Chapter 5

Almost sure convergence, convergence in probability and asymptotic normality

In the previous chapter we considered estimator of several different parameters. The hope is that as the sample size increases the estimator should get ‘closer’ to the parameter of interest. When we say closer we mean to converge. In the classical sense the sequence \( \{x_k\} \) converges to \( x \) (\( x_k \to x \)) if \( |x_k - x| \to 0 \) as \( k \to \infty \) (or for every \( \varepsilon > 0 \), there exists an \( n \) where for all \( k > n \), \( |x_k - x| < \varepsilon \)). Of course the estimators we have considered are random, that is for every \( \omega \in \Omega \) (set of all outcomes) we have a different estimate. The natural question to ask is what does convergence mean for random sequences.

5.1 Modes of convergence

We start by defining different modes of convergence.

Definition 5.1.1 (Convergence)  
- **Almost sure convergence** We say that the sequence \( \{X_t\} \) converges almost sure to \( \mu \), if there exists a set \( M \subset \Omega \), such that \( P(M) = 1 \) and for every \( \omega \in N \) we have \( X_t(\omega) \to \mu \).

In other words for every \( \varepsilon > 0 \), there exists an \( N(\omega) \) such that

\[
|X_t(\omega) - \mu| < \varepsilon,
\]

for all \( t > N(\omega) \). Note that the above definition is very close to classical convergence. We denote \( X_t \to \mu \) almost surely, as \( X_t \overset{a.s.}{\to} \mu \).

An equivalent definition, in terms of probabilities, is for every \( \varepsilon > 0 \) \( X_t \overset{a.s.}{\to} \mu \) if

\[
P(\omega; \cap_{m=1}^{\infty} \cup_{t=m}^{\infty} \{|X_t(\omega) - \mu| > \varepsilon\}) = 0.
\]

It is worth considering briefly what \( \cap_{m=1}^{\infty} \cup_{t=m}^{\infty} \{|X_t(\omega) - \mu| > \varepsilon\} \) means. If \( \cap_{m=1}^{\infty} \cup_{t=m}^{\infty} \{|X_t(\omega) - \mu| > \varepsilon\} \neq \emptyset \), then there exists an \( \omega^* \in \cap_{m=1}^{\infty} \cup_{t=m}^{\infty} \{|X_t(\omega) - \mu| > \varepsilon\} \) such that...
for some infinite sequence \( \{k_j\} \), we have \(|X_{k_j}(\omega^*) - \mu| > \varepsilon\), this means \(X_t(\omega^*)\) does not converge to \(\mu\). Now let \(\bigcap_{m=1}^\infty \bigcup_{t=m}^\infty \{\{|X_t(\omega) - \mu| > \varepsilon\} = A\), if \(P(A) = 0\), then for ‘most’ \(\omega\) the sequence \(\{X_t(\omega)\}\) converges.

- **Convergence in mean square**
  
  We say \(X_t \rightarrow \mu\) in mean square (or \(L_2\) convergence), if \(\mathbb{E}(X_t - \mu)^2 \rightarrow 0\) as \(t \rightarrow \infty\).

- **Convergence in probability**
  
  Convergence in probability cannot be stated in terms of realisations \(X_t(\omega)\) but only in terms of probabilities. \(X_t\) is said to converge to \(\mu\) in probability (written \(X_t \overset{P}{\rightarrow} \mu\)) if
  
  \[
  P(|X_t - \mu| > \varepsilon) \rightarrow 0, \quad t \rightarrow \infty.
  \]
  
  Often we write this as \(|X_t - \mu| = o_p(1)\).

  If for any \(\gamma \geq 1\) we have
  
  \[
  \mathbb{E}(X_t - \mu)^\gamma \rightarrow 0 \quad t \rightarrow \infty,
  \]

  then it implies convergence in probability (to see this, use Markov’s inequality).

- **Rates of convergence:**
  
  (i) We say the stochastic process \(\{X_t\}\) is \(|X_t - \mu| = O_p(a_t)\), if the sequence \(\{a_t^{-1}|X_t - \mu|\}\) is bounded is bounded in probability (this is bounded below). We see from the definition of boundedness, that for all \(t\), the distribution of \(a_t^{-1}|X_t - \mu|\) should mainly lie within a certain interval. In general \(a_t \rightarrow \infty\) as \(t \rightarrow \infty\). Hence in the distribution of \(a_t^{-1}|X_t - \mu|\) lies within a certain interval, then \(|X_t - \mu|\) will within an ever decreasing one.

  (ii) We say the stochastic process \(\{X_t\}\) is \(|X_t - \mu| = o_p(a_t)\), if the sequence \(\{a_t^{-1}|X_t - \mu|\}\) converges in probability to zero.

**Definition 5.1.2 (Boundedness)**

(i) **Almost surely bounded** If the random variable \(X\) is almost surely bounded, then for a positive sequence \(\{e_k\}\), such that \(e_k \rightarrow \infty\) as \(k \rightarrow \infty\) (typically \(e_k = 2^k\) is used), we have

\[
P(\omega; \{\bigcup_{k=1}^\infty \{|X(\omega)| \leq e_k\}\}) = 1.
\]

Usually to prove the above we show that

\[
P(\omega; \{\bigcup_{k=1}^\infty \{|X| \leq e_k\}\}^c) = 0.
\]

Since \((\bigcup_{k=1}^\infty \{|X| \leq e_k\})^c = \bigcap_{k=1}^\infty \{|X| > e_k\} \subset \bigcap_{k=1}^\infty \bigcup_{m=k}^\infty \{|X| > e_k\}\), to show the above we show

\[
P(\omega; \{\bigcap_{k=1}^\infty \bigcup_{m=k}^\infty \{|X(\omega)| > e_k\}\}) = 0.
\]

We note that if \((\omega; \{\bigcap_{k=1}^\infty \bigcup_{m=k}^\infty \{|X(\omega)| > e_k\}\}) \neq \emptyset\), then there exists a \(\omega^* \in \Omega\) and an infinite subsequence \(k_j\), where \(|X(\omega^*)| > e_{k_j}\), hence \(X(\omega^*)\) is not bounded (since \(e_k \rightarrow \infty\)).
To prove (5.2) we usually use the Borel Cantelli Lemma. This states that if \( \sum_{k=1}^{\infty} P(A_k) < \infty \), the events \( \{A_k\} \) occur only finitely often with probability one. Applying this to our case, if we can show that \( \sum_{m=1}^{\infty} P(\omega: \{|X(\omega)| > e_m|\}) < \infty \), then \( \{|X(\omega)| > e_m|\} \) happens only finitely often with probability one. Hence if \( \sum_{m=1}^{\infty} P(\omega: \{|X(\omega)| > e_m|\}) < \infty \), then \( X \) is a bounded random variable.

