $1 - \Phi(v) = \Phi(-v)$ and $y$ is bounded, $|\ln f(z|\theta)| \leq 2[|\ln \Phi(0)| + C(1 + |x| || \theta || || x || || \theta ||)]$, so existence of second moments of $x$ implies that $E[|\ln f(z|\theta)|]$ is finite. This part of the probit example illustrates the detailed work that may be needed to verify that moment existence assumptions like that of Lemma 2.2 are satisfied.

2.2.2. Nonlinear least squares

The identification condition for NLS is that the mean square error $E[\{y - h(x,\theta)\}^2] = -Q_o(\theta)$ have a unique minimum at $\theta_0$. As is easily shown, the mean square error
has a unique minimum at the conditional mean.\footnote{For \( m(x) = E[y|x] \) and \( a(x) \) any function with finite variance, iterated expectations gives \( E[(y - a(x))^2] = E[E[(y - m(x))^2] + 2E[(y - m(x))(m(x) - a(x))] + E[(m(x) - a(x))^2]] \geq E[(y - m(x))^2] \), with strict inequality if \( a(x) \neq m(x) \).} Since \( h(x, \theta_0) = E[y|x] \) is the conditional mean, the identification condition for NLS is that \( h(x, \theta) \neq h(x, \theta_0) \) if \( \theta \neq \theta_0 \), i.e. that \( h(x, \theta) \) is not the conditional mean when \( \theta \neq \theta_0 \). This is a natural "conditional mean" identification condition for NLS.

In some cases identification will not be sufficient for conditional mean identification. Intuitively, only parameters that affect the first conditional moment of \( y \) given \( x \) can be identified by NLS. For example, if \( \theta \) includes conditional variance parameters, or parameters of other higher-order moments, then these parameters may not be identified from the conditional mean.

As for identification, it is often easy to give primitive hypotheses for conditional mean identification. For example, in the linear model \( h(x, \theta) = x'\theta \) conditional mean identification holds if \( E[xx'] \) is nonsingular, for then \( \theta \neq \theta_0 \) implies \( x'\theta \neq x'\theta_0 \), as shown in the probit example. For another example, suppose \( x \) is a positive scalar and \( h(x, \theta) = \alpha + \beta x^2 \). As long as both \( \beta_0 \) and \( \gamma_0 \) are nonzero, the regression curve for a different value of \( \theta \) intersects the true curve at most at three \( x \) points. Thus, for identification it is sufficient that \( x \) have positive density over any interval, or that \( x \) have more than three points that have positive probability.

2.2.3. Generalized method of moments

For generalized method of moments the limit function \( Q_0(\theta) \) is a little more complicated than for MLE or NLS, but is still easy to find. By the law of large numbers, \( \hat{g}_n(\theta) \xrightarrow{p} g_0(\theta) = E[g(z, \theta)] \), so that if \( \hat{W} \xrightarrow{p} W \) for some positive semi-definite matrix \( W \), then by continuity of multiplication, \( \hat{Q}_n(\theta) \xrightarrow{p} Q_0(\theta) = -g_0(\theta)'Wg_0(\theta) \). This function has a maximum of zero at \( \theta_0 \), so \( \theta_0 \) will be identified if it is less than zero for \( \theta \neq \theta_0 \).

Lemma 2.3 (GMM identification)

If \( W \) is positive semi-definite and, for \( g_0(\theta) = E[g(z, \theta)] \), \( g_0(\theta_0) = 0 \) and \( Wg_0(\theta) \neq 0 \) for \( \theta \neq \theta_0 \) then \( Q_0(\theta) = -g_0(\theta)'Wg_0(\theta) \) has a unique maximum at \( \theta_0 \).

Proof

Let \( R \) be such that \( R'R = W \). If \( \theta \neq \theta_0 \), then \( 0 \neq Wg_0(\theta) = R'RG_0(\theta) \) implies \( Rg_0(\theta) \neq 0 \) and hence \( Q_0(\theta) = -[Rg_0(\theta)]'[Rg_0(\theta)] < Q_0(\theta_0) = 0 \) for \( \theta \neq \theta_0 \). Q.E.D.

The GMM identification condition is that if \( \theta \neq \theta_0 \) then \( g_0(\theta) \) is not in the null space of \( W \), which for nonsingular \( W \) reduces to \( g_0(\theta) \) being nonzero if \( \theta \neq \theta_0 \). A necessary order condition for GMM identification is that there be at least as many moment
functions as parameters. If there are fewer moments than parameters, then there will typically be many solutions to $g_0(\theta) = 0$.

