Chapter 5: Asymptotic Methods and Functional Central Limit Theorems

James Davidson
University of Exeter

Abstract

This chapter sketches the fundamentals of asymptotic distribution theory, and applies these specifically to questions relating to weak convergence on function spaces. These results have important applications in the analysis of nonstationary time series models. A simple case of the functional central limit theorem for processes with independent increments is stated and proved, after detailing the necessary results relating to the topology of spaces of functions and of probability measures. The concepts of weak convergence, tightness and stochastic equicontinuity, and their roles in the derivation of functional central limit theorems, are defined and reviewed. It is also shown how to extend the analysis to the vector case, and to various functionals of Brownian motion arising in nonstationary regression theory. The analysis is then widened to consider the problem of dependent increments, contrasting linear and nonparametric representations of dependence. The properties of Brownian motion and related Gaussian processes are examined, including variance-transformed processes, the Ornstein-Uhlenbeck process and fractional Brownian motion. Next, the case of functionals whose limits are characterized stochastic integrals is considered. This theory is essential to (for example) the analysis of multiple regression in integrated processes. The derivation of the Itô integral is summarized, followed by application to the weak convergence of covariances. The final section of the chapter considers increment distributions with infinite variance, and shows how weak convergence to a Lévy process generalizes the usual case of the FCLT, having a Gaussian limit.

Contents

1 Naïve Distribution Theory

We begin this chapter with a brief explanation of why its subject matter is important to econometricians. To make inferences about econometric models and their parameters, something must be known about the distributions of estimates and test statistics. What range will these random variables typically fall into when a certain hypothesis is true? How far outside this range should the statistic fall, before the chance of its realization is so small that we should conclude that the hypothesis is false?

At the elementary level, these difficult questions are often handled by postulating an idealized and simplified world satisfying the assumptions of the Classical Regression Model (CRM). In the CRM world, all the observed economic variables except one (the dependent variable) can be treated as fixed. That is to say, more precisely, 'fixed in repeated samples'. For illustration, consider the regression model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \tag{1.1}$$

Were we to have the opportunity to observe this phenomenon (the pair $\boldsymbol{y}, \boldsymbol{X}$) repeatedly, the variables $\boldsymbol{X}$ should assume the same sample values in successive drawings, and only the dependent variable $\boldsymbol{y}$ should vary. In addition, it is usually assumed that the dependent variable is normally and independently distributed with fixed variance, as $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. It would then follow that the least squares estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ has the property

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

and the regression 't ratios' and 'F statistics' would exactly follow the Student's t and F distributions when the null hypotheses are true. Our questions are then simply answered with reference to the t and F tables.

Popular as it is, this justification of inference procedures is, of course, dishonest. In economics (an essentially non-experimental discipline), cases in which a repeated sampling exercise will throw up the same pattern of explanatory variables are almost unheard of. When a new sample of firms or households is drawn, all the variables are randomly replaced, not just the dependent. However, notwithstanding that the CRM world is a fiction, the theory still yields generally valid results provided two assumptions hold. First, the sample observations must be independently distributed. Second, it is required that $\boldsymbol{u}|\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, where $\boldsymbol{u}|\boldsymbol{X}$ denotes the conditional distribution of $\boldsymbol{u}$, holding $\boldsymbol{X}$ fixed. It is sometimes thought that all is required is for u_t to be serially independent, implying no more than a notion of correct specification, but this is incorrect. The rows of $\boldsymbol{X}$ also need to be independent of each other, a condition virtually never attained in time series data. In time series it is rarely possible to assume that $\boldsymbol{x}_{t+j}$ for $j > 0$ does not depend on u_t, a shock preceding it in time, and this would invalidate the conditioning exercise.