It is worth noting that often we choose the sequence \( e_k = 2^k \), in this case \( \sum_{m=1}^{\infty} P(\omega: \{|X(\omega)| > e_m|\}) = \sum_{m=1}^{\infty} P(\omega: \{|X(\omega)| > \log 2^k|\}) \leq C \mathbb{E}(\log |X|) \). Hence if we can show that \( \mathbb{E}(\log |X|) < \infty \), then \( X \) is bounded almost surely.

(ii) **Bounded in probability** A sequence is bounded in probability, written \( X_t = O_p(1) \), if for every \( \varepsilon > 0 \), there exists a \( \delta(\varepsilon) < \infty \) such that \( P(|X_t| \geq \delta(\varepsilon)) < \varepsilon \). Roughly speaking this means that the sequence is only extremely large with a very small probability. And as the ‘largeness’ grows the probability declines.

### 5.2 Ergodicity

To motivate the notion of ergodicity we recall the strong law of large numbers (SLLN). Suppose \( \{X_t\}_t \) is an iid random sequence, and \( \mathbb{E}(|X_0|) < \infty \) then by the SLLN we have that

\[
\frac{1}{n} \sum_{j=1}^{n} X_t \overset{a.s.}{\rightarrow} \mathbb{E}(X_0),
\]

for the proof see, for example, Grimmett and Stirzaker (1994). It would be useful to generalise this result and find weaker conditions on \( \{X_t\} \) for this result to still hold true. A simple application is when we want to estimate the mean \( \mu \), and we use \( \frac{1}{n} \sum_{j=1}^{n} X_t \) as an estimator of the mean.

It can be shown that if \( \{X_t\} \) is an ergodic process then the above result holds. That is if \( \{X_t\} \) is an ergodic process then for any function \( h \) such that \( \mathbb{E}(h(X_0)) < \infty \) we have

\[
\frac{1}{n} \sum_{j=1}^{n} h(X_t) \overset{a.s.}{\rightarrow} \mathbb{E}(h(X_0)).
\]

Note that the result does not state anything about the rate of convergence. Ergodicity is normally defined in terms of measure preserving transformations. However, we do not formally define ergodicity here, but needless to say all ergodic processes are stationary. For the definition of ergodicity and a full treatment see, for example, Billingsley (1995).

However below we do state a result which characterises a general class of ergodic processes.

**Theorem 5.2.1** Suppose \( \{Z_t\} \) is an ergodic sequence (for example iid random variables) and \( g: \mathbb{R}^{\infty} \rightarrow \mathbb{R} \) is a measurable function (its really hard to think up nonmeasurable functions). Then the sequence \( \{Y_t\}_t \), where

\[
Y_t = g(Z_t, Z_{t-1}, \ldots),
\]

is an ergodic process.
PROOF. See Stout (1974), Theorem 3.5.8. □

Example 5.2.1  
(i) The process \( \{Z_t\}_t \), where \( \{Z_t\} \) are iid random variables, is probably the simplest example of an ergodic sequence.

(ii) A simple example of a time series \( \{X_t\} \) which is not independent but is ergodic is the AR(1) process. We recall that the AR(1) process satisfies the representation

\[
X_t = \phi X_{t-1} + \epsilon_t,
\]

where \( \{\epsilon_t\}_t \) are iid random variables with \( \mathbb{E}(\epsilon_t) = 0 \), \( \mathbb{E}(\epsilon_t^2) = 1 \) and \( |\phi| < 1 \). It has the unique causal solution

\[
X_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}.
\]

The solution motivates us to define the function

\[
g(x_0, x_1, \ldots) = \sum_{j=0}^{\infty} \phi^j x_j.
\]

Since \( g(\cdot) \) is bounded, it is sufficiently well behaved (thus measurable). Which implies, by using Theorem 5.2.1, that \( \{X_t\} \) is an ergodic process. We note if \( \mathbb{E}(\epsilon^2) < \infty \), then \( \mathbb{E}(X^2) < \infty \).

- The ARCH(p) process \( \{X_t\} \) defined by \( X_t = Z_t \sigma_t \) where \( \sigma_t^2 = a_0 + \sum_{j=1}^{p} a_j X_{t-j}^2 \) with \( \sum_{j=1}^{p} a_j < 1 \) is ergodic stochastic process (we look at this model in a later Chapter).

Example 5.2.2 (Application)  
If \( \{X_t\} \) is an AR(1) process with \( |a| < 1 \) and \( \mathbb{E}(\epsilon_t^2) < \infty \), then by using the ergodic theorem we have

\[
\frac{1}{n} \sum_{t=1}^{n} X_t \xrightarrow{a.s.} \mathbb{E}(X_0 X_k).
\]

5.3 Sampling properties

Often we will estimate the parameters by maximising (or minimising) a criterion. Suppose we have the criterion \( \mathcal{L}_n(a) \) (eg. likelihood, quasi-likelihood, Kullback-Leibler etc) we use as an estimator of \( a_0, \hat{a}_n \) where

\[
\hat{a}_n = \arg \max_{a \in \Theta} \mathcal{L}_n(a)
\]

and \( \Theta \) is the parameter space we do the maximisation (minimisation) over. Typically the true parameter \( a \) should maximise (minimise) the ‘limiting’ criterion \( \mathcal{L} \).

If this is to be a good estimator, as the sample size grows the estimator should converge (in some sense) to the parameter we are interesting in estimating. As we discussed above, there are various modes in which we can measure this convergence (i) almost surely (ii) in probability and (iii) in mean squared error. Usually we show either (i) or (ii) (noting that (i) implies (ii)), in time series its usually quite difficult to show (iii).
**Definition 5.3.1**  (i) An estimator \( \hat{a}_n \) is said to be almost surely consistent estimator of \( a_0 \), if there exists a set \( M \subset \Omega \), where \( P(M) = 1 \) and for all \( \omega \in M \) we have
\[
\hat{a}_n(\omega) \to a.
\]
(ii) An estimator \( \hat{a}_n \) is said to converge in probability to \( a_0 \), if for every \( \delta > 0 \)
\[
P(|\hat{a}_n - a| > \delta) \to 0 \quad T \to \infty.
\]

To prove either (i) or (ii) usually involves verifying two main things, pointwise convergence and equicontinuity.