If the moment functions are linear, say $g(z, \theta) = g(z) + G(z)\theta$, then the necessary and sufficient rank condition for GMM identification is that the rank of $WE[G(z)]$ is equal to the number of columns. For example, consider a linear instrumental variables estimator, where $g(z, \theta) = x' (y - Y' \theta)$ for a residual $y - Y' \theta$ and a vector of instrumental variables $x$. The two-stage least squares estimator of $\theta$ is a GMM estimator with $W = (\Sigma_{i=1}^n x_i x_i' / n)^{-1}$. Suppose that $E[xx']$ exists and is nonsingular, so that $W = (E[xx'])^{-1}$ by the law of large numbers. Then the rank condition for GMM identification is $E[xx']$ has full column rank, the well known instrumental variables identification condition. If $E[Y'|x] = x' \pi$ then this condition reduces to $\pi$ having full column rank, a version of the single equation identification condition [see F.M. Fisher (1976) Theorem 2.7.11]. More generally, $E[xx'] = E[E[xx']|x]$, so that GMM identification is the same as $x$ having “full rank covariance” with $E[Y'|x]$.

If $E[g(z, \theta)]$ is nonlinear in $\theta$, then specifying primitive conditions for identification becomes quite difficult. Here conditions for identification are like conditions for unique solutions of nonlinear equations (as in $E[g(z, \theta)] = 0$), which are known to be difficult. This difficulty is another reason to avoid formulating $\hat{\theta}$ as the solution to the first-order condition when analyzing consistency, e.g. to avoid interpreting MLE as a GMM estimator with $g(z, \theta) = V(\theta) \ln f(z|\theta)$. In some cases this difficulty is unavoidable, as for instrumental variables estimators of nonlinear simultaneous equations models.\(^{15}\)

Local identification analysis may be useful when it is difficult to find primitive conditions for (global) identification. If $g(z, \theta)$ is continuously differentiable and $\nabla_\theta E[g(z, \theta)] = E[\nabla_\theta g(z, \theta)]$, then by Rothenberg (1971), a sufficient condition for a unique solution of $WE[g(z, \theta)] = 0$ in a (small enough) neighborhood of $\theta_0$ is that $WE[\nabla_\theta g(z, \theta_0)]$ have full column rank. This condition is also necessary for local identification, and hence provides a necessary condition for global identification, when $E[\nabla_\theta g(z, \theta)]$ has constant rank in a neighborhood of $\theta_0$ [i.e. in Rothenberg’s (1971) “regular” case]. For example, for nonlinear 2SLS, where $\rho(z, \theta)$ is a residual and $g(z, \theta) = x' \cdot \rho(z, \theta)$, the rank condition for local identification is that $E[x' \nabla_\theta \rho(z, \theta_0)]$ has rank equal to its number of columns.

A practical “solution” to the problem of global GMM identification, that has often been adopted, is to simply assume identification. This practice is reasonable, given the difficulty of formulating primitive conditions, but it is important to check that it is not a vacuous assumption whenever possible, by showing identification in some special cases. In simple models it may be possible to show identification under particular forms for conditional distributions. The Hansen–Singleton model provides one example.

\(^{15}\)There are some useful results on identification of nonlinear simultaneous equations models in Brown (1983) and Roehrig (1989), although global identification analysis of instrumental variables estimators remains difficult.
Example 1.3 continued

Suppose that $\hat{W} = (n^{-1} \sum_{i=1}^n x_i x_i')$, so that the GMM estimator is nonlinear two-stage least squares. By the law of large numbers, if $E[xx']$ exists and is nonsingular, $\hat{W}$ will converge in probability to $W = (E[xx'])^{-1}$, which is nonsingular. Then the GMM identification condition is that there is a unique solution to $E[xp(z, \theta)] = 0$ at $\theta = \theta_0$, where $\rho(z, \theta) = \{\beta wy - 1\}$. Quite primitive conditions for identification can be formulated in a special log-linear case. Suppose that $w = \exp[a(x) + u]$ and $y = \exp[b(x) + u]$, where $(u, v)$ is independent of $x$, that $a(x) + b(x)$ is constant, and that $\rho(\theta_0) = 1$ for $\rho(\theta) = \exp[a(x) + \gamma_0 b(x)] \beta \exp(u + \gamma v)$. Suppose also that the first element is a constant, so that the other elements can be assumed to have mean zero (by "demeaning" if necessary, which is a nonsingular linear transformation, and so does not affect the identification analysis). Let $\alpha(x, y) = \exp[(\gamma - \gamma_0)b(x)]$. Then $E[\rho(z, \theta)|x] = \alpha(x, y)\rho(\theta) - 1$, which is zero for $\theta = \theta_0$, and hence $E[g(z, \theta)] = 0$. For $\theta \neq \theta_0$, $E[g(z, \theta)] = \{E[\alpha(x, y)]\rho(\theta) - 1, \text{Cov}[x', \alpha(x, y)]\rho(\theta)'\}$. This expression is nonzero if $\text{Cov}[x, \alpha(x, y)]$ is nonzero, because then the second term is nonzero if $\rho(\theta)$ is nonzero and the first term is nonzero if $\rho(\theta) = 0$. Furthermore, if $\text{Cov}[x, \alpha(x, y)] = 0$ for some $\gamma$, then all of the elements of $E[g(z, \theta)]$ are zero for all $\beta$, and one can choose $\beta > 0$ so the first element is zero. Thus, $\text{Cov}[x, \alpha(x, y)] \neq 0$ for $\gamma \neq \gamma_0$ is a necessary and sufficient condition for identification. In other words, the identification condition is that for all $\gamma$ in the parameter set, some coefficient of a nonconstant variable in the regression of $\alpha(x, y)$ on $x$ is nonzero. This is a relatively primitive condition, because we have some intuition about when regression coefficients are zero, although it does depend on the form of $b(x)$ and the distribution of $x$ in a complicated way. If $b(x)$ is a nonconstant, monotonic function of a linear combination of $x$, then this covariance will be nonzero. Thus, in this example it is found that the assumption of GMM identification is not vacuous, that there are some nice special cases where identification does hold.

