1. Asymptotic Normality

1.1. Introduction. Results to be discussed:

- Asymptotic normality for M-estimator.
- Asymptotic normality for GMM and the optimal choice of weighting matrix.
- CUE, GEL and higher order comparison.

1.2. Main Theorem of Asymptotic Normality. We start with a simple lemma.

**Lemma 1.** Fix the probability space \((\Omega, \mathcal{F}, \mathbb{P})\). Consider a sequence of events \((A_n)_{n \in \mathbb{N}} \subseteq \mathcal{F}\). Then

(a) \(1_{A_n} - 1 = o_p(1)\) if and only if \(\mathbb{P}(A_n) \to 1\);
(b) \(1_{A_n} - 1 = o_p(1)\) implies that \(c_n (1_{A_n} - 1) = o_p(1)\) for any sequence \((c_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}\).
(c) Let \((X_n)_{n \in \mathbb{N}}\) be \(\mathbb{R}\)-valued random variables. Suppose that \(1_{A_n} - 1 = o_p(1)\). Then \(X_n = o_p(1)\) if and only if \(X_n 1_{A_n} = o_p(1)\).

**Proof.** First consider part (a). Note \(1_{A_n} - 1 = o_p(1)\) implies \(\mathbb{P}(A_n) \to 1\) because of the bounded convergence theorem. On the other hand, if \(\mathbb{P}(A_n) \to 1\), then for any \(\varepsilon > 0\),

\[
\mathbb{P}(|1_{A_n} - 1| > \varepsilon) \leq \mathbb{P}(1_{A_n} = 0) = \mathbb{P}(A_n^c) \to 0,
\]

implying that \(1_{A_n} - 1 = o_p(1)\).

Now, consider (b). Suppose \(1_{A_n} - 1 = o_p(1)\). By (a), we have \(\mathbb{P}(A_n^c) \to 0\). Note that for any \(\varepsilon > 0\),

\[
\mathbb{P}(|c_n (1_{A_n} - 1)| > \varepsilon) \leq \mathbb{P}(1_{A_n} = 0) = \mathbb{P}(A_n^c) \to 0.
\]

Hence, \(c_n (1_{A_n} - 1) = o_p(1)\).

Finally, we show (c). The “only if” part of the claim is obvious. We show the “if” part, so suppose \(X_n 1_{A_n} = o_p(1)\). Fix \(\varepsilon > 0\). Then

\[
\mathbb{P}(|X_n| > \varepsilon) = \mathbb{P}(|X_n| > \varepsilon, A_n) + \mathbb{P}(|X_n| > \varepsilon, A_n^c) \\
\leq \mathbb{P}(|X_n| 1_{A_n} > \varepsilon) + \mathbb{P}(A_n^c).
\]

Since \(X_n 1_{A_n} = o_p(1)\), \(\mathbb{P}(|X_n| 1_{A_n} > \varepsilon) \to 0\). Moreover, \(\mathbb{P}(A_n^c) \to 0\). Hence, \(\mathbb{P}(|X_n| > \varepsilon) \to 0\). This means that \(X_n = o_p(1)\). \(\square\)

**Remark 1.** Part (b) of the lemma shows that when \(1_{A_n}\) converges in probability to 1, the rate of convergence is “arbitrarily fast”. Part (c) shows that in order to show some variable \(X_n = o_p(1)\), it is enough to show this in the restriction to a sequence of events \(A_n\), whose probabilities approach 1 as \(n \to \infty\). If some statement
holds in restriction to such events, many authors say that the statement holds with probability approaching 1 (w.p.a.1).

**Lemma 2.** Let \((\Theta, d)\) be a metric space and \((f_n)_{n \in \mathbb{N}}\) be \(\mathbb{R}^m\)-valued random functions on \(\Theta\). Suppose that \(\hat{\theta}_n \xrightarrow{P} \theta_0\) and for some neighborhood \(\mathcal{N}\) containing \(\theta_0\), \(\sup_{\theta \in \mathcal{N}} \|f_n(\theta)\| = o_p(1)\). Then \(f_n(\hat{\theta}_n) = o_p(1)\).

**Proof.** Let \(A_n = \{\hat{\theta}_n \in \mathcal{N}\}\). Since \(\hat{\theta}_n \xrightarrow{P} \theta_0\), we know that \(\mathbb{P}(A_n) \to 1\).

Note that \(f(\hat{\theta}_n) 1_{A_n} \leq \sup_{\theta \in \mathcal{N}} \|f_n(\theta)\| = o_p(1)\). Hence, \(f(\hat{\theta}_n) 1_{A_n} = o_p(1)\). By lemma 1, we have \(f(\hat{\theta}_n) = o_p(1)\).

Here is the main theorem for the asymptotic normality of M-estimators in a finite-dimensional setting.