### 5.4 Showing almost sure convergence of an estimator

We now consider the general case where \( L_n(a) \) is a ‘criterion’ which we maximise. Let us suppose we can write \( L_n \) as
\[
L_n(a) = \frac{1}{n} \sum_{t=1}^{n} \ell_t(a),
\]
where for each \( a \in \Theta \), \( \{\ell_t(a)\}_t \) is a ergodic sequence. Let
\[
L(a) = \mathbb{E}(\ell_t(a)),
\]
we assume that \( L(a) \) is continuous and has a unique maximum in \( \Theta \). We define the estimator \( \hat{a}_n \) where \( \hat{a}_n = \arg \min_{a \in \Theta} L_n(a) \).

**Definition 5.4.1 (Uniform convergence)** \( L_n(a) \) is said to almost surely converge uniformly to \( L(a) \), if
\[
\sup_{a \in \Theta} |L_n(a) - L(a)| \overset{a.s.}{\to} 0.
\]
In other words there exists a set \( M \subset \Omega \) where \( P(M) = 1 \) and for every \( \omega \in M \),
\[
\sup_{a \in \Theta} |L_n(\omega, a) - L(a)| \to 0.
\]

**Theorem 5.4.1 (Consistency)** Suppose that \( \hat{a}_n = \arg \max_{a \in \Theta} L_n(a) \) and \( a_0 = \arg \max_{a \in \Theta} L(a) \) is the unique minimum. If \( \sup_{a \in \Theta} |L_n(a) - L(a)| \overset{a.s.}{\to} 0 \) as \( n \to \infty \) and \( L(a) \) has a unique maximum. Then \( \hat{a}_n \overset{a.s.}{\to} a_0 \) as \( n \to \infty \).

**Proof.** We note that by definition we have \( L_n(a_0) \leq L_n(\hat{a}_n) \) and \( L(\hat{a}_n) \leq L(a_0) \). Using this inequality we have
\[
L_n(a_0) - L(a_0) \leq L_n(\hat{a}_n) - L(a_0) \leq L_n(\hat{a}_n) - L(\hat{a}_n).
\]
Therefore from the above we have
\[
|L_n(\hat{a}_T) - L(a_0)| \leq \max \{|L_n(a_0) - L(a_0)|, |L_n(\hat{a}_T) - L(\hat{a}_n)|\} \leq \sup_{a \in \Theta} |L_n(a) - L(a)|.
\]
Hence since we have uniform converge we have $$|\mathcal{L}_n(\hat{a}_n) - \mathcal{L}(a_0)| \overset{a.s.}{\to} 0$$ as $$n \to \infty$$. Now since $$\mathcal{L}(a)$$ has a unique maximum, we see that $$|\mathcal{L}_n(\hat{a}_n) - \mathcal{L}(a_0)| \overset{a.s.}{\to} 0$$ implies $$\hat{a}_n \overset{a.s.}{\to} a_0$$. □

We note that directly establishing uniform convergence is not easy. Usually it is done by assuming the parameter space is compact and showing point wise convergence and stochastic equicontinuity, these three facts imply uniform convergence. Below we define stochastic equicontinuity and show consistency under these conditions.

**Definition 5.4.2** The sequence of stochastic functions $$\{f_n(a)\}_n$$ is said to be stochastically equicontinuous if there exists a set $$M \in \Omega$$ where $$P(M) = 1$$ and for every $$\omega \in M$$ and and $$\varepsilon > 0$$, there exists a $$\delta$$ and such that for every $$\omega \in M$$

$$\sup_{|a_1 - a_2| \leq \delta} |f_n(\omega, a_1) - f_n(\omega, a_2)| \leq \varepsilon,$$

for all $$n > N(\omega)$$.

A sufficient condition for stochastic equicontinuity of $$f_n(a)$$ (which is usually used to prove equicontinuity), is that $$f_n(a)$$ is almost surely Lipschitz continuous. In other words, there exists a random variable $$K$$ which is almost surely bounded, where for all $$\omega \in M$$ ($$P(M) = 1$$) we have

$$\sup_{a_1, a_2 \in \Theta} |f_n(\omega, a_1) - f_n(\omega, a_2)| < K(\omega)\|a_1 - a_2\|.$$

In the following theorem we state sufficient conditions for almost sure uniform convergence. It is worth noting this is the Arzela-Ascoli theorem for random variables.

**Theorem 5.4.2 (The stochastic Ascoli Lemma)** Suppose the parameter space $$\Theta$$ is compact, for every $$a \in \Theta$$ we have $$\mathcal{L}_n(a) \overset{a.s.}{\to} \mathcal{L}(a)$$ and $$\mathcal{L}_n(a)$$ is stochastic equicontinuous. Then

$$\sup_{a \in \Theta} |\mathcal{L}_n(a) - \mathcal{L}(a)| \overset{a.s.}{\to} 0$$ as $$n \to \infty$$.

We use the theorem below.

**Corollary 5.4.1** Suppose that $$\hat{a}_n = \arg \max_{a \in \Theta} \mathcal{L}_n(a)$$ and $$a_0 = \arg \max_{a \in \Theta} \mathcal{L}(a)$$, moreover $$\mathcal{L}(a)$$ has a unique maximum. If

(i) we have point wise convergence, that is for every $$a \in \Theta$$ we have $$\mathcal{L}_n(a) \overset{a.s.}{\to} \mathcal{L}(a)$$.

(ii) The parameter space $$\Theta$$ is compact.

(iii) $$\mathcal{L}_n(a)$$ is stochastic equicontinuous.

Then $$\hat{a}_n \overset{a.s.}{\to} a_0$$ as $$n \to \infty$$.

We prove Theorem 5.4.2 in the section below, but it can be omitted on first reading.

**5.4.1 Proof of Theorem 5.4.2 (The stochastic Ascoli theorem)**

We now show that stochastic equicontinuity and almost pointwise convergence imply uniform convergence. We note that on its own, pointwise convergence is a much weaker condition than uniform convergence, since for pointwise convergence the rate of convergence can be different for each parameter.
Before we continue a few technical points. We recall that we are assuming almost pointwise convergence. This means for each parameter $a \in \Theta$ there exists a set $N_a \in \Omega$ (with $P(N_a) = 1$) such that for all $\omega \in N_a$ $L_t(\omega, a) \rightarrow L(a)$. In the following lemma we unify this set. That is show (using stochastic equicontinuity) that there exists a set $N \in \Omega$ (with $P(N) = 1$) such that for all $\omega \in N$ $L_t(\omega, a) \rightarrow L(a)$.

