

Chapter 13

Large Samples and Dependence

With IID data, the joint distribution of the sample is pinned down by the common marginal of the observations. The statistical learning problem becomes one of learning the marginal. With dependent data the joint distribution is more complex. The fundamental statistical problem from page 213 has to be extended as follows:

Problem 13.0.1 We observe a draw $\mathbf{z}_{\mathcal{D}}$ from some joint distribution $P_{\mathcal{D}}$, which is assumed to belong to a given class of distributions $\mathcal{P}$ but is otherwise unknown. We wish to infer some features of $P_{\mathcal{D}}$ from the observation $\mathbf{z}_{\mathcal{D}}$.

Let's look at how we might tackle this problem, starting with least squares.

13.1 Large Sample Least Squares

Our treatment of least squares in this chapter permits dependent data but at the same time requires that the sample size is large. Large samples allow us to drop some of the strict parametric assumptions on the error term we made when we considered finite sample inference (see, e.g., assumption 12.1.5 on page 332). In this sense, the theory developed below is also useful for cross-sectional environments with no correlation between observations.

13.1.1 Setup and Assumptions

Let's go back to the linear model assumption of chapter 12. That is, we assume that our data $(y_1, \mathbf{x}_1), \dots, (y_T, \mathbf{x}_T)$ are generated by the linear model

$$y_t = \mathbf{x}_t^T \boldsymbol{\beta} + u_t, \quad t = 1, \dots, T, \quad (13.1)$$

where β is a K -vector of unknown coefficients, and u_t is an unobservable shock. Observations are indexed by t rather than n to remind us that observations are dependent. The sample size will be denoted by T .

We let $\mathbf{y}$ be the $T \times 1$ vector of observed outputs, so that y_t is the t th element of $\mathbf{y}$, and $\mathbf{u}$ be the vector of shocks, so that u_t is the t th element of $\mathbf{u}$. We let $\mathbf{X}$ be the $T \times K$ matrix $\mathbf{X} := (x_{tk})$, where $1 \leq t \leq T$ and $1 \leq k \leq K$.

We will estimate the parameter vector β via least squares. The OLS estimate is therefore unchanged at

$$\hat{\beta}_T = \left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \right]^{-1} \cdot \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t y_t$$

(Recall (11.17) on page 306.) The expression for the sampling error in (12.4) can be expanded into sums to obtain

$$\hat{\beta}_T - \beta = \left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \right]^{-1} \cdot \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t u_t \quad (13.2)$$

In this section we drop the exogeneity assumption $\mathbb{E}[\mathbf{u} | \mathbf{X}] = \mathbf{0}$ (assumption 12.1.3) because it excludes too many dynamic models. For example, we know from §12.2.2 that exogeneity fails when we try to estimate the simple AR(1) model $y_{t+1} = \beta y_t + u_{t+1}$ by setting $x_t = y_{t-1}$, thereby producing the regression model

$$y_t = \beta x_t + u_t, \quad t = 1, \dots, T \quad (13.3)$$

For this specification of (13.1), the regressor is correlated with lagged values of the shock.

Assumption 13.1.1 The matrix $\mathbf{X}$ is full column rank with probability one and the sequence $\{\mathbf{x}_t\}$ is stationary. Moreover

- (i) $\Sigma_{\mathbf{x}} := \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top]$ exists and is positive definite, and
- (ii) the sequence $\{\mathbf{x}_t\}$ satisfies $\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \xrightarrow{P} \Sigma_{\mathbf{x}}$ as $T \rightarrow \infty$.

The full column rank condition is the same as assumption 11.1.1 on page 302 and ensures that expressions like the matrix inverse in (13.2) make sense. Condition (i) is borrowed from the identical condition given in (11.10) on page 302 and ensures that the limit in (13.2) exists. Condition (ii) gives convergence.

Example 13.1.1 Let $\{x_t\}$ be the Markov process in example 7.2.11 on page 195. To

repeat,

$$x_{t+1} = a|x_t| + (1 - a^2)^{1/2}w_{t+1} \quad \text{with} \quad -1 < a < 1 \quad \text{and} \quad \{w_t\} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

As discussed in example 7.2.11, the model has a unique, globally stable stationary distribution π_∞ . If $\mathcal{L}(x_0) = \pi_\infty$, then the process $\{x_t\}$ is stationary (see fact 7.2.5 on page 191) and all of the conditions in assumption 13.1.1 are satisfied (see ex. 13.4.3).

Assumption 13.1.2 (Weak exogeneity) The shocks $\{u_t\}$ are IID. Moreover

- (i) $\mathbb{E}[u_t] = 0$ and $\mathbb{E}[u_t^2] = \sigma^2$ for all t , and
- (ii) u_t is independent of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t$ for all t .

Assumption 13.1.2 permits dependence between current shocks and future regressors. This is desirable in a time series setting, since current shocks usually feed into future state variables.

Example 13.1.2 In the AR(1) regression (13.3), assumption 13.1.2 holds whenever the shock process $\{u_t\}$ is IID, because the contemporaneous and lagged regressors $x_1, \dots, x_t$ are equal to the lagged state variables $y_0, \dots, y_{t-1}$, which in turn are functions of only y_0 and $u_1, \dots, u_{t-1}$, and therefore independent of u_t .

One consequence of assumption 13.1.2 is that

$$\mathbb{E}[u_s u_t | \mathbf{x}_1, \dots, \mathbf{x}_t] = \begin{cases} \sigma^2 & \text{if } s = t \\ 0 & \text{if } s < t \end{cases} \quad (13.4)$$

The proof is an exercise (ex. 13.4.4).

A useful implication of assumptions 13.1.1 and 13.1.2 is that linear functions of $\{\mathbf{x}_t u_t\}$ form a martingale difference sequence (MDS). This will allow us to apply the sample path results in theorem 7.3.1.