If the sample were truly independent, as in a randomly drawn cross-section for example, conditioning $\boldsymbol{u}$ on $\boldsymbol{X}$ is merely equivalent to conditioning u_t on $\boldsymbol{x}_t$ for each $t = 1, \ldots, T$ (T denoting sample size). In this case, while the CRM does not hold and $\hat{\boldsymbol{\beta}}$ is not normally distributed (unconditionally), it is still the case that $\hat{\boldsymbol{\beta}}|\boldsymbol{X}$ is normal. The t and F statistics for the regression follow the t and F distributions exactly, since their conditional distributions are free of dependence on $\boldsymbol{X}$, and hence equivalent to their unconditional distributions. We may 'act as if' the CRM assumptions hold. However, when either the conditional normality assumption fails, or the sample is in any way dependent (both conditions endemic in econometric data sets), the only recourse available is to large-sample (asymptotic) approximations.

2 Asymptotic Theory

The asymptotic approach to econometric inference is to derive approximate distributions under weaker assumptions than in the CRM setup, where the approximation improves with sample size. These arguments invoke a collection of theorems on stochastic convergence. In this section the scene is set with a brief resume of the fundamental ideas, starting with the axiomatic probability model. For further reading on these topics, see for example Davidson (1994). Another accessible text aimed at nonspecialists is Pollard (2002).

The standard representation of a probability space (a mathematical model of a random experiment) is the triple $(\Omega, \mathcal{F}, P)$, where Ω is the sample space (the collection of all random objects under consideration), $\mathcal{F}$ is a σ-field, being the collection of random events (subsets of Ω) to which probabilities are to be assigned, and P is the probability measure, such that $P(A) \in [0, 1]$ is the probability of the event A, for each $A \in \mathcal{F}$. Recall that a σ-field is a class of subsets of Ω having the properties

(a) $\Omega \in \mathcal{F}$.

(b) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ where $A^c = \Omega - A$.

(c) if $A_1, A_2, A_3, \ldots \in \mathcal{F}$ (an infinite collection) then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

If $\mathcal{C}$ is any class of subsets of Ω the notation $\sigma(\mathcal{C})$ represents the smallest σ-field containing $\mathcal{C}$. This is called the 'σ-field generated by $\mathcal{C}$'. A probability measure (p.m.) $P : \mathcal{F} \mapsto [0, 1]$ is then a set function having the properties $P(\Omega) = 1$, and

$$P\big(\bigcup_{i=1}^{\infty} A_i\big) = \sum_{\iota=1}^{\infty} P(A_i)$$

for disjoint collections $A_1, A_2, A_3, \ldots \in \mathcal{F}$.

It is worth being reminded of why a probability space is defined in this manner. We need a way to assign probabilities to all sets of interest, but these are usually too numerous to allow a rule to assign each one individually. Hence, we assign probabilities to a class $\mathcal{C}$ of 'basic' events, and then extend these probabilities to elements of $\sigma(\mathcal{C})$ using the rules of set algebra. The *extension theorem* is the fundamental result in probability, stating that if the class $\mathcal{C}$ is rich enough, probabilities can be uniquely assigned to all the members of $\sigma(\mathcal{C})$. $\mathcal{C}$ is called a *determining class* for P. However, to go beyond $\sigma(\mathcal{C})$ is to run the risk of encountering so called 'non-measurable' sets. In infinite spaces it is not feasible to simply let $\mathcal{F}$ be the power set of Ω without running into contradictions.

Often, $(\Omega, \mathcal{F}, P)$ is to be thought of as the 'fundamental' probability space for a particular random experiment, where the outcomes are not necessarily numerical magnitudes. The outcomes are then mapped into a 'derived' space, by the act of measurement. The best known example of a derived probability space, on which random variables live, is $(\mathbb{R}, \mathcal{B}, \mu)$, where $\mathbb{R}$ denotes the real line, and $\mathcal{B}$, called the *Borel field* of $\mathbb{R}$, is the σ-field generated by the set of the half-lines $(-\infty, x]$ for $x \in \mathbb{R}$. The fact that this is a rich enough collection is evidenced by the fact that $\mathcal{B}$ is also the σ-field generated by the open sets of $\mathbb{R}$, containing also the intervals, the closed sets, and much more besides. A random variable (r.v.) can be thought of as a measurable mapping $X : \Omega \mapsto \mathbb{R}$ where 'measurable' means that $X^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{B}$, and the probabilities are assigned by the rule $\mu(A) = P(X^{-1}(A))$ for each $A \in \mathcal{B}$. A fundamental result, since the half-lines form a determining class for this space, is that specifying a cumulative distribution function (c.d.f.) $F(x) = \mu((-\infty, x])$ is sufficient to define $\mu(A)$ uniquely for every $A \in \mathcal{B}$.