**Lemma 5.4.1** Suppose the sequence $\{L_n(a)\}_n$ is stochastically equicontinuous and also pointwise convergent (that is $L_n(a)$ converges almost surely to $L(a)$), then there exists a set $M \in \Omega$ where $\omega \in N$ and $a \in \Theta$ we have

$$|L_n(\omega, a) - L(a)| \rightarrow 0.$$ 

**Proof.** Enumerate all the rationals in the set $\Theta$ and call this sequence $\{a_i\}_i$. Then for every $a_i$ there exists a set $M_{a_i}$ where $P(M_{a_i}) = 1$, such that for every $\omega \in M_{a_i}$ we have $|L_T(\omega, a_i) - L(a_i)| \rightarrow 0$. Define $M = \cap M_{a_i}$, since the number of sets is countable $P(M) = 1$ and for every $\omega \in M$ and $a_i$ we have $L_n(\omega, a_i) \rightarrow L(a_i)$.

Since we have stochastic equicontinuity, there exists a set $\bar{M}$ (with $P(M) = 1$), such that for every $\omega \in \bar{M}$, $\{L_n(\omega, \cdot)\}$ is equicontinuous. Let $\bar{M} = \bar{M} \cap \cap M_{a_i}$, we will show that for all $a \in \Theta$ and $\omega \in \bar{M}$ we have $L_n(\omega, a) \rightarrow L(a)$. By stochastic equicontinuity for every $\omega \in \bar{M}$ and $\epsilon/3 > 0$, there exists a $\delta > 0$ such that

$$\sup_{|b_1 - b_2| \leq \delta} |L_n(\omega, b_1) - L_n(\omega, b_2)| \leq \epsilon/3, \quad (5.6)$$

for all $n > N(\omega)$. Furthermore by definition of $\bar{M}$ for every rational $a_j \in \Theta$ and $\omega \in \bar{M}$ we have

$$|L_T(\omega, a_j) - L(a_j)| \leq \epsilon/3, \quad (5.7)$$

where $n > N'(\omega)$. Now for any given $a \in \Theta$, there exists a rational $a_i$ such that $\|a - a_j\| \leq \delta$. Using this, (5.6) and (5.7) we have

$$|L_n(\omega, a) - L(a)| \leq |L_n(\omega, a) - L_n(\omega, a_i)| + |L_n(\omega, a_i) - L(a_i)| + |L(a) - L(a_i)| \leq \epsilon,$$

for $n > \max(N(\omega), N'(\omega))$. To summarise for every $\omega \in \bar{M}$ and $a \in \Theta$, we have $|L_n(\omega, a) - L(a)| \rightarrow 0$. Hence we have pointwise convergence for every realisation in $\bar{M}$. \hfill \Box

We now show that equicontinuity implies uniform convergence.

**Proof of Theorem 5.4.2.** Using Lemma 5.4.1 we see that there exists a set $\bar{M} \in \Omega$ with $P(\bar{M}) = 1$, where $L_n$ is equicontinuous and also pointwise convergent. We now show uniform convergence on this set. Choose $\epsilon/3 > 0$ and let $\delta$ be such that for every $\omega \in \bar{M}$ we have

$$\sup_{|a_1 - a_2| \leq \delta} |L_T(\omega, a_1) - L_T(\omega, a_2)| \leq \epsilon/3, \quad (5.8)$$

for all $n > n(\omega)$. Since $\Theta$ is compact it can be divided into a finite number of open sets. Construct the sets $\{O_i\}_{i=1}^p$, such that $\Theta \subset \cup_{i=1}^p O_i$ and $\sup_{x,y,i} \|x - y\| \leq \delta$. Let $\{a_i\}_{i=1}^p$ be such that $a_i \in O_i$. We note that for every $\omega \in \bar{M}$ we have $L_n(\omega, a_i) \rightarrow L(a_i)$, hence for every $\epsilon/3,$
there exists an \( n_i(\omega) \) such that for all \( n > n_i(\omega) \) we have \( |L_T(\omega, a_i) - L(a_i)| \leq \varepsilon/3 \). Therefore, since \( p \) is finite (due to compactness), there exists a \( \tilde{n}(\omega) \) such that

\[
\max_{1 \leq i \leq p} |L_n(\omega, a_i) - L(a_i)| \leq \varepsilon/3,
\]

for all \( n > \tilde{n}(\omega) = \max_{1 \leq i \leq p}(n_i(\omega)) \). For any \( a \in \Theta \), choose the \( i \), such that open set \( O_i \) such that \( a \in O_i \). Using (5.8) we have

\[
|L_T(\omega, a) - L_T(\omega, a_i)| \leq \varepsilon/3,
\]

for all \( n > \tilde{n}(\omega) \). Altogether this gives

\[
|L_T(\omega, a) - L(a)| \leq |L_T(\omega, a) - L_T(\omega, a_i)| + |L_T(\omega, a_i) - L(a_i)| + |L(a) - L(a_i)| \leq \varepsilon,
\]

for all \( n \geq \max(n(\omega), \tilde{n}(\omega)) \). We observe that \( \max(n(\omega), \tilde{n}(\omega)) \) and \( \varepsilon/3 \) does not depend on \( a \), therefore for all \( n \geq \max(n(\omega), \tilde{n}(\omega)) \) and we have \( \sup_a |L_n(\omega, a) - L(a)| < \varepsilon \). This gives for every \( \omega \in \bar{M} \) \( (\mathbb{P}(\bar{M}) = 1) \), \( \sup_a |L_n(\omega, a) - L(a)| \to 0 \), thus we have almost sure uniform convergence. \( \square \)

### 5.5 Almost sure convergence of the least squares estimator for an AR\((p)\) process

In Chapter ?? we will consider the sampling properties of many of the estimators defined in Chapter ??, However to illustrate the consistency result above we apply it to the least squares estimator of the autoregressive parameters.

To simply notation we only consider estimator for \( \text{AR}(1) \) models. Suppose that \( X_t \) satisfies \( X_t = \phi X_{t-1} + \varepsilon_t \). To estimate \( \phi \) we use the least squares estimator defined below. Let

\[
L_n(a) = \frac{1}{n-1} \sum_{t=2}^{n} (X_t - aX_{t-1})^2,
\]

we use \( \hat{\phi}_n \) as an estimator of \( \phi \), where

\[
\hat{\phi}_n = \arg \max_{a \in \Theta} L_T(a), \tag{5.10}
\]

where \( \Theta = [-1, 1] \).