**Theorem 1.** (*Theorems 3.1 and 4.1 [NM]*) Let \((Q_n)_{n \in \mathbb{N}}\) be a sequence of random functions on \(\Theta \subseteq \mathbb{R}^m\) and \(\hat{\theta}_n = \arg \max_{\theta \in \Theta} Q_n(\theta)\). Suppose that

(a) \(\hat{\theta}_n \xrightarrow{P} \theta_0 \in \text{int}(\Theta)\)

(b) \(Q_n\) is twice continuously differentiable in a neighborhood \(\mathcal{N}\) of \(\theta_0\)

(c) \(c_n \nabla_{\theta} Q_n(\theta_0) \xrightarrow{d} N(0, \Sigma)\), for some real sequence \((c_n)_{n \in \mathbb{N}} \subset \mathbb{R}^+\) (typically \(c_n = n^{1/2}\))

(d) there is \(H(\theta)\) that is continuous at \(\theta_0\) and \(\sup_{\theta \in \mathcal{N}} \|\nabla_{\theta} Q_n(\theta) - H(\theta)\| \xrightarrow{P} 0\)

(e) \(H = H(\theta_0)\) is nonsingular.

Then

(i) \(c_n \left(\hat{\theta}_n - \theta_0\right) = H^{-1} c_n \nabla_{\theta} Q_n(\theta_0) + o_p(1)\); consequently, \(c_n \left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, H^{-1} \Sigma H^{-1})\).

(ii) Let \(\hat{H}_n = \nabla_{\theta} Q_n(\hat{\theta}_n)\). We have \(\hat{H}_n \xrightarrow{P} H\) and \(\hat{H}_n^{-1} \xrightarrow{P} H^{-1}\). If we can find estimators \(\Sigma_n\) with \(\Sigma_n \xrightarrow{P} \Sigma\), then the asymptotic variance can be estimated by \(\hat{H}_n^{-1} \Sigma_n \hat{H}_n^{-1}\).

**Proof.** Step 1 (Linearization). We define a sequence of events \(A_n = \{\hat{\theta}_n \in \mathcal{N}\}\).

On \(A_n\), we have

\[\nabla_{\theta} Q_n(\hat{\theta}_n) = 0,\]

because \(Q_n\) is differentiable on \(\mathcal{N}\) and \(\hat{\theta}_n \in \mathcal{N}\) is a maximizer of \(Q\). By assumption, \(\nabla_{\theta} Q_n(\cdot)\) is also differentiable. By the mean value theorem, there exists \(\hat{\theta}_n\) falling between \(\theta_0\) and \(\hat{\theta}_n\) such that

\[0 = \nabla_{\theta} Q_n(\hat{\theta}_n) = \nabla_{\theta} Q_n(\theta_0) + (\hat{\theta}_n - \theta_0)^T \nabla_{\theta} Q_n(\hat{\theta}_n)\] on \(A_n\).

Now, let \(B_n = \{\nabla_{\theta} Q_n(\hat{\theta}_n)\) is nonsingular\}. Hence, on the event \(B_n\), \(\nabla_{\theta} Q_n(\hat{\theta}_n)\) is invertible. We can then solve the above equation on \(A_n \cap B_n\) and get

\[\hat{\theta}_n - \theta_0 = - \left(\nabla_{\theta} Q_n(\hat{\theta}_n)\right)^{-1} \nabla_{\theta} Q_n(\theta_0)\] on \(A_n \cap B_n\).
This equation also holds when we multiply both side by \( c_n \). We hence have

\[
(0.1) \quad c_n (\hat{\theta}_n - \theta_0) 1_{A_n \cap B_n} = - (\nabla_{\theta} Q_n (\hat{\theta}_n))^{-1} c_n \nabla_{\theta} Q_n (\theta_0) 1_{A_n \cap B_n}.
\]

**Step 2 (Handling bad situations).** We now consider the behavior of \( A_n \) and \( B_n \). Since \( \hat{\theta}_n \overset{p}{\to} \theta_0 \), we have \( \mathbb{P} (A_n) \to 0 \). By condition (d), we have

\[
\| \nabla_{\theta} Q_n (\hat{\theta}_n) - H (\hat{\theta}_n) \|_{A_n} \leq \sup_{\theta \in \mathcal{N}} \| \nabla_{\theta} Q_n (\theta) - H (\theta) \| = o_p (1),
\]

where the inequality holds because on \( A_n \), \( \hat{\theta}_n \in \mathcal{N} \). Note that \( \hat{\theta}_n \overset{p}{\to} \theta_0 \) also implies \( \hat{\theta}_n \overset{p}{\to} \theta_0 \). Since \( H \) is continuous at \( \theta_0 \), we use the continuous mapping theorem to get \( H (\hat{\theta}_n) \overset{p}{\to} H (\theta_0) \). Since \( 1_{A_n} \overset{p}{\to} 1 \), we have \( H (\hat{\theta}_n) 1_{A_n} \overset{p}{\to} H (\theta_0) \). Combining this convergence with the display above, we have

\[
(0.2) \quad \nabla_{\theta} Q_n (\hat{\theta}_n) 1_{A_n} \overset{p}{\to} H (\theta_0).
\]

Denote \( \eta = |\det [H (\theta_0)]| \). Since \( H (\theta_0) \) is non-singular (condition (e)), we have \( \eta > 0 \). By the continuous mapping theorem,

\[
|\det [\nabla_{\theta} Q_n (\hat{\theta}_n)]| 1_{A_n} \overset{p}{\to} |\det [H (\theta_0)]| = \eta.
\]