2.1 Stochastic Convergence

Given this background, we can now describe the basic toolbox of results for asymptotic analysis. The essential idea is that of a random sequence, $X_1, X_2, \ldots, X_T, \ldots = \{X_t\}_{t=1}^{\infty}$, and the essential problem whether, and how, such a sequence might converge as T increases. The more familiar case is the sequence of constants $\{a_t\}_{t=1}^{\infty}$; say, $a_t = t$, or $a_t = 1/t$. If for every $\varepsilon > 0$ there exists an integer N_ε such that $|a_T - a| < \varepsilon$ for all $T > N_\varepsilon$, then we say 'a_T converges to a', and write $a_T \to a$. The first of our examples does not converge, on this criterion, but the second one converges to 0.

By contrast, there are several different ways to capture the notion of the convergence of a sequence of r.v.s. The basic approach is to define certain associated nonstochastic sequences, and consider whether these converge. Let μ_T represent the probability measure associated with X_T, such that $\mu_T(A) = P(X_T \in A)$, where A is any set of real numbers to which probabilities are to be assigned. Here are four contrasting convergence concepts.

1. *Almost sure convergence*

   $X_T(\omega) \to X(\omega)$ for every $\omega \in C$, where $C \in \mathcal{F}$ and $P(C) = 1$. Write $X_T \overset{\text{a.s.}}{\to} X$.

2. *Convergence in mean square*

   $E(X_T - X)^2 \to 0$. Write $X_T \overset{\text{ms}}{\to} X$.

3. *Convergence in probability*

   $P(|X_T - X| < \varepsilon) \to 1$ for all $\varepsilon > 0$. Write $X_T \overset{\text{pr}}{\to} X$.

4. *Convergence in distribution* (weak convergence of probability measures)

   $\mu_T(A) \to \mu(A)$ for every $A \in \mathcal{F}$ such that $\mu(\delta A) = 0$, where δA denotes the boundary points of A. Equivalently, $F_T(x) \to F(x)$ at all continuity points of F. Write $\mu_T \Rightarrow \mu$ or $X_T \overset{\text{d}}{\to} X$, where $X \sim \mu$.

Almost sure convergence and convergence in mean square both imply convergence in probability, and convergence in probability implies weak convergence, and is equivalent to weak convergence when the limit is a constant. Otherwise, none of the reverse implications hold. There is an important distinction to be made between weak convergence and the other modes, since this specifies a limiting distribution but not a limiting r.v. In other words, if $X_T \overset{\text{pr}}{\to} X$ this implies that (say) $|X_{2T} - X_T| \overset{\text{pr}}{\to} 0$, such that when the sample is large enough, the effect of doubling it is negligible. However, it is *not* the case that $X_T \overset{\text{d}}{\to} X$ implies $|X_{2T} - X_T| \overset{\text{d}}{\to} 0$. What converges in this case is the sequence of probability measures, not a sequence of random variables. The conventional, rather imprecise, notation implies that the limit is the distribution of of a specified r.v. X, and sometimes this is written in the more explicit form '$X_T \overset{\text{d}}{\to} \text{N}(0, \sigma^2)$' or similar.

The sequence with typical index T that satisfies these convergence criteria is usually a sequence of sample statistics or estimators, hence functions of T data points. The data points themselves also constitute real sequences, which we often distinguish with the typical index t. The following are the most important asymptotic results concerning the sequences generated by constructing averages of data sequences $\{U_1, U_2, \ldots, U_T\}$ as T increases. Let

$$\bar{U}_T = \frac{1}{T} \sum_{t=1}^{T} U_t.$$

1. *The weak (strong) law of large numbers* (W(S)LLN)

   If $E|U_t| < \infty$ for each $t \geq 1$, then under suitable additional regularity conditions

$$\bar{U}_T - E(\bar{U}_T) \overset{\text{pr}}{\to} 0 \ (\overset{\text{a.s.}}{\to} 0) \qquad (2.1)$$