How can we show that this is consistent?

- In the case of least squares for AR processes, \( \hat{a}_T \) has the explicit form

\[
\hat{\phi}_n = \frac{\frac{1}{n-1} \sum_{t=2}^{n} X_t X_{t-1}}{\frac{1}{n-1} \sum_{t=1}^{n-1} X_t^2}.
\]

Now by just applying the ergodic theorem to the numerator and denominator we get \( \hat{\phi}_n \overset{\text{a.s.}}{\to} \phi \).

It is worth noting, that \( \left| \frac{\frac{1}{n-1} \sum_{t=2}^{n} X_t X_{t-1}}{\frac{1}{n-1} \sum_{t=1}^{n-1} X_t^2} \right| < 1 \) is not necessarily true.
However we will tackle the problem in a rather artificial way and assume that it does not have an explicit form and instead assume that \( \hat{\phi}_n \) is obtained by minimising \( L_n(a) \) using a numerical routine. In general this is the most common way of minimising a likelihood function (usually explicit solutions do not exist).

In order to derive the sampling properties of \( \hat{\phi}_n \) we need to directly study the likelihood function \( L_n(a) \). We will do this now in the least squares case.

We will first show almost sure convergence, which will involve repeated use of the ergodic theorem. We will then demonstrate how to show convergence in probability. We look at almost sure convergence as its easier to follow. Note that almost sure convergence implies convergence in probability (but the converse is not necessarily true).

The first thing to do it let

\[
\ell_t(a) = (X_t - aX_{t-1})^2.
\]

Since \( \{X_t\} \) is an ergodic process (recall Example 5.2.1(ii)) by using Theorem 5.2.1 we have for \( a \), that \( \{\ell_t(a)\}_t \) is an ergodic process. Therefore by using the ergodic theorem we have

\[
L_n(a) = \frac{1}{n-1} \sum_{t=2}^{n} \ell_t(a) \xrightarrow{a.s.} \mathbb{E}(\ell_0(a)).
\]

In other words for every \( a \in [-1,1] \) we have that \( L_n(a) \xrightarrow{a.s.} \mathbb{E}(\ell_0(a)) \) (almost sure pointwise convergence).

Since the the parameter space \([-1,1]\) is compact and \( a \) is the unique minimum of \( \ell(\cdot) \) in the parameter space, then all that remains is to show show stochastic equicontinuity, from this we deduce almost sure uniform convergence.

To show stochastic equicontinuity we expand \( L_T(a) \) and use the mean value theorem to obtain

\[
L_n(a_1) - L_n(a_2) = \nabla L_T(\bar{a})(a_1 - a_2),
\]

where \( \bar{a} \in \text{[min}[a_1, a_2], \text{max}[a_1, a_2]] \) and

\[
\nabla L_n(\bar{a}) = \frac{-2}{n-1} \sum_{t=2}^{n} X_{t-1}(X_t - \bar{a}X_{t-1}).
\]

Because \( \bar{a} \in [-1,1] \) we have

\[
|\nabla L_n(\bar{a})| \leq D_n, \text{ where } D_n = \frac{2}{n-1} \sum_{t=2}^{n}(|X_{t-1}X_t| + X_{t-1}^2).
\]

Since \( \{X_t\}_t \) is an ergodic process, then \( \{|X_{t-1}X_t| + X_{t-1}^2\} \) is an ergodic process. Therefore, if \( \text{var}(\epsilon_0) < \infty \), by using the ergodic theorem we have

\[
D_n \xrightarrow{a.s.} 2\mathbb{E}(|X_{t-1}X_t| + X_{t-1}^2).
\]

56
Let \( D := 2\mathbb{E}(|X_{t-1}X_t| + X_t^2) \). Therefore there exists a set \( M \subset \Omega \), where \( \mathbb{P}(M) = 1 \) and for every \( \omega \in M \) and \( \varepsilon > 0 \) we have
\[
|D_T(\omega) - D| \leq \delta^*,
\]
for all \( n > N(\omega) \). Substituting the above into (5.11) we have
\[
|L_n(\omega, a_1) - L_n(\omega, a_2)| \leq D_n(\omega)|a_1 - a_2| \leq (D + \delta^*)|a_1 - a_2|,
\]
for all \( n \geq N(\omega) \). Therefore for every \( \varepsilon > 0 \), there exists a \( \delta := \varepsilon/(D + \delta^*) \) such that
\[
\sup_{|a_1 - a_2| \leq \varepsilon/(D + \delta^*)} |L_n(\omega, a_1) - L_n(\omega, a_2)| \leq \varepsilon,
\]
for all \( n \geq N(\omega) \). Since this is true for all \( \omega \in M \) we see that \( \{L_n(a)\} \) is stochastically equicontinuous.

**Theorem 5.5.1** Let \( \hat{\phi}_n \) be defined as in (5.10). Then we have \( \hat{\phi}_n \overset{a.s.}{\rightarrow} \phi \).

**PROOF.** Since \( \{L_n(a)\} \) is almost sure equicontinuous, the parameter space \([-1, 1]\) is compact and we have pointwise convergence of \( L_n(a) \overset{a.s.}{\rightarrow} \mathcal{L}(a) \), by using Theorem 5.4.1 we have that \( \hat{\phi}_n \overset{a.s.}{\rightarrow} a \), where \( a = \min_{a \in \Theta} \mathcal{L}(a) \). Finally we need to show that \( a = \phi \). Since
\[
\mathcal{L}(a) = \mathbb{E}(\ell_0(a)) = -\mathbb{E}(X_1 - aX_0)^2,
\]
we see by differentiating \( \mathcal{L}(a) \) with respect to \( a \), that it is minimised at \( a = \mathbb{E}(X_0X_1)/\mathbb{E}(X_0^2) \), hence \( a = \mathbb{E}(X_0X_1)/\mathbb{E}(X_0^2) \). To show that this is \( \phi \), we note that by the Yule-Walker equations
\[
X_t = \phi X_{t-1} + \epsilon_t \quad \Rightarrow \quad \mathbb{E}(X_tX_{t-1}) = \phi\mathbb{E}(X_{t-1}^2) + \mathbb{E}(\epsilon_t X_{t-1}).
\]
Therefore \( \phi = \mathbb{E}(X_0X_1)/\mathbb{E}(X_0^2) \), hence \( \hat{\phi}_n \overset{a.s.}{\rightarrow} \phi \). \( \square \)

We note that by using a very similar methods we can show strong consistency of the least squares estimator of the parameters in an AR(\( p \)) model.