By the definition of \( \overset{p}{\to} \), we have

\[
\mathbb{P} \left( |\det [\nabla_{\theta} Q_n (\hat{\theta}_n)]| 1_{A_n} < \eta/2 \right) \leq \mathbb{P} \left( |\det [\nabla_{\theta} Q_n (\hat{\theta}_n)]| 1_{A_n} - \eta > \eta/2 \right) \to 0.
\]

Note that \( |\det [\nabla_{\theta} Q_n (\hat{\theta}_n)]| 1_{A_n} \geq \eta/2 \) necessarily implies that \( \nabla_{\theta} Q_n (\hat{\theta}_n) \) is nonsingular. So we have

\[
\mathbb{P} (A_n \cap B_n) = \mathbb{P} (A_n, \nabla_{\theta} Q_n (\hat{\theta}_n) \text{ is nonsingular}) \\
\geq \mathbb{P} (A_n, |\det [\nabla_{\theta} Q_n (\hat{\theta}_n)]| 1_{A_n} \geq \eta/2) \\
= \mathbb{P} (|\det [\nabla_{\theta} Q_n (\hat{\theta}_n)]| 1_{A_n} \geq \eta/2) \\
\to 1.
\]

In other words, \( 1_{A_n \cap B_n} - 1 = o_p (1) \). By part (b) of the previous lemma, we see that

\[
(0.4) \quad c_n (1_{A_n \cap B_n} - 1) = o_p (1).
\]

**Step 3 (Linear Representation).** Note that (0.2) and \( 1_{A_n \cap B_n} - 1 = o_p (1) \) implies that \( \nabla_{\theta} Q_n (\hat{\theta}_n) 1_{A_n \cap B_n} \overset{p}{\to} H (\theta_0) \). Then by the nonsingularity of \( H (\theta_0) \) and the continuous mapping theorem, we have

\[
(0.5) \quad (\nabla_{\theta} Q_n (\hat{\theta}_n))^{-1} 1_{A_n \cap B_n} \overset{p}{\to} H (\theta_0)^{-1} = H^{-1}.
\]
Note that \( c_n(\hat{\theta}_n - \theta_0) = c_n(\hat{\theta}_n - \theta_0)1_{A_n \cap B_n} + (\hat{\theta}_n - \theta_0)c_n(1_{A_n \cap B_n} - 1). \)

But \( \hat{\theta}_n - \theta_0 = o_p(1) \) and \( c_n(1_{A_n \cap B_n} - 1) = o_p(1). \) Hence,

\[
(0.6) \quad c_n(\hat{\theta}_n - \theta_0) = c_n(\hat{\theta}_n - \theta_0)1_{A_n \cap B_n} + o_p(1).
\]

On the other hand, by (0.5),

\[
(\nabla_{\theta}Q_n(\hat{\theta}_n))^{-1}c_n\nabla_{\theta}Q_n(\theta_0)1_{A_n \cap B_n} = (H^{-1} + o_p(1))c_n\nabla_{\theta}Q_n(\theta_0) = H^{-1}c_n\nabla_{\theta}Q_n(\theta_0) + o_p(1)\frac{c_n\nabla_{\theta}Q_n(\theta_0)}{o_p(1)}
\]

\[
(0.7) \quad = H^{-1}c_n\nabla_{\theta}Q_n(\theta_0) + o_p(1)
\]

Combining (0.1), (0.7) and (0.6), we derive

\[
c_n(\hat{\theta}_n - \theta_0) = H^{-1}c_n\nabla_{\theta}Q_n(\theta_0) + o_p(1).
\]

This finishes the proof of the first claim in (i). The second claim then follows from Slutsky’s theorem and condition (c).

**Step 4 (Asymptotic variance estimation).** By condition (d), we use Lemma 2 to derive \( \nabla_{\theta}Q_n(\hat{\theta}_n) - H(\hat{\theta}_n) = o_p(1). \) But \( H(\hat{\theta}_n) - H = o_p(1) \) by continuous mapping. Hence, \( \nabla_{\theta}Q_n(\hat{\theta}_n) - H = o_p(1). \) Other claims follows easily from continuous mapping.

**Remark 2.** (Important) Note that we prove the theorem without saying what is \( \theta_0 \), except requiring it to be the limit of \( \hat{\theta}_n \). In particular, we do not require \( \theta_0 \) to be the “true parameter” of our model, simply because we have not even specified a model. In view of the consistency result of M-estimator (see lecture note 2), \( \theta_0 \) is defined as the well-separated maximum of the limit criterion function \( Q(\theta) \).

If we have a model, such as a likelihood model or a moment equality model, the true parameter often is the well-separated maximum of \( Q(\theta) \). In this case, \( \theta_0 \) in this theorem coincide with the true parameter in our model. But what if our model is misspecified? Under misspecification, the notion of “true parameter” is an empty concept. Nevertheless, \( Q_n \) is well-defined, \( Q \) is well-defined, and \( \theta_0 \) is still well-defined as the well-separated maximum of \( Q \). In this case, we call \( \theta_0 \) “the pseudo-true parameter”, which is defined through our estimation procedure. The theorem goes through and gives a characterization of the sampling variability of the estimator \( \hat{\theta}_n \) of the pseudo-true parameter \( \theta_0 \).