Lemma 13.1.1 If assumptions 13.1.1 and 13.1.2 both hold, then, for any constant vector $\mathbf{a} \in \mathbb{R}^K$, the sequence $\{m_t\}$ defined by $m_t = \mathbf{a}^\top \mathbf{x}_t u_t$ is

- (i) stationary with $\mathbb{E}[m_t^2] = \sigma^2 \mathbf{a}^\top \boldsymbol{\Sigma}_x \mathbf{a}$ for all t , and
- (ii) an MDS with respect to the filtration defined by

$$\mathcal{F}_t := \{\mathbf{x}_1, \dots, \mathbf{x}_t, \mathbf{x}_{t+1}, u_1, \dots, u_t\} \quad (13.5)$$

Proof. First let's check part (i). That $\{m_t\}$ is stationary follows from the assumption that $\{u_t\}$ and $\{\mathbf{x}_t\}$ are stationary. Regarding the second moment $\mathbb{E}[m_1^2]$, we have

$$\mathbb{E}[m_1^2] = \mathbb{E}[\mathbb{E}[u_1^2 (\mathbf{a}^\top \mathbf{x}_1)^2 | \mathbf{x}_1]] = \mathbb{E}[(\mathbf{a}^\top \mathbf{x}_1)^2 \mathbb{E}[u_1^2 | \mathbf{x}_1]]$$

From independence of u_1 and $\mathbf{x}_1$, the inner expectation is σ^2 . Moreover

$$\begin{aligned} (\mathbf{a}^\top \mathbf{x}_1)^2 &= \mathbf{a}^\top \mathbf{x}_1 \mathbf{a}^\top \mathbf{x}_1 = \mathbf{a}^\top \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{a} \\ \therefore \mathbb{E}[m_1^2] &= \mathbb{E}[\mathbf{a}^\top \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{a} \sigma^2] = \sigma^2 \mathbf{a}^\top \mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^\top] \mathbf{a} = \sigma^2 \mathbf{a}^\top \boldsymbol{\Sigma}_x \mathbf{a} \end{aligned}$$

To check part (ii), note that $\{m_t\}$ is adapted to $\{\mathcal{F}_t\}$, since $m_t := u_t \mathbf{a}^\top \mathbf{x}_t$ is a function of variables in $\mathcal{F}_t$. Moreover we have

$$\mathbb{E}[m_{t+1} | \mathcal{F}_t] = \mathbb{E}[u_{t+1} \mathbf{a}^\top \mathbf{x}_{t+1} | \mathcal{F}_t] = \mathbf{a}^\top \mathbf{x}_{t+1} \mathbb{E}[u_{t+1} | \mathcal{F}_t] = \mathbf{a}^\top \mathbf{x}_{t+1} \mathbb{E}[u_{t+1}] = 0$$

(Here the second equality follows from the fact that $\mathbf{x}_{t+1} \in \mathcal{F}_t$, while the third follows from the independence in assumption 13.1.2.) This confirms that $\{m_t\}$ is an MDS with respect to $\{\mathcal{F}_t\}$. $\square$

13.1.2 Consistency

Under the conditions of §13.1.1, the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is consistent for $\boldsymbol{\beta}$:

Theorem 13.1.1 *If assumptions 13.1.1 and 13.1.2 hold, then*

$$\hat{\boldsymbol{\beta}}_T \xrightarrow{p} \boldsymbol{\beta} \quad \text{as } T \rightarrow \infty$$

Proof. It suffices to show that the expression on the right-hand side of (13.2) converges in probability to $\mathbf{0}$. As a first step, let's show that $\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t u_t \xrightarrow{p} \mathbf{0}$. In view of fact 6.1.1 on page 161, it suffices to show that, for any $\mathbf{a} \in \mathbb{R}^K$, we have

$$\mathbf{a}^\top \left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t u_t \right] \xrightarrow{p} \mathbf{a}^\top \mathbf{0} = 0 \quad (13.6)$$

If we define $m_t := \mathbf{a}^\top \mathbf{x}_t u_t$, then the left-hand side of (13.6) can be written as $T^{-1} \sum_{t=1}^T m_t$. Since $\{m_t\}$ is a stationary MDS (lemma 13.1.1), the convergence $T^{-1} \sum_{t=1}^T m_t \xrightarrow{p} 0$ follows from theorem 7.3.1 (page 200).

Now let's return to the expression on the right-hand side of (13.2). By assumption 13.1.1 and fact 6.2.1, we see that

$$\left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \right]^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_x^{-1} \quad \text{as } T \rightarrow \infty \quad (13.7)$$

Appealing to fact 6.2.1 once more, we obtain

$$\hat{\beta}_T - \beta = \left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \right]^{-1} \cdot \frac{1}{T} \sum_{t=1}^T u_t \mathbf{x}_t \xrightarrow{p} \Sigma_{\mathbf{x}}^{-1} \mathbf{0} = \mathbf{0} \quad \square$$

13.1.2.1 Consistency of $\hat{\sigma}_T^2$

In (12.5) we estimated the error variance σ^2 using $\hat{\sigma}^2 := \text{RSS}/(N - K)$, with N being the sample size. Here the sample size is T , and since T is assumed to be large relative to K , we have $1/(T - K) \approx 1/T$. Hence in this chapter we set $\hat{\sigma}_T^2 = \text{RSS}/T$ instead. (None of the following theory is affected if we use $\text{RSS}/(T - K)$, however.)

Theorem 13.1.2 *If assumptions 13.1.1 and 13.1.2 hold, then*

$$\hat{\sigma}_T^2 \xrightarrow{p} \sigma^2 \quad \text{as } T \rightarrow \infty$$

Proof. By the definition of $\hat{\sigma}_T^2$ and the linear model assumption (13.1),

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_{t=1}^T (y_t - \mathbf{x}_t^\top \hat{\beta}_T)^2 = \frac{1}{T} \sum_{t=1}^T \left[u_t + \mathbf{x}_t^\top (\beta - \hat{\beta}_T) \right]^2 \quad (13.8)$$

Expanding out the square gives

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_{t=1}^T u_t^2 + 2(\beta - \hat{\beta}_T)^\top \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t u_t + (\beta - \hat{\beta}_T)^\top \left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \right] (\beta - \hat{\beta}_T)$$

By assumption 13.1.2 and the law of large numbers, the first term on the right-hand side converges in probability to σ^2 . Hence it suffices to show that the second and third term converge in probability to zero as $T \rightarrow \infty$. These results follow from repeated applications of fact 6.2.1 on page 170 and convergence results we have already established. The details are left as an exercise. $\square$

13.1.3 Asymptotic Normality of $\hat{\beta}$

Next we show that $\sqrt{T}(\hat{\beta}_T - \beta)$ is asymptotically normal. From this information we can develop asymptotic tests and confidence intervals.