### 5.6 Convergence in probability of an estimator

We described above almost sure (strong) consistency where we showed \( \hat{a}_T \overset{a.s.}{\rightarrow} a_0 \). Sometimes its not possible to show strong consistency (when ergodicity etc. cannot be verified). Often, as an alternative, weak consistency where \( \hat{a}_T \overset{P}{\rightarrow} a_0 \) (convergence in probability), is shown. This requires a weaker set of conditions, which we now describe:

(i) The parameter space \( \Theta \) should be compact.

(ii) We pointwise convergence: for every \( a \in \Theta \) \( L_n(a) \overset{P}{\rightarrow} \mathcal{L}(a) \).
(iii) Equicontinuity in probability, that is for every $\epsilon > 0$ there exists a $\delta$ such that

$$\lim_{n \to \infty} \mathbb{P} \left( \sup_{|a_1 - a_2| \leq \delta} |\mathcal{L}_n(a_1) - \mathcal{L}(a_2)| > \epsilon \right) \to 0.$$ 

If the above conditions are satisfied we have $\hat{a}_T \xrightarrow{p} a_0$.

Verifying conditions (ii) and (iii) may look a little daunting but actually with the use of Chebyshev’s (or Markov’s) inequality it can be quite straightforward. For example if we can show that for every $a \in \Theta$

$$\mathbb{E}(\mathcal{L}_T(a) - \mathcal{L}(a))^2 \to 0 \quad T \to \infty.$$ 

Therefore by applying Chebyshev’s inequality we have for every $\epsilon > 0$ that

$$P(|\mathcal{L}_T(a) - \mathcal{L}(a)| > \epsilon) \leq \frac{\mathbb{E}(\mathcal{L}_T(a) - \mathcal{L}(a))^2}{\epsilon^2} \to 0 \quad T \to \infty.$$ 

Thus for every $a \in \Theta$ we have $\mathcal{L}_T(a) \xrightarrow{D} \mathcal{L}(a)$.

To show (iii) we often use the mean value theorem $\mathcal{L}_T(a)$. Using the mean value theorem we have

$$|\mathcal{L}_T(a_1) - \mathcal{L}(a_2)| \leq \sup_a \|\nabla_a \mathcal{L}_T(a)\|_2 \|a_1 - a_2\|.$$ 

Now if it can be shown that $\sup_a \|\nabla_a \mathcal{L}_T(a)\|_2$ is bounded in probability (usually we show that $\mathbb{E}(\sup_a \|\nabla_a \mathcal{L}_T(a)\|_2) < \infty$), then we have (ii).

5.7 Asymptotic normality of an estimator

The first central limit theorem goes back to the asymptotic distribution of sums of binary random variables (these have a binomial distribution and Bernoulli showed that they could be approximated to a normal distribution). This result was later generalised to sums of iid random variables. However from mid 20th century to late 20th century several advances have been made for generalising the results to dependent random variables. These include generalisations to random variables which have $n$-dependence, mixing properties, cumulant properties, near-epoch dependence etc. In this section we will concentrate on a central limit theorem for martingales. Our reason for choosing this flavour of CLT is that it can be applied in various estimation settings - as it can often be shown that the derivative of a criterion at the true parameter is a martingale.

Let us suppose that

$$\hat{a}_n = \arg \max_{a \in \Theta} \mathcal{L}_n(a),$$

where

$$\mathcal{L}_n(a) = \frac{1}{n} \sum_{t=1}^{n} \ell_t(a),$$

and for each $a \in \Theta$, $\{\ell_t(a)\}_t$ are identically distributed random variables.
In this section we shall show asymptotic normality of $\sqrt{n}(\hat{a}_n - a_0)$. The reason for normalising by $\sqrt{n}$ is that $(\hat{a}_n - a_0) \xrightarrow{a.s.} 0$ as $n \to \infty$, hence in terms of distributions it converges towards the point mass at zero. Therefore we need to increase the magnitude of the difference $\hat{a}_n - a_0$. We can show that $(\hat{a}_n - a_0) = O(n^{-1/2})$, therefore $\sqrt{n}(\hat{a}_n - a) = O(1)$.

We often use $\nabla L_n(a)$ to denote the partial derivative of $L_n(a)$ with respect to $a$ ($\nabla L_n(a) = \frac{\partial L_n(a)}{\partial a_1}, \ldots, \frac{\partial L_n(a)}{\partial a_p}$). Since $\hat{a}_T = \text{arg max } L_n(a)$, we observe that $\nabla L_n(\hat{a}_n) = 0$. Now expanding $\nabla L_n(\hat{a}_n)$ about $a_0$ (the true parameter) we have

$$\nabla L_n(\hat{a}_n) = \nabla L_n(a_0) + (\hat{a}_n - a_0)\nabla^2 L_n(\hat{a})$$

$$\Rightarrow (\hat{a}_n - a_0) = -\{\nabla^2 L_n(\hat{a})\}^{-1}\nabla L_n(a_0)$$

To show asymptotically normality of $\sqrt{n}(\hat{a}_n - a_0)$, first asymptotic normality of $\nabla L_n(a_0)$ is shown, second it is shown that $\nabla^2 L_n(\hat{a}) \xrightarrow{P} \mathbb{E}(\nabla^2 \ell_0(a_0))$, together they yield asymptotically normality of $\sqrt{n}(\hat{a}_n - a_0)$. In many cases $\nabla L_n(a_0)$ is a martingale, hence the martingale central limit theorem is usually applied to show asymptotic normality of $\nabla L_n(a_0)$. We start by defining a martingale and stating the martingale central limit theorem.

**Definition 5.7.1** \{Z_t; t = 1, \ldots, \infty\} are called martingale differences if

$$\mathbb{E}(Z_t|Z_{t-1}, Z_{t-2}, \ldots) = 0.$$  

An example is the sequence \{X_{t-1}\epsilon_t\}_t considered above. Because $\epsilon_t$ and $X_{t-1}, X_{t-2}, \ldots$ are independent then $\mathbb{E}(X_{t-1}\epsilon_t|X_{t-1}, X_{t-2}, \ldots) = 0$. The sequence $\{S_T\}_T$, where

$$S_T = \sum_{k=1}^{T} Z_t$$

are called martingales if \{Z_t\} are martingale differences.