**Problem 1.** (Asymptotic Normality of MLE) Read Theorem 3.3 of [NM]. Discuss the role of each assumption.
Problem 2. (MLE under misspecification) Again, consider Theorem 3.3 of \[NM\]. Do not assume that \(\text{“the hypotheses of \([Theorem\] 2.5 are satisfied”}\). Instead, assume that \(^n\rightarrow \theta_0\). This is the situation in which we conduct MLE with density function \(f(z|\theta)\) without assuming this is the density of the data. As a result, \(\theta_0\) is not the “true parameter” but only the pseudo-true parameter. Keep other assumptions in the theorem. Derive the asymptotic distribution of \(n^{1/2} (\hat{\theta}_n - \theta_0)\).

This theorem establishes asymptotic normality of M-estimators for twice differentiable \(Q_n(\cdot)\) functions. The differentiability is crucial to our proof, since the linearization step relies on it. Such smoothness requirement can be weakened with more techniques from the empirical process theory. Maximum likelihood models and GMM typically satisfy the differentiability condition. Nonsmoothness shows up quite often in semiparametric models.

The asymptotic variance has the form \(H^{-1}\Sigma H^{-1}\). The regularity conditions for asymptotic normality also justify using \(\hat{H}_n = \nabla_{\theta_0} Q_n(\hat{\theta}_n)\) as an estimator for \(H\); so we have solved part of the problem of estimating the asymptotic variance for free. On the other hand, the estimation of \(\Sigma\) is model specific; partly because the variance \(\Sigma\) itself is model specific. We discuss how characterize and estimate \(\Sigma\) in the following examples.

Example 1. Suppose that \(Q_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} q(X_t, \theta)\), and \((X_t)_{t\in\mathbb{N}}\) are IID. Then \(Q(\theta) = \text{plim}_{n\to\infty} Q_n(\theta) = \mathbb{E}[q(X_t, \theta)]\). We also have \(\nabla_{\theta_0} Q_n(\theta_0) = \frac{1}{n} \sum_{t=1}^{n} \nabla_{\theta} q(X_t, \theta_0)\).

Under regularity conditions (see e.g. Lemma 3.6 of \[NM\]), we can interchange the order of differentiation and expectation, and have
\[
\mathbb{E}[\nabla_{\theta} q(X_t, \theta_0)] = \nabla_{\theta} \mathbb{E}[q(X_t, \theta)] = \nabla_{\theta} Q(\theta_0).
\]

Since \(\theta_0\) is an interior maximizer of \(Q(\cdot)\), we must have \(\nabla_{\theta} Q(\theta_0) = 0\). Therefore, \((\nabla_{\theta} q(X_t, \theta_0))_{t\in\mathbb{N}}\) is an IID sequence of random variables with zero mean. If we also assume that \(\mathbb{E} \left[ \| \nabla_{\theta} q(X_t, \theta) \|^2 \right] < \infty\), then by CLT,
\[
\frac{1}{n^{1/2}} \nabla_{\theta} Q_n(\theta_0) \xrightarrow{d} N(0, \Sigma(\theta_0))
\]
where \(\Sigma(\theta) = \mathbb{E} \left[ \nabla_{\theta} q(X_t, \theta) \nabla_{\theta} q(X_t, \theta)^\top \right]\).

A natural candidate for estimating \(\Sigma(\theta_0)\) is \(\hat{\Sigma}_n(\hat{\theta}_n)\) where
\[
\hat{\Sigma}_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \nabla_{\theta} q(X_t, \theta) \nabla_{\theta} q(X_t, \theta)^\top.
\]

We want to show that \(\hat{\Sigma}_n(\hat{\theta}_n) - \Sigma(\theta_0) = o_p(1)\). We have at least two (very similar) ways of thinking. Firstly, Lemma 4.3 of \[NM\] can be used to show this result. For the reader’s convenience, we reproduce their lemma here but under our
notation: “If \((X_t)_{t \in \mathbb{N}}\) is IID, a \((X_t, \theta)\) is continuous at \(\theta_0\) with probability one and there exists a neighborhood \(N\) of \(\theta_0\) such that \(\mathbb{E} [\sup_{\theta \in N} \|a(X_t, \theta)\|] < \infty\), then for any \(\hat{\theta}_n \xrightarrow{p} \theta_0\), \(n^{-1} \sum_{t=1}^{n} a(X_t, \hat{\theta}_n) \xrightarrow{p} \mathbb{E} [a(X_t, \theta_0)]\).” To apply this result for our purpose here, take \(a(X_t, \theta) = \nabla \theta q(X_t, \theta) \nabla \theta q(X_t, \theta)^\top\).