2. *The central limit theorem* (CLT)

   If $EU_t^2 < \infty$ for each $t \geq 1$, then under suitable additional regularity conditions

$$\frac{\bar{U}_T - E(\bar{U}_T)}{\sqrt{E(\bar{U}_T - E(\bar{U}_T))^2}} \overset{\text{d}}{\to} \text{N}(0, 1) \qquad (2.2)$$

Two points to note. First, (2.1) does not imply that $E(\bar{U}_T)$ must be a constant independent of T, or even that it is a convergent sequence, although of course this is often the case. Second, it is often the case that

$$E(\bar{U}_T - E(\bar{U}_T))^2 = \sigma^2/T$$

where σ^2 is the common variance of the data points, and then (2.2) can be restated with the customary '$\sqrt{T}$' normalization. Our version simply emphasizes that it is the sample average, re-normalized to have zero mean and unit variance, that is the sequence of interest here.

The laws of large numbers are rather straightforward and intuitive. Few would doubt that a sample average usually approaches a limiting value as the sample increases, simply because the marginal contribution of the last term is inevitably getting smaller relative to the whole. It is perhaps of more interest to note those situations where the convergence fails, primarily when $E|U_t| = \infty$. The Cauchy distribution is a well-known counter-example, in which new sample drawings can be large enough, with high enough probability, that the average never settles down to a fixed value – and is, in fact, just another Cauchy variate.

On the other hand, many people find the fact that the normal (or Gaussian) 'bell curve' arises by aggregating many independent zero-mean shocks to be mysterious at an intuitive level, even though the mathematics is quite simple and transparent. One way to appreciate the mechanism at work is through the fact that, among those possessing a variance, the Gaussian is the unique distribution to be invariant under the summation of independent drawings. As is well known, the *characteristic function* (ch.f.) of any random variable U, defined as

$$\phi_U(\lambda) = E(e^{i\lambda U}) \tag{2.3}$$

is an equivalent representation of the distribution. If $U_1, \ldots, U_T$ are independent drawings from this distribution, and $S_T = (U_1 + \cdots + U_T)/a_T$ for some $a_T > 0$, note that

$$\begin{aligned}\phi_{S_T}(\lambda) &= E(e^{i\lambda U_1/a_T}) \cdots E(e^{i\lambda U_T/a_T}) \\ &= \phi_{U_1}(\lambda/a_T) \cdots \phi_{U_T}(\lambda/a_T).\end{aligned} \tag{2.4}$$

This identity raises the interesting question of whether, for some sequence a_T, the functional forms of $\phi_{S_T}(\lambda)$ and $\phi_U(\lambda)$ are the same. As is well known, the Gaussian distribution with mean 0 and variance σ^2 has ch.f. $\phi_U(\lambda) = e^{-\sigma^2\lambda^2/2}$, and in this case it is easily seen that setting $a_T = \sqrt{T}$ yields the desired result:

$$\phi_{S_T}(\lambda) = (e^{-\sigma^2(\lambda/\sqrt{T})^2/2})^T = e^{-\sigma^2\lambda^2/2}. \tag{2.5}$$

This fact helps us to appreciate how the Gaussian distribution acts as an 'attractor' for sums of independent r.v.s, normalized by the square root of the sample size. A formal proof of the CLT may be obtained by considering the Taylor's expansion of $\phi_{S_T}(\lambda)$ and showing that the first and second order terms match those of (2.5), while the higher order terms are of small order in T. Note that the variances of the r.v.s must be finite, and of course the '$\sqrt{T}$' normalization corresponds to the familiar summation rule for variances of independent sums.

We will avoid specifying regularity conditions for the LLN and CLT in detail here, since there are so many different ways to formulate them. Various specific cases are cited in the sequel. Let it suffice to say two things at this point. First, if the sequence elements are identically and independently distributed, then no extra conditions are required. However, if the sequence elements are either heterogeneously distributed, or serially dependent, or both, then a range of different sufficient restrictions can be demonstrated. A condition that frustrates the CLT is where a finite number of terms of the sum are influential enough to affect the whole. The well-known *Lindeberg condition* for the CLT rules out this possibility. *Uniform integrability* is a distinct but related restriction relevant to the LLN, ruling out certain pathological cases where the absolute moments depend excessively on extreme values in the limit. However, requiring that the order of existing moments be slightly larger than 1 in the LLNs, or larger than 2 in the CLT, is a simple way to avoid a failure of either of these weaker but more subtle conditions.