Theorem 13.1.3 *If assumptions 13.1.1 and 13.1.2 hold, then*

$$\sqrt{T}(\hat{\beta}_T - \beta) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2 \Sigma_{\mathbf{x}}^{-1}\right) \quad \text{as } T \rightarrow \infty$$

Proof. Expression (13.2) gives

$$\sqrt{T}(\hat{\beta}_T - \beta) = \left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \right]^{-1} \cdot T^{-1/2} \sum_{t=1}^T u_t \mathbf{x}_t \quad (13.9)$$

Let $\mathbf{z}$ be a random variable satisfying $\mathcal{L}(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma_{\mathbf{x}})$. Suppose we can show that

$$T^{-1/2} \sum_{t=1}^T u_t \mathbf{x}_t \xrightarrow{d} \mathbf{z} \quad \text{as } T \rightarrow \infty \quad (13.10)$$

If (13.10) is valid, then, applying assumption 13.1.1 along with fact 6.2.2, we obtain

$$\sqrt{T}(\hat{\beta}_T - \beta) = \left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \right]^{-1} \cdot T^{-1/2} \sum_{t=1}^T u_t \mathbf{x}_t \xrightarrow{d} \Sigma_{\mathbf{x}}^{-1} \mathbf{z}$$

Clearly $\Sigma_{\mathbf{x}}^{-1} \mathbf{z}$ is Gaussian with zero mean. By symmetry of $\Sigma_{\mathbf{x}}^{-1}$ (since $\Sigma_{\mathbf{x}}$ is symmetric) the variance of $\Sigma_{\mathbf{x}}^{-1} \mathbf{z}$ is

$$\Sigma_{\mathbf{x}}^{-1} \text{var}[\mathbf{z}] \Sigma_{\mathbf{x}}^{-1} = \Sigma_{\mathbf{x}}^{-1} \sigma^2 \Sigma_{\mathbf{x}} \Sigma_{\mathbf{x}}^{-1} = \sigma^2 \Sigma_{\mathbf{x}}^{-1}$$

This completes the proof of theorem 13.1.3, conditional on (13.10). Let's now check that (13.10) is valid.

By the Cramér–Wold device (fact 6.1.6 on page 167), it suffices to show that for any $\mathbf{a} \in \mathbb{R}^K$ we have

$$\mathbf{a}^\top \left[T^{-1/2} \sum_{t=1}^T u_t \mathbf{x}_t \right] \xrightarrow{d} \mathbf{a}^\top \mathbf{z} \quad (13.11)$$

Fixing $\mathbf{a}$ and letting $m_t := u_t \mathbf{a}^\top \mathbf{x}_t$, the expression on the left of (13.11) can be rewritten as $T^{-1/2} \sum_{t=1}^T m_t$. Since $\mathcal{L}(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma^2 \Sigma_{\mathbf{x}})$, to establish (13.11) we need to show that

$$T^{-1/2} \sum_{t=1}^T m_t \xrightarrow{d} \mathcal{N}(0, \sigma^2 \mathbf{a}^\top \Sigma_{\mathbf{x}} \mathbf{a}) \quad (13.12)$$

From lemma 13.1.1 we know that $\{m_t\}$ is stationary with $\mathbb{E}[m_t^2] = \sigma^2 \mathbf{a}^\top \Sigma_{\mathbf{x}} \mathbf{a}$ and an MDS with respect to the filtration given in (13.5). By the martingale difference CLT on page 200, the result in (13.12) holds whenever

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[m_t^2 \mid \mathcal{F}_{t-1}] \xrightarrow{p} \sigma^2 \mathbf{a}^\top \Sigma_{\mathbf{x}} \mathbf{a} \quad \text{as } T \rightarrow \infty \quad (13.13)$$

Since $\mathbf{x}_t \in \mathcal{F}_{t-1}$, we have

$$\mathbb{E}[m_t^2 | \mathcal{F}_{t-1}] = \mathbb{E}[u_t^2 (\mathbf{a}^\top \mathbf{x}_t)^2 | \mathcal{F}_{t-1}] = (\mathbf{a}^\top \mathbf{x}_t)^2 \mathbb{E}[u_t^2 | \mathcal{F}_{t-1}] = \sigma^2 (\mathbf{a}^\top \mathbf{x}_t)^2$$

Another way to write the last expression is $\sigma^2 \mathbf{a}^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{a}$. The left-hand side of (13.13) is therefore

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[m_t^2 | \mathcal{F}_{t-1}] = \frac{1}{T} \sum_{t=1}^T (\sigma^2 \mathbf{a}^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{a}) = \sigma^2 \mathbf{a}^\top \left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top \right] \mathbf{a}$$

which converges in probability to $\sigma^2 \mathbf{a}^\top \Sigma_{\mathbf{x}} \mathbf{a}$ by assumption 13.1.1 and fact 6.2.1. This verifies (13.13), completing the proof of theorem 13.1.3. $\square$

Example 13.1.3 Consider again the scalar linear Gaussian AR(1) model $x_{t+1} = ax_t + w_{t+1}$ with $|a| < 1$ and $\{w_t\}$ IID and standard normal. Let $\{x_t\}$ be stationary. As discussed in §12.2.2, the OLS estimator of a is

$$\hat{a}_T := \frac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^\top \mathbf{x}} \quad \text{where} \quad \mathbf{y} := (x_1, \dots, x_T) \text{ and } \mathbf{x} := (x_0, \dots, x_{T-1})$$

Both assumption 13.1.1 and assumption 13.1.2 are satisfied, so $\sqrt{T}(\hat{a}_T - a)$ converges in distribution to $N(0, \sigma^2 \Sigma_{\mathbf{x}}^{-1})$. In this case, $\sigma^2 = 1$ because the shocks are standard normal. Furthermore $\Sigma_{\mathbf{x}}^{-1}$ reduces to $1/\mathbb{E}[x_1^2]$, where the expectation is under the stationary distribution. By (7.18) on page 190, we see that the stationary distribution is $N(0, 1/(1 - a^2))$. Hence the inverse of $\mathbb{E}[x_1^2]$ is $1 - a^2$, and

$$\sqrt{T}(\hat{a}_T - a) \xrightarrow{d} N(0, 1 - a^2) \quad (13.14)$$

13.1.4 Large Sample Tests

In §12.1.4.1 we considered the problem of testing a hypothesis about an individual coefficient β_k . Let's consider this problem again in the large sample setting. The hypothesis to be tested is

$$H_0: \beta_k = \beta_k^0$$

In §12.1.4.1 we showed that if the error terms are normally distributed, then the expression $(\hat{\beta}_k - \beta_k) / \text{se}(\hat{\beta}_k)$ is t -distributed with $N - K$ degrees of freedom. In the large sample case, we can use the CLT to show that the same statistic is asymptotically normal.¹ The next theorem gives details.

1. This isn't surprising because the t -distribution converges to the standard normal distribution as the degrees of freedom converges to infinity. However, we can't use this result directly because our model assumptions are different.