Alternatively, note that
\[
\tilde{\Sigma}_n \left( \hat{\theta}_n \right) - \Sigma (\theta_0) = \underbrace{\tilde{\Sigma}_n \left( \hat{\theta}_n \right) - \hat{\Sigma}_n \left( \hat{\theta}_n \right)}_{\text{use lemma 2 with } f_n(\cdot) = \Sigma_n(\cdot) - \Sigma(\cdot)} + \underbrace{\hat{\Sigma}_n \left( \hat{\theta}_n \right) - \Sigma (\theta_0)}_{\text{use continuous mapping}}
\]
we can use the hint below each term on the right-hand-side of the above display to show that each of them is \(o_p(1)\). To do so, we need to verify (1) \(\sup_{\theta \in N} |\Sigma_n (\theta) - \Sigma (\theta)| = o_p(1)\) for some neighborhood around \(\theta_0\) (2) \(\Sigma (\cdot)\) is continuous at \(\theta\). These conditions are in the same spirit as condition (d) of Theorem 1. In the IID case, by Lemma 2.4 of [NM] (Tauchen 1985), it suffices to assume that \(q(X_t, \theta)\) is continuously differentiable with probability 1 in some neighborhood \(N\) of \(\theta_0\) and \(\mathbb{E} \left[ \sup_{\theta \in N} \|\nabla q(X_t, \theta)\|^2 \right] < \infty\). (Note that Lemma 2.4 of [NM] requires the index set to be compact. But we can apply the argument to the closure of \(N\).)

Example 2. Suppose that \(Q_n (\theta) = \frac{1}{n} \sum_{t=1}^{n} q(X_t, \theta)\), but now, suppose that \((X_t)_{t \in \mathbb{N}}\) is i.i.d. For correctly specified models, it is still reasonable to assume that \(\theta_0\) maximizes \(\mathbb{E} [q(X_t, \theta)]\) for each \(t\), implying that \(\nabla q(X_t, \theta_0) = 0\). Therefore, \((q(X_t, \theta))_{t \in \mathbb{N}}\) is a zero-mean i.i.d sequence. By Lindeberg’s CLT (lecture note 1)
\[
n^{1/2} \nabla \theta Q_n (\theta_0) = \frac{1}{n^{1/2}} \sum_{t=1}^{n} \nabla \theta q(X_t, \theta_0) \xrightarrow{d} N(0, \Sigma (\theta_0)),
\]
where
\[
\Sigma (\theta) = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E} \left[ \nabla \theta q(X_t, \theta) \nabla \theta q(X_t, \theta)^\top \right].
\]
Again, we propose estimating \(\Sigma (\theta_0)\) with \(\hat{\Sigma}_n \left( \hat{\theta}_n \right)\). Similar as in the previous example, it is sufficient to have (1) \(\sup_{\theta \in N} |\Sigma_n (\theta) - \Sigma (\theta)| = o_p(1)\) for some neighborhood \(N\) around \(\theta_0\) (2) \(\Sigma (\cdot)\) is continuous at \(\theta\). We now discuss these two conditions. For (1), by using LLN of i.i.d sequence, it is easy to show that \(\tilde{\Sigma}_n (\theta) - \Sigma (\theta) = o_p(1)\) for each \(\theta\). We need to establish the stochastic equicontinuity for \((\tilde{\Sigma}_n (\theta) - \Sigma (\theta) : \theta \in N)\). This requires tools for i.i.d sequences. For (2), we note that it is easy to get conditions such that \(\mathbb{E} \left[ \nabla \theta q(X_t, \theta) \nabla \theta q(X_t, \theta)^\top \right]\) is continuous in \(\theta\) for each \(t\) (See the problem below). Hence, for each \(n \in \mathbb{N},\)

\Nu {10}{12}
\[
\frac{1}{n} \sum_{t=1}^{n} \mathbb{E} \left[ \nabla_{\theta} q(X_t, \theta) \nabla_{\theta} g(X_t, \theta)^{\top} \right]
\] is continuous in \( \theta \). But the limit of continuous functions may not be continuous; hence the continuity of \( \Sigma(\theta) \) requires a little bit more. The bottom line is that when the data is not identically distributed but still independent, we can still use the same estimators as in the IID case, after being careful about our regularity conditions.

**Problem 3.** Let \( f : \Omega \times \Theta \rightarrow \mathbb{R} \) be a random function and \( \theta_0 \in \Theta \). Suppose that (a) the mapping \( \theta \mapsto f(\theta) \) is continuous at \( \theta_0 \) with probability 1; (b) there exists some random variable \( Y \) and a neighborhood (or open ball) \( N \) of \( \theta_0 \) such that \( \sup_{\theta \in N} |f(\theta)| \leq Y \) and \( \mathbb{E}[Y] < \infty \). Show the following:

(i) Interpret condition (a) as follows: for any \( \theta_n \rightarrow \theta_0 \), \( f(\theta_n) \xrightarrow{a.s.} f(\theta_0) \).

(ii) Show that \( \mathbb{E}[f(\theta_n)] \rightarrow \mathbb{E}[f(\theta_0)] \) by using the dominated convergence theorem.

(iii) Conclude that the mapping \( \theta \mapsto \mathbb{E}[f(\theta)] \) is continuous at \( \theta_0 \).