2.2.4. Classical minimum distance

The analysis of CMD identification is very similar to that for GMM. If $\hat{\theta} \overset{P}{\rightarrow} \pi_0$ and $\hat{W} \overset{P}{\rightarrow} W$, $W$ positive semi-definite, then $\hat{Q}(\theta) = -[\hat{\theta} - h(\theta)]'W[\hat{\theta} - h(\theta)] \overset{P}{\rightarrow} [\pi_0 \cdot h(\theta)']W[\pi_0 - h(\theta)] = Q_0(\theta)$. The condition for $Q_0(\theta)$ to have a unique maximum (of zero) at $\theta_0$ is that $h(\theta_0) = \pi_0$ and $h(\theta) - h(\theta_0)$ is not in the null space of $W$ if $\theta \neq \theta_0$, which reduces to $h(\theta) \neq h(\theta_0)$ if $W$ is nonsingular. If $h(\theta)$ is linear in $\theta$ then there is a readily interpretable rank condition for identification, but otherwise the analysis of global identification is difficult. A rank condition for local identification is that the rank of $W \cdot \nabla_\theta h(\theta_0)$ equals the number of components of $\theta$.

16 It is well known that $\text{Cov}[x, f(x)] \neq 0$ for any monotonic, nonconstant function $f(x)$ of a random variable $x$. 
2.3 Uniform convergence and continuity

Once conditions for identification have been found and compactness of the parameter set has been assumed, the only other primitive conditions for consistency required by Theorem 2.1 are those for uniform convergence in probability and continuity of the limiting objective function. This subsection gives primitive hypotheses for these conditions that, when combined with identification, lead to primitive conditions for consistency of particular estimators.

For many estimators, results on uniform convergence of sample averages, known as uniform laws of large numbers, can be used to specify primitive regularity conditions. Examples include MLE, NLS, and GMM, each of which depends on sample averages. The following uniform law of large numbers is useful for these estimators.

Let \( a(z, \theta) \) be a matrix of functions of an observation \( z \) and the parameter \( \theta \), and for a matrix \( A = [a_{jk}] \), let \( \| A \| = (\sum_{j,k} a_{jk}^2)^{1/2} \) be the Euclidean norm.

**Lemma 2.4**

If the data are i.i.d., \( \Theta \) is compact, \( a(z, \theta) \) is continuous at each \( \theta \in \Theta \) with probability one, and there is \( d(z) \) with \( \| a(z, \theta) \| \leq d(z) \) for all \( \theta \in \Theta \) and \( E[d(z)] < \infty \), then \( E[a(z, \theta)] \) is continuous and \( \sup_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} a(z_i, \theta) - E[a(z, \theta)] \rightarrow P \rightarrow 0 \).

The conditions of this result are similar to assumptions of Wald's (1949) consistency proof, and it is implied by Lemma 1 of Tauchen (1985).

The conditions of this result are quite weak. In particular, they allow for \( a(z, \theta) \) to not be continuous on all of \( \Theta \) for given \( z \). Consequently, this result is useful even when the objective function is not continuous, as for Manski's (1975) maximum score estimator and the simulation-based estimators of Pakes (1986) and McFadden (1989). Also, this result can be extended to dependent data. The conclusion remains true if the i.i.d. hypothesis is changed to strict stationarity and ergodicity of \( z_i \).