Theorem 13.1.4 Let assumptions 13.1.1 and 13.1.2 hold, and let

$$\text{se}(\hat{\beta}_k^T) := \sqrt{\hat{\sigma}_T^2 v_k(\mathbf{X})}$$

Under the null hypothesis H_0 , we have

$$z_k^T := \frac{\hat{\beta}_k^T - \beta_k^0}{\text{se}(\hat{\beta}_k^T)} \xrightarrow{d} N(0, 1) \quad \text{as } T \rightarrow \infty \quad (13.15)$$

In the theorem, $\hat{\sigma}_T$ is as defined in (13.8) on page 359 and $v_k(\mathbf{X})$ is the (k, k) th element of $(\mathbf{X}^T \mathbf{X})^{-1}$.

Proof. Recall from theorem 13.1.3 that $\sqrt{T}(\hat{\beta}_T - \beta) \xrightarrow{d} \mathbf{z}$, where $\mathbf{z}$ is a random vector with distribution $N(\mathbf{0}, \sigma^2 \Sigma_{\mathbf{x}}^{-1})$ and β is the true parameter vector. It follows that

$$\sqrt{T}(\hat{\beta}_k^T - \beta_k) = \mathbf{e}_k^T [\sqrt{T}(\hat{\beta}_T - \beta)] \xrightarrow{d} \mathbf{e}_k^T \mathbf{z}$$

The distribution of $\mathbf{e}_k^T \mathbf{z}$ is $N(0, \mathbf{e}_k^T \text{var}[\mathbf{z}] \mathbf{e}_k) = N(0, \sigma^2 \mathbf{e}_k^T \Sigma_{\mathbf{x}}^{-1} \mathbf{e}_k)$, so

$$\frac{\sqrt{T}(\hat{\beta}_k^T - \beta_k)}{\sqrt{\sigma^2 \mathbf{e}_k^T \Sigma_{\mathbf{x}}^{-1} \mathbf{e}_k}} \xrightarrow{d} N(0, 1) \quad (13.16)$$

Applying (13.7) and (v) of fact 6.2.1 on page 170,

$$T v_k(\mathbf{X}) = T \mathbf{e}_k^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{e}_k = \mathbf{e}_k^T \left[\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^T \right]^{-1} \mathbf{e}_k \xrightarrow{p} \mathbf{e}_k^T \Sigma_{\mathbf{x}}^{-1} \mathbf{e}_k$$

By theorem 13.1.2 we have $\hat{\sigma}_T^2 \xrightarrow{p} \sigma^2$, and hence

$$\sqrt{\hat{\sigma}_T^2 T v_k(\mathbf{X})} \xrightarrow{p} \sqrt{\sigma^2 \mathbf{e}_k^T \Sigma_{\mathbf{x}}^{-1} \mathbf{e}_k}$$

Combining this with (13.16) yields

$$\frac{\sqrt{T}(\hat{\beta}_k^T - \beta_k)}{\sqrt{\hat{\sigma}_T^2 T v_k(\mathbf{X})}} \xrightarrow{d} N(0, 1)$$

Assuming H_0 and canceling $\sqrt{T}$ gives (13.15). □

13.2 MLE for Markov Processes

Let's now turn to nonlinear estimation in a time series setting, using maximum likelihood. We will cover just the foundations of building and maximizing the likelihood.

13.2.1 The Likelihood Function

If we have IID observations from some common density, then, as we saw in §8.3.1, the likelihood function for that data is the product of the marginals evaluated at the sample. If, instead, the data contains dependence, the likelihood is more complex.

One intermediate case is Markov processes, the joint distribution of which was given in (7.17) on page 189. Suppose that the transition density p_θ depends on some unknown parameter vector $\theta \in \Theta$. Suppose further that the process has a unique stationary density π_∞^θ for all θ , and that $\mathbf{x}_1$ is a draw from this stationary density. Then the log-likelihood function is given by

$$\ell(\theta) = \ln \pi_\infty^\theta(\mathbf{x}_1) + \sum_{t=1}^{T-1} \ln p_\theta(\mathbf{x}_{t+1} | \mathbf{x}_t)$$

In practice, it is usual to drop the first term in this expression, particularly when the data size is large, since the influence of a single element is likely to be negligible. In addition, even though the stationary density is formally defined by (7.19), for many processes, there is no known analytical expression for this density.² Here we'll follow this convention and, abusing notation slightly, write

$$\ell(\theta) = \sum_{t=1}^{T-1} \ln p_\theta(\mathbf{x}_{t+1} | \mathbf{x}_t) \quad (13.17)$$

13.2.1.1 Example: The ARCH Case

Let's look at a relatively simple case: the ARCH model from (7.5) on page 183. By (13.17), the log-likelihood function is

$$\ell(a, b) = \sum_{t=1}^{T-1} \left\{ -\frac{1}{2} \ln(2\pi(a + bx_t^2)) - \frac{x_{t+1}^2}{2(a + bx_t^2)} \right\} \quad (13.18)$$

2. It is still possible to compute the density numerically. See, for example, Stachurski and Martin (2008).

Rearranging, dropping terms that don't depend on a or b , and multiplying by 2 (an increasing transformation), we can rewrite this (abusing notation again) as

$$\ell(a, b) = - \sum_{t=1}^{T-1} \left\{ \ln z_t + \frac{x_{t+1}^2}{z_t} \right\} \quad \text{where} \quad z_t := a + bx_t^2 \quad (13.19)$$

Let's run some simulations to see what this function looks like. In the simulations we will set $T = 500$ and $a = b = 0.5$. Thus, we imagine the situation where, unbeknown to us, the true parameter values are $a = b = 0.5$, and we observe a time series $x_1, \dots, x_{500}$ generated by these parameters. In order to estimate a and b , we form the likelihood function (13.19), and obtain the MLEs $\hat{a}$ and $\hat{b}$ as the vector $(\hat{a}, \hat{b})$ that maximizes $\ell(a, b)$.

Four different simulations of ℓ are given in figure 13.1. In each figure, a separate data set $x_1, \dots, x_{500}$ is generated using the true parameter values $a = b = 0.5$ and the function ℓ in (13.19) is then plotted using contour lines and a color map. Lighter colors refer to larger values. The horizontal axis is a values, and the vertical axis is b values. The code for producing one of these figures (modulo randomness) is given in listing 11. The function `arch_like(theta, data)` represents ℓ in (13.19), with `theta` corresponding to (a, b) and `data` corresponding to the time series $x_1, \dots, x_T$.

In each of the four simulations, a rough guess of the MLEs can be obtained just by looking for maximizers in the figures. For example, in simulation (a), the MLEs look to be around $\hat{a} = 0.44$ and $\hat{b} = 0.61$. To get more accurate estimates, we can use some form of analytical or numerical optimization. For this problem we don't have any analytical expressions for the MLEs because setting the two partial derivatives of ℓ in (13.19) to zero does not yield neat expressions for $\hat{a}$ and $\hat{b}$. However, there are many numerical routines we can use to obtain the MLEs for a given data set.