Standard treatments of the CLT assume stationary process increments, but it is an important fact that this is not a necessary restriction. In particular, conditions of the form

$$E(U_t^2) = t^\alpha\sigma^2 \tag{2.6}$$

can be accommodated, for any $\alpha > -1$. In other words, the variances may diverge to infinity, and even converge to 0 provided the variance sequence is not actually summable. This form of

CLT is especially useful for deriving certain forms of the functional CLT, as explained in Section 6.

Regularity conditions restricting the dependence are of many sorts. Some involve assuming the process is *linear* (i.e., has an infinite-order MA representation with independent shocks) and placing restrictions on the coefficients. Others, such as mixing and near-epoch dependence, are purely non-parametric conditions imposing 'short memory'. The martingale difference assumption is a sufficient restriction on dependence for all these results, a very useful fact for econometric applications in particular. We say more about these cases in the discussion of the FCLT, see Section 5.

Certain supplementary results, the handmaidens of the LLN and CLT so to speak, are constantly invoked in asymptotic analysis.

1. *Slutsky's Theorem* If $X_T \overset{\text{pr}}{\to} a$, and $g(\cdot)$ is continuous at a, then $\operatorname{plim} g(X_T) = g(a)$.

2. *Cramér's Theorem* If $Y_T \xrightarrow{\text{d}} Y$ and $X_T \xrightarrow{\text{pr}} a$ then

   (i) $X_T + Y_T \xrightarrow{\text{d}} a + Y$

   (ii) $X_T Y_T \xrightarrow{\text{d}} aY$

   (iii) $\dfrac{Y_T}{X_T} \xrightarrow{\text{d}} \dfrac{Y}{a}$ when $a \neq 0$.

3. *Continuous Mapping Theorem* (CMT) If $X_T \xrightarrow{\text{d}} X$ and $g(\cdot)$ is continuous, then $g(X_T) \xrightarrow{\text{d}} g(X)$.

Versions of these results for random vectors are easy extensions, using the following result in particular.

4. *Cramér–Wold Theorem* Let $\boldsymbol{X}_T$ be a sequence of random vectors. Then $\boldsymbol{X}_T \xrightarrow{\text{d}} \boldsymbol{X}$ if and only if for all conformable fixed vectors $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}'\boldsymbol{X}_T \xrightarrow{\text{d}} \boldsymbol{\lambda}'\boldsymbol{X}$.

Additional background on all these results, including formal statements, proofs and mathematical details, can be found in a number of specialist texts with an econometric emphasis and motivation such as McCabe and Tremayne (1993), Davidson (1994) and White (1999).

2.2 Application to Regression

Now, to return to the problem posed in Section 1. Letting T denote sample size, and $\hat{\boldsymbol{\beta}}$ the least squares estimator as before, write

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{T}\right)^{-1} \frac{\boldsymbol{X}'u}{\sqrt{T}}. \tag{2.7}$$

Subject to regularity conditions, the matrix $T^{-1}\boldsymbol{X}'\boldsymbol{X}$ converges in probability to its mean, $\boldsymbol{M}_{xx}$, thanks to the WLLN. This matrix must be nonsingular. Subject to further regularity conditions, the vector $\boldsymbol{X}'\boldsymbol{u}/\sqrt{T}$ is jointly normally distributed in the limit according to the vector generalization of the CLT, making use of the Cramér-Wold theorem. The variance matrix of this vector under the limiting distribution is shown to equal $\sigma^2\boldsymbol{M}_{xx}$, making use of the assumed uncorrelatedness of $\boldsymbol{x}_t$ and u_t and the result (using the consistency of $\hat{\boldsymbol{\beta}}$ and the Slutsky theorem) that $s^2 \overset{\text{pr}}{\to} \sigma^2$.