The two conditions imposed on \( a(z, \theta) \) are a continuity condition and a moment existence condition. These conditions are very primitive. The continuity condition can often be verified by inspection. The moment existence hypothesis just requires a data-dependent upper bound on \( \| a(z, \theta) \| \) that has finite expectation. This condition is sometimes referred to as a "dominance condition", where \( d(z) \) is the dominating function. Because it only requires that certain moments exist, it is a "regularity condition" rather than a "substantive restriction".

It is often quite easy to see that the continuity condition is satisfied and to specify moment hypotheses for the dominance condition, as in the examples.

---

\(^{17}\) The conditions of Lemma 2.4 are not sufficient for measurability of the supremum in the conclusion, but are sufficient for convergence of the supremum in outer measure. Convergence in outer measure is sufficient for consistency of the estimator in terms of outer measure, a result that is useful when the objective function is not continuous, as previously noted.

\(^{18}\) Strict stationarity means that the distribution of \( (z_i, z_{i+1}, \ldots, z_{i+m}) \) does not depend on \( i \) for any \( m \), and ergodicity implies that \( n^{-1} \sum_{i=1}^{n} a(z_i) \rightarrow E[a(z)] \) for (measurable) functions \( a(z) \) with \( E[|a(z)|] < \infty \).
Example 1.1 continued

For the Cauchy location-scale likelihood, continuity of \( \ln f(z|\theta) = \ln C - \ln \sigma - \ln(1 + ((z - \mu)/\sigma)^2) \) is obvious. Also, as in the Example 1.1 discussion in Section 2.2.1, for any \( \Theta \) where \( \theta \) is bounded and \( \sigma \geq 0 \) is bounded away from zero, the dominance condition of Lemma 2.4 is satisfied for \( a(z, \theta) = \ln f(z|\theta) \) and \( d(z) = C_1 + \ln(C_2 + C_3|z|^2) \), for certain positive constants \( C_1, C_2, \) and \( C_3 \). Thus, by the conclusion of Lemma 2.4, \( E[\ln f(z|\theta)] \) is continuous and the average log-likelihood converges uniformly in probability to the expected log-likelihood.

Example 1.2 continued

For the probit example, continuity of \( \ln f(z|\theta) = y \ln \Phi(x'|\theta) + (1 - y) \ln \Phi(-x'|\theta) \) is obvious, while the dominance condition of Lemma 2.4 follows as in Section 2.2.1, with \( C(1 + \|x\|^2) = d(z) \). Then the conclusion of Lemma 2.4 applies to \( a(z, \theta) = \ln f(z|\theta) \).

Example 1.3 continued

In the Hansen–Singleton example, the GMM objective function depends on \( \theta \) through the average moment functions \( \hat{g}_i(\theta) = n^{-1} \sum_{i=1}^n g(z_i, \theta) = n^{-1} \sum_{i=1}^n x_i \times (\beta w_i y_i^\gamma - 1) \). Consequently, as shown below for general GMM estimators, uniform convergence of the objective function and continuity of the limit will hold if the hypotheses of Lemma 2.4 are satisfied with \( a(z, \theta) \) equal to each element of \( g(z, \theta) \). By inspection, each element of \( g(z, \theta) \) is continuous. Also, assuming \( \Theta \) is specified so that \( \beta \) and \( \gamma \) are bounded, and letting \( \beta_l, \beta_u, \) and \( \gamma_l, \gamma_u \) denote upper and lower bounds, respectively,

\[
\|g(z, \theta)\| \leq \|x\| \left[ 1 + (|\beta_l| + |\beta_u|)\|w\|(|y|^\gamma_u + |y|^\gamma_l) \right], \quad \theta \in \Theta.
\]

Thus, the dominance condition will be satisfied if each of \( \|x\|\|w\||y|^\gamma_u, \|x\|\|w\||y|^\gamma_l \) and \( \|x\| \) have finite expectations. In this example existence of \( E[\|x\|\|w\||y|^\gamma_u] \) and \( E[\|x\|\|w\||y|^\gamma_l] \) may place bounds on how large or small \( \gamma \) can be allowed to be.

Lemma 2.4 is useful, but it only applies to stationary data and to sample averages. There are many examples of models and estimators in econometrics where more general uniform convergence results are needed. It is possible to formulate necessary and sufficient conditions for uniform convergence using a stochastic equicontinuity condition. Stochastic equicontinuity is an important concept in recent developments in asymptotic theory, is used elsewhere in this chapter, and is fully discussed in Andrews’ chapter in this volume. However, because this concept is somewhat more technical, and not needed for many results, we have placed the discussion of uniform convergence and stochastic equicontinuity in Section 2.7, and left all description of its other uses until needed in Section 7.