**Example 3.** Suppose that \( Q_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} q(X_t, \theta) \), but now, suppose that \( (X_t)_{t \in \mathbb{N}} \) is a dependent sequence. We maintain the assumption that \( \mathbb{E}[\nabla_{\theta} q(X_t, \theta_0)] = 0 \). Under regularity conditions, which we will discuss more in the lecture on limit theorems for dependent data, we have

\[
n^{1/2} \sum_{t=1}^{n} \nabla_{\theta} q(X_t, \theta_0) \xrightarrow{d} N(0, \Sigma(\theta_0)),
\]

where

\[
\Sigma(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left[ \left( \sum_{t=1}^{n} \nabla_{\theta} q(X_t, \theta) \right) \left( \sum_{t=1}^{n} \nabla_{\theta} q(X_t, \theta) \right)^{\top} \right].
\]

The asymptotic variance \( \Sigma(\theta_0) \) is the long run variance of the sequence \( (\nabla_{\theta} q(X_t, \theta_0))_{t \in \mathbb{N}} \), we will discuss the estimation of this quantity later.

**1.3. GMM under correct specification.** We now consider the asymptotic result for GMM. In this case,

\[
Q_n(\theta) = -\bar{g}_n(\theta)^{\top} W_n \bar{g}_n(\theta),
\]

where \( \bar{g}_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} q(X_t, \theta) \), \( (X_t)_{1 \leq t \leq n} \) is data and \( \theta \) is the parameter.

**Theorem 2.** (Theorem 3.2 [NM]) Suppose that \( \hat{\theta}_n = \arg \max_{\theta \in \Theta} Q_n(\theta) \) and \( W_n \xrightarrow{p} W \), \( W \) is positive semi-definite, \( \hat{\theta}_n \xrightarrow{p} \theta_0 \), and

(a) \( \theta_0 \in \text{int}(\Theta) \)

(b) \( \bar{g}_n(\theta) \) is continuously differentiable in a neighborhood \( N \) of \( \theta_0 \)

(c) \( n^{1/2} \bar{g}_n(\theta_0) \xrightarrow{d} N(0, S) \)

(d) there is \( G(\theta) \) that is continuous at \( \theta_0 \) and \( \sup_{\theta \in N} \| \nabla_{\theta} \bar{g}_n(\theta) - G(\theta) \| \xrightarrow{p} 0 \)

(e) for \( G = G(\theta_0) \), and \( G'WG \) is nonsingular.
Then $n^{1/2} \left( \hat{\theta}_n - \theta_0 \right) = - (G^T W G)^{-1} G^T W n^{1/2} \bar{g}_n(\theta_0) + o_p(1)$ and consequently, $n^{1/2} \left( \hat{\theta}_n - \theta_0 \right) \xrightarrow{d} N \left( 0, (G^T W G)^{-1} G^T W S W G \left( G^T W G \right)^{-1} \right)$.

**Proof.** (Sketchy) We only describe the main steps. By the first order condition and the mean value theorem, w.p.a.1, with some $\bar{\theta}$ between $\hat{\theta}_n$ and $\theta_0$,

$$0 = \nabla_{\theta} \bar{g}_n \left( \bar{\theta} \right)^T W_n \bar{g}_n \left( \bar{\theta} \right)$$

$$= \nabla_{\theta} \bar{g}_n \left( \hat{\theta}_n \right)^T W_n \left( \bar{g}_n(\theta_0) + \nabla_{\theta} \bar{g}_n \left( \hat{\theta}_n \right) (\bar{\theta} - \theta_0) \right).$$

By condition (d), $\nabla_{\theta} \bar{g}_n \left( \hat{\theta}_n \right)^T W_n \nabla_{\theta} \bar{g}_n \left( \hat{\theta}_n \right) = G^T W G + o_p(1)$. By condition (e), we see that w.p.a.1 $\nabla_{\theta} \bar{g}_n \left( \hat{\theta}_n \right)^T W_n \nabla_{\theta} \bar{g}_n \left( \hat{\theta}_n \right)$ is invertible. Hence, we can solve the equation above

$$n^{1/2} \left( \hat{\theta}_n - \theta_0 \right) = - \left( \nabla_{\theta} \bar{g}_n \left( \hat{\theta}_n \right)^T W_n \nabla_{\theta} \bar{g}_n \left( \hat{\theta}_n \right) \right)^{-1} \nabla_{\theta} \bar{g}_n \left( \hat{\theta}_n \right)^T W_n n^{1/2} \bar{g}_n(\theta_0)$$

$$= - \left( (G^T W G)^{-1} + o_p(1) \right) \left( G^T W + o_p(1) \right) n^{1/2} \bar{g}_n(\theta_0)$$

$$= - (G^T W G)^{-1} G^T W n^{1/2} \bar{g}_n(\theta_0) + o_p(1).$$

This gives us the asymptotic linear representation of our estimator. By Slutsky’s theorem and condition (c), we get the asymptotic distribution of $n^{1/2} \left( \hat{\theta}_n - \theta_0 \right)$ as stated in the theorem. \qed

**Remark 3.** The assumptions for Theorem 2 is slightly weaker than Theorem 1 in that the former only involves the first-order derivative $\nabla_{\theta} \bar{g}_n(\theta)$ while the latter involves the second-order derivative $\nabla_{\theta}^2 Q_n(\theta)$. The reason is that the GMM calculation exploits the special quadratic form of its sample criterion function.

**Problem 4.** Apply Theorem 1 to show that it gives the same asymptotic linear representation and asymptotic distribution of $n^{1/2} \left( \hat{\theta}_n - \theta_0 \right)$ as in Theorem 2. (Do not bother verifying regularity conditions.)