The simplest approach is to use one of R's inbuilt optimization routines. For example, given the definition of `arch_like` in listing 11 and a sequence of observations $x_1, \dots, x_T$ stored in a vector `xdata`, the function `arch_like` can be optimized numerically via the commands

```
start_theta <- c(0.65, 0.35) # An initial guess of (a,b)
neg_like <- function(theta) {
  return(-arch_like(theta, xdata)) # xdata is the data
}
opt <- optim(start_theta, neg_like, method="BFGS")
```

Here `optim` is an built-in R function for numerical optimization of multivariate functions. Most built-in functions perform minimization rather than maximization and `optim` is no exception. For this reason the function that we pass to `optim` is `neg_like`, which is -1 times ℓ . The first argument to `optim` is a vector of starting values (a guess

of the MLEs). The last argument tells `optim` to use the BFGS routine, which is variation on the Newton–Raphson algorithm. The return value of `optim` is a list, and the approximate minimizing vector is one element of this list (called `par`).

In this particular setup, for most realizations of the data and starting values, you will find that the algorithm converges to a good approximation to the global optimizer. However, there’s no guarantee that it will. In case of problems, it’s useful to know how these kinds of algorithms work, and how to code up simple implementations on your own. The next section will get you started.

13.2.2 The Newton–Raphson Algorithm

The Newton–Raphson algorithm is a *root-finding* algorithm. In other words, given a function $g: \mathbb{R} \rightarrow \mathbb{R}$, the algorithm searches for points $\bar{s} \in \mathbb{R}$ such that $g(\bar{s}) = 0$. Any root-finding algorithm can be used to optimize differentiable functions because, for differentiable functions, interior optimizers are always roots of the objective function’s first derivative.

To describe the algorithm, let’s begin with the root-finding problem and then specialize to optimization. To begin, let $g: \mathbb{R} \rightarrow \mathbb{R}$, and let s_0 be some initial point in $\mathbb{R}$ that we think (hope) is somewhere near a root. We don’t know how to jump from s_0 straight to a root of g (otherwise, there would be no problem to solve), but what we can do is move to the root of the function that forms the *tangent line* to g at s_0 . In other words, we replace g with its linear approximation around s_0 , which is given by

$$\tilde{g}(s) := g(s_0) + g'(s_0)(s - s_0) \quad (s \in \mathbb{R})$$

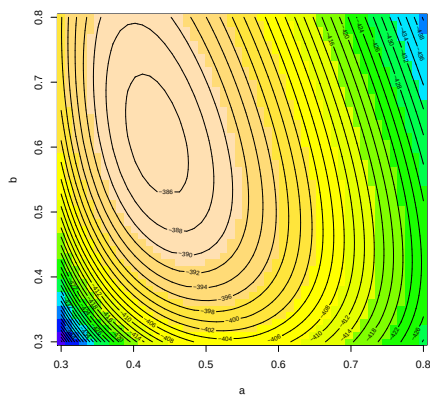
and solve for the root of $\tilde{g}$. This point is represented as s_1 in figure 13.2, and the value is easily seen to be $s_1 := s_0 - g(s_0)/g'(s_0)$. The point s_1 is taken as our next guess of the root, and the procedure is repeated, taking the tangent of g at s_1 , solving for the root, and so on. This generates a sequence of points $\{s_k\}$ satisfying

$$s_{k+1} = s_k - \frac{g(s_k)}{g'(s_k)}$$

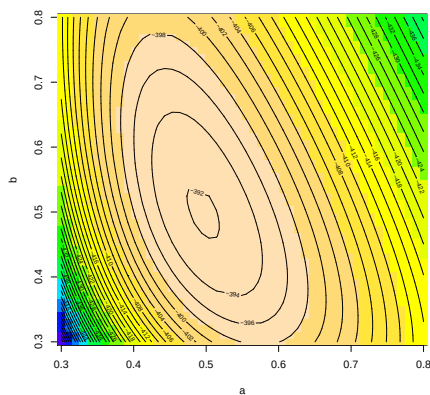
There are various results telling us that when g is suitably well-behaved and s_0 is sufficiently close to a given root $\bar{s}$, then sequence $\{s_k\}$ will converge to $\bar{s}$.³

To move from general root-finding to the specific problem of optimization, suppose now that $g: \mathbb{R} \rightarrow \mathbb{R}$ is a twice differentiable function we wish to maximize. We know that if s^* is a maximizer of g , then $g'(s^*) = 0$. Hence it is natural to begin

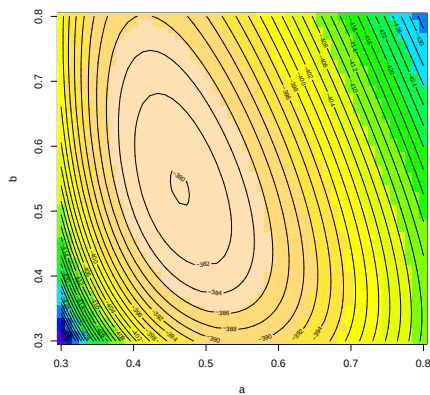
3. In practical situations we often have no way of knowing whether the conditions are satisfied, and there have been many attempts to make the procedure more robust. The R optimization routine described above is a child of this process.



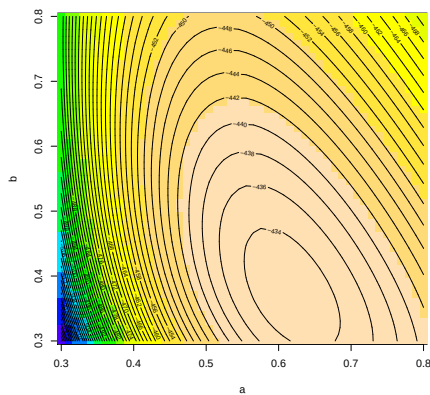
(a) Simulation 1



(b) Simulation 2



(c) Simulation 3



(d) Simulation 4

Figure 13.1 Simulations of the function (13.19) with $T = 500$

Listing 11 Code for figure 13.1 (R)

```

arch_like <- function(theta, data) {
  Y <- data[-1]          # All but first element
  X <- data[-length(data)] # All but last element
  Z <- theta[1] + theta[2] * X^2
  return(-sum(log(Z) + Y^2 / Z))
}

sim_data <- function(a, b, n=500) {
  x <- numeric(n)
  x[1] = 0
  w = rnorm(n)
  for (t in 1:(n-1)) {
    x[t+1] = sqrt(a + b * x[t]^2) * w[t]
  }
  return(x)
}

xdata <- sim_data(0.5, 0.5) # True parameters

K <- 50
a <- seq(0.3, 0.8, length=K)
b <- seq(0.3, 0.8, length=K)
M <- matrix(nrow=K, ncol=K)
for (i in 1:K) {
  for (j in 1:K) {
    theta <- c(a[i], b[j])
    M[i,j] <- arch_like(theta, xdata)
  }
}
image(a, b, M, col=topo.colors(12))
contour(a, b, M, nlevels=40, add=T)