Let us define $S_T$ as

$$S_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} Y_t,$$

where $\mathcal{F}_t = \sigma(Y_t, Y_{t-1}, \ldots)$, $\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = 0$ and $\mathbb{E}(Y_t^2) < \infty$. In the following theorem adapted from Hall and Heyde (1980), Theorem 3.2 and Corollary 3.1, we show that $S_T$ is asymptotically normal.

**Theorem 5.7.1** Let $\{S_T\}_T$ be defined as in (6.30). Further suppose

$$\frac{1}{T} \sum_{t=1}^{T} Y_t^2 \xrightarrow{P} \sigma^2,$$

where $\sigma^2$ is a finite constant, for all $\varepsilon > 0$, 

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(Y_t^2 I(|Y_t| > \varepsilon \sqrt{T})|\mathcal{F}_{t-1}) \xrightarrow{P} 0,$$  

59
(this is known as the conditional Lindeberg condition) and
\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(Y_t^2 | \mathcal{F}_{t-1}) \xrightarrow{p} \sigma^2.
\]  \hspace{1cm} (5.16)

Then we have
\[
S_T \xrightarrow{d} \mathcal{N}(0, \sigma^2).
\]  \hspace{1cm} (5.17)

### 5.8 Asymptotic normality of the least squares estimator

In this section we show asymptotic normality of the least squares estimator of the AR(1) defined in (5.9).

We call that the least squares estimator is \( \hat{\phi}_n = \arg \max_{a \in [-1, 1]} \mathcal{L}_n(a) \). Recalling the criterion
\[
\mathcal{L}_n(a) = \frac{1}{n-1} \sum_{t=2}^{n} (X_t - aX_{t-1})^2,
\]
the first and the second derivative is
\[
\nabla \mathcal{L}_n(a) = - \frac{2}{n-1} \sum_{t=2}^{n} X_{t-1}(X_t - aX_{t-1}) = - \frac{2}{n-1} \sum_{t=2}^{n} X_{t-1} \epsilon_t
\]
and \( \nabla^2 \mathcal{L}_n(a) = \frac{2}{n-1} \sum_{t=2}^{n} X_{t-1}^2 \).

Therefore by using (5.12) we have
\[
(\hat{\phi}_n - \phi) = - (\nabla^2 \mathcal{L}_n)^{-1} \nabla \mathcal{L}_n(\phi).
\]  \hspace{1cm} (5.18)

Since \( \{X_t^2\} \) are ergodic random variables, by using the ergodic theorem we have \( \nabla^2 \mathcal{L}_n \stackrel{a.s.}{\to} 2 \mathbb{E}(X_0^2) \). This with (5.18) implies
\[
\sqrt{n}(\hat{\phi}_n - \phi) = - \frac{(\nabla^2 \mathcal{L}_n)^{-1}}{\sqrt{n} \nabla \mathcal{L}_n(\phi)} \stackrel{a.s.}{\to} \mathcal{N}(0, \sigma^2).
\]  \hspace{1cm} (5.19)

To show asymptotic normality of \( \sqrt{n}(\hat{\phi}_n - \phi) \), will show asymptotic normality of \( \sqrt{n} \nabla \mathcal{L}_n(\phi) \).

We observe that
\[
\nabla \mathcal{L}_n(\phi) = - \frac{2}{n-1} \sum_{t=2}^{n} X_{t-1} \epsilon_t,
\]
is the sum of martingale differences, since \( \mathbb{E}(X_{t-1} \epsilon_t | X_{t-1}) = X_{t-1} \mathbb{E}(\epsilon_t | X_{t-1}) = X_{t-1} \mathbb{E}(\epsilon_t) = 0 \) (here we used Definition 5.7.1). In order to show asymptotic of \( \nabla \mathcal{L}_n(\phi) \) we will use the martingale central limit theorem.

We now use Theorem 5.7.1 to show that \( \sqrt{n} \nabla \mathcal{L}_n(\phi) \) is asymptotically normal, which means we have to verify conditions (5.14)-(5.16). We note in our example that \( Y_t := X_{t-1} \epsilon_t \), and that the series \( \{X_{t-1} \epsilon_t\}_t \) is an ergodic process. Furthermore, since for any function \( g \), \( \mathbb{E}(g(X_{t-1} \epsilon_t) | \mathcal{F}_{t-1}) = \mathbb{E}(g(X_{t-1} \epsilon_t) | X_{t-1}) = \mathbb{E}(g(X_{t-1} \epsilon_t) | X_{t-1}) \) we need only to condition on \( X_{t-1} \) rather than the entire sigma-algebra \( \mathcal{F}_{t-1} \).
C1 : By using the ergodicity of \( \{X_{t-1} \xi_t\}_t \) we have
\[
\frac{1}{n} \sum_{t=1}^{n} Y_t^2 = \frac{1}{n} \sum_{t=1}^{n} X_{t-1}^2 \xi_t^2 \xrightarrow{P} E(X_{t-1}^2) E(\xi_t^2) = \sigma^2.
\]

C2 : We now verify the conditional Lindeberg condition.
\[
\frac{1}{n} \sum_{t=1}^{n} E(Y_t^2 I(\{|Y_t| > \varepsilon \sqrt{n}\}|\mathcal{F}_{t-1}) = \frac{1}{n} \sum_{t=1}^{n} E(X_{t-1}^2 \xi_t^2 I(\{|X_{t-1} \xi_t| > \varepsilon \sqrt{n}\}|X_{t-1})
\]

We now use the Cauchy-Schwartz inequality for conditional expectations to split \( X_{t-1}^2 \xi_t^2 \) and \( I(|X_{t-1} \xi_t| > \varepsilon) \). We recall that the Cauchy-Schwartz inequality for conditional expectations is \( E(X_t Y_t|\mathcal{G}) \leq \|E(X_t^2|\mathcal{G})E(Y_t^2|\mathcal{G})\|^{1/2} \) almost surely. Therefore
\[
\begin{align*}
\frac{1}{n} \sum_{t=1}^{n} E(Y_t^2 I(|Y_t| > \varepsilon \sqrt{n})|\mathcal{F}_{t-1}) & \leq \frac{1}{n} \sum_{t=1}^{n} \left\{ E(X_{t-1}^4 \xi_t^4 I_{X_{t-1}}) E(I(|X_{t-1} \xi_t| > \varepsilon \sqrt{n})^2|X_{t-1}) \right\}^{1/2} \\
& \leq \frac{1}{n} \sum_{t=1}^{n} X_{t-1}^2 \xi_t^2 \left\{ E(I(|X_{t-1} \xi_t| > \varepsilon \sqrt{n})^2|X_{t-1}) \right\}^{1/2}.
\end{align*}
\]

We note that rather than use the Cauchy-Schwartz inequality we can use a generalisation of it called the Hölder inequality. The Hölder inequality states that if \( p^{-1} + q^{-1} = 1 \), then \( E(XY) \leq \|E(X^p)\|^{1/p} \|E(Y^q)\|^{1/q} \) (the conditional version also exists). The advantage of using this inequality is that one can reduce the moment assumptions on \( X_t \).