**Theorem 3.** (Efficient choice of weighting matrix) Consider the same setting as Theorem 2. Suppose that $S$ is nonsingular. Let $\theta^*_n$ be GMM estimators with weighting matrix $W_n^* \xrightarrow{p} S^{-1}$. Then $\text{Avar}(n^{1/2}(\hat{\theta}_n - \theta_0)) - \text{Avar}(n^{1/2}(\theta^*_n - \theta_0))$ is a positive semi-definite matrix.

**Proof.** Recall from Theorem 2 the asymptotic linear representations: with $B(W) = - (G^T W G)^{-1} G^T W$

$$n^{1/2} \left( \hat{\theta}_n - \theta_0 \right) = B(W) n^{1/2} \bar{g}_n(\theta_0) + o_p(1)$$

$$n^{1/2} \left( \theta^*_n - \theta_0 \right) = B(S^{-1}) n^{1/2} \bar{g}_n(\theta_0) + o_p(1).$$
But the efficiency claim is then immediate:

\[ n^{1/2} \left( \hat{\theta}_n - \theta^*_n \right) = (B(W) - B(S^{-1})) n^{1/2} \bar{g}_n(\theta_0) + o_p(1). \]

The asymptotic covariance between \( n^{1/2} (\theta^*_n - \theta_0) \) and \( n^{1/2} (\hat{\theta}_n - \theta^*_n) \) is

\[
Avar \left( n^{1/2} \left( \hat{\theta}_n - \theta^*_n \right) , n^{1/2} \left( \theta^*_n - \theta_0 \right) \right) 
= (B(W) - B(S^{-1})) \times Avar \left( n^{1/2} \bar{g}_n(\theta_0) \right) \times B(S^{-1})^\top 
= (B(W) - B(S^{-1})) \times S \times B(S^{-1})^\top.
\]

But

\[
B(W) \times S \times B(S^{-1})^\top = \left( G^\top W G \right)^{-1} G^\top W S S^{-1} G (G^\top S^{-1} G)^{-1} 
\]

\[
= \left( G^\top S^{-1} G \right)^{-1}. 
\]

Note that the right-hand-side of the above display no longer depends on \( W \). In particular, \( B(W) \times S \times B(S^{-1})^\top = B(S^{-1}) \times S \times B(S^{-1})^\top \). Hence,

\[
Avar \left( n^{1/2} \left( \hat{\theta}_n - \theta^*_n \right) , n^{1/2} \left( \theta^*_n - \theta_0 \right) \right) = 0;
\]

in words, we have shown that \( n^{1/2} (\hat{\theta}_n - \theta^*_n) \) and \( n^{1/2} (\theta^*_n - \theta_0) \) are asymptotically independent (because they are asymptotically normal with zero covariance). The efficiency claim is then immediate:

\[
Avar \left( n^{1/2} \left( \hat{\theta}_n - \theta_0 \right) \right) = Avar \left( n^{1/2} (\theta^*_n - \theta_0) + n^{1/2} (\hat{\theta}_n - \theta^*_n) \right) 
= Avar \left( n^{1/2} (\theta^*_n - \theta_0) \right) + Avar \left( n^{1/2} (\hat{\theta}_n - \theta^*_n) \right) 
\geq \text{matrix sense } Avar \left( n^{1/2} (\theta^*_n - \theta_0) \right) .
\]

To implement the efficient choice of the weighting matrix, we only need to have a consistent estimator \( \hat{S}_n \) for \( S \); we then set \( W_n^* = \hat{S}_n^{-1} \). If \( (X_t)_{1 \leq t \leq n} \) are independent, we define

\[
\hat{S}_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} g(X_t, \theta) g(X_t, \theta)^\top ,
\]

set \( \hat{S}_n = \hat{S}_n(\hat{\theta}_{1,n}) \), where \( \hat{\theta}_{1,n} \) is a consistent estimator of \( \theta_0 \), typically obtained from a GMM estimation with weighting matrix being the identity matrix. Under mild conditions as discussed before, \( \hat{S}_n(\hat{\theta}_{1,n}) \xrightarrow{p} S \). We then use \( \hat{S}_n(\hat{\theta}_{1,n})^{-1} \) as the weighting matrix to get the efficient GMM estimator \( \hat{\theta}_{2,n} = \arg \min_\theta \bar{g}_n(\theta)^\top \hat{S}_n(\hat{\theta}_{1,n})^{-1} \bar{g}_n(\theta)\).
In theory, \( \hat{\theta}_{2n} \) is a more precise estimator than \( \hat{\theta}_{1n} \), and it makes sense to iterate this procedure:

\[
\hat{\theta}_{k+1,n} = \arg \min_\theta \tilde{g}_n (\theta)^\top \hat{S}_n \left( \hat{\theta}_{k,n} \right)^{-1} \tilde{g}_n (\theta).
\]

The estimator \( \hat{\theta}_{2,n} \) is called the “two-step efficient GMM estimator”, and \( \hat{\theta}_{k,n} \) is called an “iterated” GMM estimator. In the dependent case, \( \hat{S}_n (\theta) \) should be defined as a long-run variance estimator of the sequence \( g(X_t, \theta) \).