```

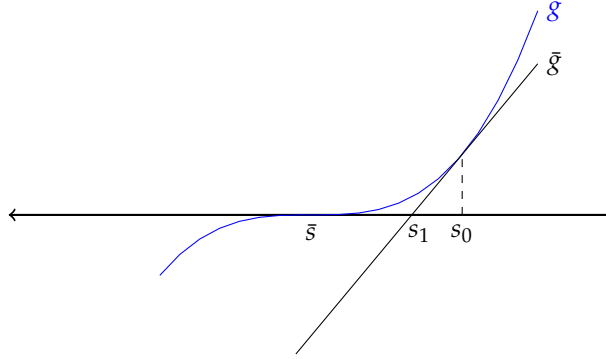


Figure 13.2 First step of the Newton–Raphson algorithm

our search for maximizers by looking for roots to this equation. This can be done by applying the Newton–Raphson algorithm to g' , which yields the sequence

$$s_{k+1} = s_k - \frac{g'(s_k)}{g''(s_k)} \quad (13.20)$$

We can extend this algorithm to the multivariate case as well. Let's suppose that g is a function of two arguments. In particular, suppose that g is twice differentiable and $g: \mathbb{R}^2 \rightarrow \mathbb{R}$. The **gradient vector** and **Hessian** of g at $(x, y) \in \mathbb{R}^2$ are defined as

$$\nabla g(x, y) := \begin{pmatrix} g'_1(x, y) \\ g'_2(x, y) \end{pmatrix} \quad \text{and} \quad \nabla^2 g(x, y) := \begin{pmatrix} g''_{11}(x, y) & g''_{12}(x, y) \\ g''_{21}(x, y) & g''_{22}(x, y) \end{pmatrix}$$

Here g'_i is the first partial of g with respect to its i th argument, g''_{ij} is the second cross-partial, and so on.

By analogy with (13.20), the Newton–Raphson algorithm for this two dimensional case is the algorithm that generates the sequence $\{(x_k, y_k)\}$ defined by

$$(x_{k+1}, y_{k+1}) = (x_k, y_k) - [\nabla^2 g(x_k, y_k)]^{-1} \nabla g(x_k, y_k) \quad (13.21)$$

from some initial guess (x_0, y_0) .⁴

For the sake of the exercise, let's apply this to maximization of (13.19). Let z_t be as

4. We are assuming that the Hessian matrix is nonsingular.

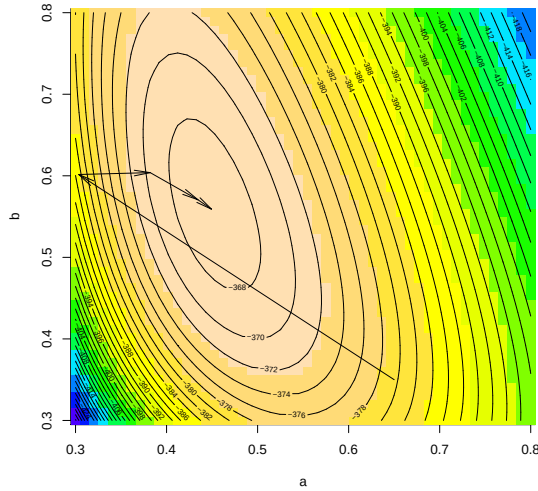


Figure 13.3 Newton–Raphson iterates

defined in (13.19). The first partials are then

$$\frac{\partial \ell}{\partial a}(a, b) = \sum_{t=1}^{T-1} \left[\frac{x_{t+1}^2}{z_t^2} - \frac{1}{z_t} \right], \quad \frac{\partial \ell}{\partial b}(a, b) = \sum_{t=1}^{T-1} x_t^2 \left[\frac{x_{t+1}^2}{z_t^2} - \frac{1}{z_t} \right]$$

while the second partials are

$$\frac{\partial^2 \ell}{\partial a^2}(a, b) = \sum_{t=1}^{T-1} \left[\frac{1}{z_t^2} - 2 \frac{x_{t+1}^2}{z_t^3} \right], \quad \frac{\partial^2 \ell}{\partial b^2}(a, b) = \sum_{t=1}^{T-1} x_t^4 \left[\frac{1}{z_t^2} - 2 \frac{x_{t+1}^2}{z_t^3} \right]$$

The cross-partial is

$$\frac{\partial^2 \ell}{\partial a \partial b}(a, b) = \sum_{t=1}^{T-1} x_t^2 \left[\frac{1}{z_t^2} - 2 \frac{x_{t+1}^2}{z_t^3} \right]$$

From these expressions we can form the gradient vector and the Hessian, pick an initial guess, and iterate according to (13.21). Figure 13.3 shows four iterations of this procedure, starting from $(a_0, b_0) = (0.65, 0.35)$.⁵ In this case the convergence is quick, and we are already close to the global optimum.

Replication of this figure (modulo randomness) is left as an exercise.

5. As before, the simulation uses $a = b = 0.5$ and $T = 500$.

13.3 Further Reading

Suitable topics for study after reading this chapter include GARCH models, hidden Markov models, Markov switching models, stochastic volatility models, factor models, and cointegration. High-quality textbook treatments of these and other topics in time series econometrics include Martin et al. (2012) and Hamilton (1994). Useful chapters on time series econometrics can be found in Hansen (2015) and Hayashi (2000). For a recent survey, see Diebold et al. (2009). For financial applications, see, for example, Ruppert (2011).

13.4 Exercises

Ex. 13.4.1 Let $\{x_t\}$ be the scalar AR(1) process $x_{t+1} = ax_t + w_{t+1}$, where $|a| < 1$ and $\{w_t\}$ is IID and standard normal. Let x_0 be drawn from the stationary distribution $\pi_\infty = N(0, 1/(1 - a^2))$. Show that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T x_t \xrightarrow{d} N(0, v) \quad \text{where} \quad v := \frac{1}{(1 - a)^2} \quad (13.22)$$

(Note how v increases in a , meaning that greater correlation makes it harder to estimate the mean of x_t via the sample mean. The intuition is that dependent data are less informative than independent data.)