Returning to (5.20), and studying \( E(I(|X_{t-1} \xi_t| > \varepsilon)^2|X_{t-1}) \) we use that \( E(I(A)) = P(A) \) and the Chebyshev inequality to show
\[
\begin{align*}
E(I(|X_{t-1} \xi_t| > \varepsilon \sqrt{n})^2|X_{t-1}) &= E(I(|X_{t-1} \xi_t| > \varepsilon \sqrt{n})|X_{t-1}) \\
&= \frac{P_e(|\xi_t| > \varepsilon \sqrt{n}/X_{t-1})}{X_{t-1}} \leq \frac{X_{t-1}^2 \text{var}(\xi_t)}{\varepsilon^2 n}.
\end{align*}
\]

Substituting (5.21) into (5.20) we have
\[
\begin{align*}
\frac{1}{n} \sum_{t=1}^{n} E(Y_t^2 I(|Y_t| > \varepsilon \sqrt{n})|\mathcal{F}_{t-1}) & \leq \frac{1}{n} \sum_{t=1}^{n} X_{t-1}^2 \xi_t^2 \left\{ \frac{X_{t-1}^2 \text{var}(\xi_t)}{\varepsilon^2 n} \right\}^{1/2} \\
& \leq \frac{E(\xi_t^4)^{1/2}}{\varepsilon n^{3/2}} \sum_{t=1}^{n} |X_{t-1}|^3 \frac{E(\xi_t^4)^{1/2}}{\varepsilon n^{1/2}} \\
& \leq \frac{E(\xi_t^4)^{1/2}}{\varepsilon n^{1/2}} \frac{1}{n} \sum_{t=1}^{n} |X_{t-1}|^3.
\end{align*}
\]
If \( \mathbb{E}(\xi^4) < \infty \), then \( \mathbb{E}(X_t^4) < \infty \), therefore by using the ergodic theorem we have \( \frac{1}{n} \sum_{t=1}^{n} |X_{t-1}|^3 \rightarrow^{a.s.} \mathbb{E}(|X_0|^3) \). Since almost sure convergence implies convergence in probability we have

\[
\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}(Y_t^2 I(|Y_t| > \varepsilon \sqrt{n})|\mathcal{F}_{t-1}) \leq \mathbb{E}(\xi^4)^{1/2} \frac{1}{\varepsilon n^{1/2}} \frac{1}{n} \sum_{t=1}^{n} |X_{t-1}|^3 \rightarrow^{P} \mathbb{E}(|X_0|^3)
\]

\( P \rightarrow 0 \).

Hence condition (5.15) is satisfied.

**C3**: We need to verify that

\[
\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}(Y_t^2|\mathcal{F}_{t-1}) \rightarrow^{P} \sigma^2.
\]

Since \( \{X_t\}_t \) is an ergodic sequence we have

\[
\frac{1}{n} \sum_{t=1}^{n} \mathbb{E}(Y_t^2|\mathcal{F}_{t-1}) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}(X_t^2|\mathcal{F}_{t-1}) = \frac{1}{n} \sum_{t=1}^{n} X_{t-1}^2 \mathbb{E}(\xi^2|\mathcal{F}_{t-1}) = \mathbb{E}(\xi^2) \frac{1}{n} \sum_{t=1}^{n} X_{t-1}^2 \rightarrow^{a.s.} \mathbb{E}(X_0^2)
\]

\( P \rightarrow \mathbb{E}(\xi^2)\mathbb{E}(X_0^2) = \sigma^2 \),

hence we have verified condition (5.16).

Altogether conditions C1-C3 imply that

\[
\sqrt{n} \nabla L_n(\phi) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_{t-1} \xi_t \rightarrow^{D} \mathcal{N}(0, \sigma^2).
\] (5.22)

Recalling (5.19) and that \( \sqrt{n} \nabla L_n(\phi) \rightarrow^{D} \mathcal{N}(0, \sigma^2) \) we have

\[
\sqrt{n}(\hat{\phi}_n - \phi) = - \sqrt{n} \nabla L_n(\phi)^{-1} \sqrt{n} \nabla L_n(\phi) \rightarrow^{a.s.} (2\mathbb{E}(X_0^2))^{-1} \mathcal{N}(0, \sigma^2)^{-1} \rightarrow^{D} \mathcal{N}(0, \sigma^2^{-1})
\] (5.23)

Using that \( \mathbb{E}(X_0^2) = \sigma^2 \), this implies that

\[
\sqrt{n}(\hat{\phi}_n - \phi) \rightarrow^{D} \mathcal{N}(0, \frac{1}{\sigma^2}).
\] (5.24)

Thus we have derived the limiting distribution of \( \hat{\phi}_n \).
Remark 5.8.1 We recall that

\[ \hat{\phi}_n - \phi = - (\nabla^2 L_n)^{-1} \nabla L_n(\phi) = \frac{-2}{n-1} \sum_{t=2}^{n} \varepsilon_t X_{t-1}, \]  

(5.25)

and that \( \text{var}(\frac{-2}{n-1} \sum_{t=2}^{n} \varepsilon_t X_{t-1}) = \frac{-2}{n-1} \sum_{t=2}^{n} \text{var}(\varepsilon_t X_{t-1}) = O(\frac{1}{n}) \). This implies

\[ (\hat{\phi}_n - \phi) = O_p(n^{-1/2}). \]

Indeed the results also holds almost surely

\[ (\hat{\phi}_n - \phi) = O(n^{-1/2}). \]  

(5.26)

The same result is true for autoregressive processes of arbitrary finite order. That is

\[ \sqrt{n}(\hat{\phi}_n - \phi) \xrightarrow{D} N(0, 2E(\Gamma_{p}^{-1})). \]  

(5.27)