### 1.4. GMM, CUE, GEL and higher order properties.

A closely related, but different estimator is Hansen, Heaton and Yaron’s (JBES, 1996) continuous updating estimator (CUE), defined as follows

\[
\hat{\theta}_{n}^{CUE} = \arg \min_\theta \tilde{g}_n (\theta)^\top \hat{S}_n (\theta)^{-1} \tilde{g}_n (\theta),
\]

\[
\hat{S}_n (\theta) = \frac{1}{n} \sum_{t=1}^n g(X_t, \theta) g(X_t, \theta)^\top \quad \text{(consider IID case here)}
\]

The key difference between CUE and iterated GMM is that in CUE, the minimization over the parameter \( \theta \) runs through the weighting function \( \hat{S}_n (\theta)^{-1} \), but in iterated GMM, it is fixed before the minimization. The CUE is still an M-estimator with \( Q_n (\theta) = \tilde{g}_n (\theta)^\top \hat{S}_n (\theta)^{-1} \tilde{g}_n (\theta) \). It is an easy exercise to see that \( \hat{\theta}_n \) is consistent, asymptotically normal, and its asymptotic variance is the same as the efficient 2-step GMM and iterated GMM estimators.

The CUE actually belongs to a more general class of estimators, which are labelled as generalized empirical likelihood (GEL) estimators. To define GEL estimators, let \( \rho (\cdot) \) be a concave function such as

\[
\rho (v) = \rho_0 - v - \frac{1}{2} v^2 \quad \text{(quadratic)}
\]

\[
\rho (v) = \ln (1 - v)
\]

\[
\rho (v) = -e^v.
\]

Given the function \( \rho (\cdot) \), the corresponding GEL estimator is defined as

\[
\hat{\theta}_n^{GEL} = \arg \min_\theta \sup \lambda \sum_{i=1}^n \rho \left( \lambda^\top g(X_t, \theta) \right).
\]

While GMM only requires the minimization w.r.t. \( \theta \), the GEL requires two “loops” of optimization: the inner loop maximization over the auxiliary parameter \( \lambda \) and the outer loop minimization over the parameter \( \theta \). Note that the inner loop maximization is relatively easy because the objective function is concave in \( \lambda \); the concavity means that standard optimization procedures tend to be stable and converge fairly fast. So, GEL is harder to compute, but not too much.
GEL has special names for various choices of $\rho(\cdot)$:

$$\text{GEL} = \begin{cases} 
\text{CUE} & \rho \text{ is quadratic} \\
\text{Empirical Likelihood (EL)} & \rho(v) = \ln(1 - v) \\
\text{Exponential Tilting (ET)} & \rho(v) = -e^v
\end{cases}$$

Note that when $\rho(\cdot)$ is quadratic, the inner loop maximization is easily solved analytically, which leads to the simple objective function in (0.8).

For IID data, Newey and Smith (2004) show that GEL estimators are consistent and asymptotically normal. Moreover, they all have the same asymptotic variance as efficient GMM estimators (2-step or iterated). If we only look at consistency, asymptotic normality and the asymptotic variance, then we see no difference among these estimators. Newey and Smith compare the higher order bias of these estimators. They find four sources of bias for GMM

$$\text{Bias(GMM)} = B_I + B_G + B_S + B_W$$

where

- $B_I$ : Bias of the infeasible efficient GMM with $W_n = S^{-1}$
- $B_G$ : Bias due to estimation error of the Jacobian $G = \mathbb{E}[\nabla \theta g(X_t, \theta_0)]$
- $B_S$ : Bias due to estimation error of the matrix $S = \mathbb{E}\left[g(X_t, \theta_0)g(X_t, \theta_0)^\top\right]$
- $B_W$ : Bias due to the estimation error of the preliminary estimator.

and they show

$$\text{Bias(GEL)} = B_I + \left(1 + \frac{1}{2}\rho''(0)\right)B_S$$

Note that for CUE, $\rho$ is quadratic, $\rho''(0) = 0$; for EL, $\rho''(0) = -2$; for ET, $\rho''(0) = -1$. We hence have

$$\text{Bias(GMM)} = B_I + B_G + B_S + B_W$$
$$\text{Bias(CUE)} = B_I + B_S$$
$$\text{Bias(EL)} = B_I$$
$$\text{Bias(ET)} = B_I + \frac{1}{2}B_S.$$
The discussion above is for IID data. When the data are dependent, GEL estimators based on (0.9) are still consistent and asymptotically normal, but the asymptotic variance is greater (in matrix sense) than that of the efficient 2-step and iterated GMM. See for example Kitamura (2006). The CUE can be easily modified to achieve the efficient asymptotic variance, by re-defining $\hat{S}_n(\theta)$ as the long-run variance estimator of the $(g(X_t, \theta))_{1 \leq t \leq n}$. EL estimators can also be modified to achieve the efficient asymptotic variance; the modified estimator is called “blockwise empirical likelihood (BEL)”. See Kitamura (2006) for more details. Note that the Newey-Smith result is not directly applicable here because their theory was written for independent data.

**Further readings** (optional):

Hall 2005, “Generalized Method of Moments” Chapters 3, 4, 6