Ex. 13.4.2 Recall that the sample variance is a consistent and asymptotically normal estimator of the variance of a distribution when the data are IID. A proof was given in example 9.2.11 on page 264. Now assume the conditions of exercise 13.4.1. While the data sequence $\{x_t\}$ is no longer IID, it is a stable Markov process. Using the proof in example 9.2.11 as a starting point, prove that asymptotic normality still holds. In particular, show that

$$\sqrt{T} \left\{ s_T^2 - \text{var}[x_t] \right\} \xrightarrow{d} N(0, \gamma^2) \quad \text{where} \quad \gamma^2 := 2 \frac{1 + a^2}{(1 - a^2)^3}$$

Note that this exercise involves substantial calculation.

Ex. 13.4.3 In example 13.1.1 it was claimed that the threshold process studied in that example satisfies all of the conditions of assumption 13.1.1. Verify that this is the case.

Ex. 13.4.4 Verify the claim that (13.4) holds when assumption 13.1.2 is valid.

Ex. 13.4.5 Let $K \times 1$ random vector $\hat{\theta}_T$ be an estimator of θ . Suppose that this estimator is asymptotically normal, in the sense that

$$\sqrt{T}(\hat{\theta}_T - \theta) \xrightarrow{d} N(\mathbf{0}, \mathbf{C})$$

where $\mathbf{C}$ is positive definite. By exercise 3.5.36, for such a $\mathbf{C}$ there exists a symmetric $K \times K$ matrix $\mathbf{D}$ such that $\mathbf{D}\mathbf{C}\mathbf{D} = \mathbf{I}$. Let $\hat{\mathbf{D}}_T$ be a consistent estimator of $\mathbf{D}$. Show that

$$T\|\hat{\mathbf{D}}_T(\hat{\theta}_T - \theta)\|^2 \xrightarrow{d} \chi^2(K) \quad (13.23)$$

Incidentally, (13.23) can be used to test the null hypothesis that $\theta = \theta_0$. Under the null hypothesis, we have

$$T\|\hat{\mathbf{D}}_T(\hat{\theta}_T - \theta_0)\|^2 \xrightarrow{d} \chi^2(K)$$

Fixing α and letting c be the $1 - \alpha$ quantile of the $\chi^2(K)$ distribution, we reject the null if $T\|\hat{\mathbf{D}}_T(\hat{\theta}_T - \theta_0)\|^2 > c$. This test is asymptotically of size α .

Ex. 13.4.6 Consider once more the scalar linear Gaussian AR(1) model $x_{t+1} = ax_t + w_{t+1}$ with $|a| < 1$ and $\{w_t\}$ IID and standard normal. In example 13.1.3 we saw that the OLS estimator $\hat{a}_T$ is a consistent and asymptotically normal estimator of a , with asymptotic variance $1 - a^2$. Let's now look at another consistent and asymptotically normal estimate of a and compare its performance with the OLS estimator in terms of asymptotic variance (with small variance being better).

For this second estimator, we can exploit the fact that the sample variance converges to the stationary variance of x_t , which in this case is $1/(1 - a^2)$. Once we estimate the variance with the sample variance, we can solve for the value of a . The exercise is to formalize this idea, and show that

$$\sqrt{T}(a_T^* - a) \xrightarrow{d} N\left(0, (1 - a^2) \left(1 + \frac{1 - a^2}{2a^2}\right)\right) \quad (13.24)$$

where a_T^* is the estimator just described. Comments:

- The exercise shows that the OLS estimator has lower asymptotic variance than a_T^* (cf. (13.14) on page 361). We say that the OLS estimator is **relatively more efficient**.
- In attempting the exercise, you should first consult exercise 13.4.2 on page 370.

13.4.1 Solutions to Selected Exercises

Solution to Ex. 13.4.1. Let $\{x_t\}$ be as in the statement of the problem. The claim is that (13.22) holds. To see this, apply (7.27) on page 194. (All the relevant conditions are satisfied; see, in particular, example 7.2.10 on page 195. Observe also that the stationary distribution $\pi_\infty = N(0, 1/(1-a^2))$ has finite moments of all orders.) This gives

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T x_t = \sqrt{T} \left[\frac{1}{T} \sum_{t=1}^T x_t \right] = \sqrt{T} \left[\frac{1}{T} \sum_{t=1}^T x_t - \int s \pi_\infty(s) ds \right] \xrightarrow{d} N(0, v)$$

where, from (7.28),

$$v = \text{var}[x_0] + 2 \sum_{t=1}^{\infty} \text{cov}[x_0, x_t]$$

Recall here that $\mathcal{L}(x_0) = N(0, 1/(1-a^2))$, so this becomes

$$v = \frac{1}{1-a^2} + 2 \sum_{t=1}^{\infty} a^t \mathbb{E}[x_0^2] = \frac{1}{1-a^2} + \frac{2}{1-a^2} \sum_{t=1}^{\infty} a^t = \frac{1}{1-a^2} + \frac{2a}{(1-a)(1-a^2)}$$

Here the first equality follows from the discussion of correlation in §7.2.4. Simplifying the right-hand side gives $v = 1/(1-a)^2$ as claimed. $\square$

Solution to Ex. 13.4.2. In this exercise we are assuming that $x_{t+1} = ax_t + w_{t+1}$ where $|a| < 1$, $\{w_t\}$ is IID and standard normal, and $\mathcal{L}(x_0) = \pi_\infty = N(0, 1/(1-a^2))$. (The stationary distribution was obtained on page 193.) The aim is to show that

$$\sqrt{T} \left\{ s_T^2 - \frac{1}{1-a^2} \right\} \xrightarrow{d} N(0, \gamma^2) \quad \text{where} \quad \gamma^2 := 2 \frac{1+a^2}{(1-a^2)^3} \quad (13.25)$$

Letting $v := 1/(1-a^2)$, note from the manipulations in example 9.2.11 that it suffices to prove that

$$\sqrt{T} \left\{ \frac{1}{T} \sum_{t=1}^T x_t^2 - v \right\} \xrightarrow{d} N(0, \gamma^2) \quad (13.26)$$

All of the conditions of theorem 7.2.2 on page 193 are satisfied, and we apply the resulting CLT in (7.27) with $h(x) = x^2$ to get (13.26) with

$$\gamma^2 := \mathbb{E}[(x_0^2 - v)^2] + 2 \sum_{t=1}^{\infty} \mathbb{E}[(x_0^2 - v)(x_t^2 - v)] \quad (13.27)$$

To check (13.25) it only remains to show that this expression for γ^2 matches that in

(13.25). To see this, consider the first term on the right-hand side of (13.27). We have

$$\mathbb{E}[(x_0^2 - v)^2] = \mathbb{E}[x_0^4] - 2v\mathbb{E}[x_0^2] + v^2 = 2v^2$$

(The fourth moment of the distribution $N(0, \sigma^2)$ is $3\sigma^4$.) For the second term on the right-hand side of (13.27), we have

$$\mathbb{E}[(x_0^2 - v)(x_t^2 - v)] = \mathbb{E}[x_0^2 x_t^2] - v^2 = \mathbb{E}\left\{x_0^2 [a^t x_0 + s(w)]^2\right\} - v^2$$

for $s(w) := \sum_{i=1}^t a^{t-i} w_i$. Note that $s(w)$ and x_0 is independent. (For a Markov process, the initial condition is always assumed to be independent of the shocks.) Working through all the algebra, we get

$$\mathbb{E}[(x_0^2 - v)(x_t^2 - v)] = 2a^{2t}v^2$$

Returning to (13.27) we have

$$\gamma^2 = 2v^2 + 2 \sum_{t=1}^{\infty} 2a^{2t}v^2 = 2v^2 + 4 \frac{a^2}{1-a^2} v^2 \quad (13.28)$$

Using $v = 1/(1-a^2)$ and working through the remaining algebra will get you to the expression for γ^2 in (13.25). $\square$

Solution to Ex. 13.4.3. Let $\pi_{\infty}(s) = 2\phi(s)\Phi(qs)$, where $q := a(1-a^2)^{-1/2}$, ϕ is the standard normal density and Φ is the standard normal CDF. The process $\{x_t\}$ studied in example 13.4.3 is identically distributed as a result of our assumption that $\mathcal{L}(x_0) = \pi_{\infty}$ (see fact 7.2.5 on page 191). It remains to check the conditions on Σ_x . In the present case,

$$\Sigma_x = \mathbb{E}[x_t^2] = \int_{-\infty}^{\infty} s^2 \pi_{\infty}(s) ds = \int_{-\infty}^{\infty} s^2 2\phi(s)\Phi(qs) ds$$

To verify that Σ_x is positive definite, we need to check that the term on the right-hand side is strictly positive. This is clearly true, because the function inside the integral is strictly positive everywhere but zero. To be careful, we should also check that Σ_x is finite, and this is also true because $\Phi(qs) \leq 1$, and hence

$$\int_{-\infty}^{\infty} s^2 2\phi(s)\Phi(qs) ds \leq \int_{-\infty}^{\infty} s^2 2\phi(s) ds = 2 \int_{-\infty}^{\infty} s^2 \phi(s) ds = 2$$

Last, we need to show that

$$\frac{1}{T} \sum_{t=1}^T x_t^2 \xrightarrow{p} \Sigma_{\mathbf{x}} = \mathbb{E}[x_t^2] \quad (13.29)$$

Since the conditions of theorem 7.2.2 (page 193) are satisfied, the convergence in (13.29) is valid. $\square$

Solution to Ex. 13.4.4. Suppose that assumption 13.1.2 (page 357) is valid. We need to show that

$$\mathbb{E}[u_s u_t \mid \mathbf{x}_1, \dots, \mathbf{x}_t] = \begin{cases} \sigma^2 & \text{if } s = t \\ 0 & \text{if } s < t \end{cases}$$

On one hand, if $s = t$, then $\mathbb{E}[u_t^2 \mid \mathbf{x}_1, \dots, \mathbf{x}_t] = \mathbb{E}[u_t^2] = \sigma^2$ by independence. On the other hand, if $s < t$, then

$$\begin{aligned} \mathbb{E}[u_s u_t \mid \mathbf{x}_1, \dots, \mathbf{x}_t] &= \mathbb{E}[\mathbb{E}[u_s u_t \mid \mathbf{x}_1, \dots, \mathbf{x}_t, u_s] \mid \mathbf{x}_1, \dots, \mathbf{x}_t] \\ &= \mathbb{E}[u_s \mathbb{E}[u_t \mid \mathbf{x}_1, \dots, \mathbf{x}_t, u_s] \mid \mathbf{x}_1, \dots, \mathbf{x}_t] \\ &= \mathbb{E}[u_s \mathbb{E}[u_t] \mid \mathbf{x}_1, \dots, \mathbf{x}_t] \\ &= \mathbb{E}[u_s 0 \mid \mathbf{x}_1, \dots, \mathbf{x}_t] = 0 \end{aligned} \quad \square$$

Solution to Ex. 13.4.5. By assumption we have $\hat{\mathbf{D}}_T \xrightarrow{p} \mathbf{D}$ and $\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{z}$, where $\mathcal{L}(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{C})$. Applying Slutsky's theorem, we obtain

$$\hat{\mathbf{D}}_T \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{D} \mathbf{z} \quad (13.30)$$

Clearly $\mathbf{D} \mathbf{z}$ is normally distributed with mean $\mathbf{0}$. Moreover

$$\text{var}[\mathbf{D} \mathbf{z}] = \mathbf{D} \text{var}[\mathbf{z}] \mathbf{D}^\top = \mathbf{D} \mathbf{C} \mathbf{D} = \mathbf{I}$$

In other words, $\mathbf{D} \mathbf{z}$ is standard normal. As a result, $\|\mathbf{D} \mathbf{z}\|^2 \sim \chi^2(K)$. Applying the continuous mapping theorem to (13.30), we obtain

$$\|\hat{\mathbf{D}}_T \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta})\|^2 \xrightarrow{d} \|\mathbf{D} \mathbf{z}\|^2 \sim \chi^2(K)$$

This is equivalent to (13.23). $\square$

Solution to Ex. 13.4.6. Let $v := 1/(1 - a^2)$, the asymptotic variance of x_t . We can solve this for a to get

$$a = g(v) \quad \text{for} \quad g(v) := \sqrt{1 - \frac{1}{v}}$$

We take as our estimator a_T^* the statistic $g(s_T^2)$, where s_T^2 is the sample variance. We then have

$$\sqrt{T}(a_T^* - a) = \sqrt{T}\{g(s_T^2) - g(v)\}$$

Applying the delta method (see page 173) combined with the result in exercise 13.4.2 we have

$$\sqrt{T}(a_T^* - a) \xrightarrow{d} N(0, g'(v)^2 \gamma^2) \quad \text{where} \quad \gamma^2 := 2 \frac{1 + a^2}{(1 - a^2)^3}$$

Using the definition of v and a bit of algebra leads to the expression on the right-hand side of (13.24). $\square$