The Slutsky Theorem (the result that the plim of the inverse matrix is the inverse of the plim) and the Cramér Theorem (the result that $T(\boldsymbol{X}'\boldsymbol{X})^{-1}$ can be replaced by $\boldsymbol{M}_{xx}^{-1}$ in the limiting distribution) can now be combined with the CLT and WLLN to yield the conclusion

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2\boldsymbol{M}_{xx}^{-1}).$$

Since the limiting covariance matrix $\sigma^2\boldsymbol{M}_{xx}^{-1}$ is consistently estimated by $Ts^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, as just indicated, and since the Student's t family approaches the normal as the degrees of freedom increase, the spurious 'Student's t' result can therefore be justified as an approximation to the distribution actually attained by the 't-ratios' in large samples. The discussion must then focus on the variety of regularity conditions required for this large-sample result to hold.

In independent samples, these conditions are comparatively simple to express. Letting $\boldsymbol{x}_t'$ denote the tth row of $\boldsymbol{X}$, and u_t the corresponding element of $\boldsymbol{u}$, two basic requirements are

$$E(\boldsymbol{x}_t u_t) = 0$$

and

$$E(\boldsymbol{x}_t\boldsymbol{x}_t' u_t^2) = \sigma^2\boldsymbol{M}_{xx} < \infty.$$

If it is possible to assume that the data are identically distributed, then both $E(u_t^2)$ and $E(\boldsymbol{x}_t\boldsymbol{x}_t')$, if they exist, are finite constants not depending on t. If the data are heterogeneously distributed, a conveniently stated sufficient condition for the CLT is

$$E|\boldsymbol{\lambda}'\boldsymbol{x}_t u_t|^{2+\delta} \leq B$$

for all conformable vectors λ of unit length, some $\delta > 0$, and all $t = 1, \ldots T$, $T \geq 1$, where B is a finite bound (see Davidson 2000, 3.5.1).

In time series, matters are always more complicated, because serial independence of $(\boldsymbol{x}_t, u_t)$ is an improbable restriction, as noted above. We are forced to introduce asymptotic results for dependent processes. One might do this by assuming that the independent variables follow some well-known model such as the VAR, so that the amount of dependence depends in a fairly simple way on the autoregressive root closest to the unit circle. However, such assumptions would often be heroic, requiring much more detailed knowledge of the generation mechanism of the sample than we would ever need under serial independence. It is also possible to cite mixing and/or near-epoch dependence conditions (see Sections 5.4 and 5.5), although when this is done the conditions are not easily verified without specifying an explicit model. Asymptotics can therefore be problematic in time series, and there is also the rarely cited fact that dependence must slow the rate of convergence in the CLT, reducing the effective sample size. Subject to these caveats, asymptotic theory may provide a workable if imperfect solution to the inference problem.

2.3 Autoregressive and Unit Root Processes

We now examine a particularly well-known example of a time series model in a bit more detail. We focus on the simple case,

$$x_t = \lambda x_{t-1} + u_t, \qquad u_t \sim iid(0, \sigma_u^2)$$

where $x_0 = 0$; generalizations with more lags and more variables will not contribute much to the important insights we seek. If $|\lambda| < 1$, it is well-known that the autoregressive series is asymptotically stationary with a finite variance $E(x_t^2) = \sigma_u^2/(1 - \lambda^2)$, and that its dependence (serial correlation) declines exponentially. Applying the arguments in the previous section yields the result

$$\sqrt{T}(\hat{\lambda}-\lambda)=\frac{T^{-1/2}\sum_{t=2}^{T}x_{t-1}u_t}{T^{-1}\sum_{t=2}^{T}x_{t-1}^2}\underset{asy}{\sim}\mathrm{N}(0,1-\lambda^2). \tag{2.8}$$

However, it is evident that this formula is tending to break down as λ approaches 1. It appears that at that point, $\sqrt{T}(\hat{\lambda}-\lambda)$ is approaching 0 in probability, since the variance is tending to zero.

To analyze the problem in more detail, consider first the denominator of (2.8). With $\lambda = 1$ we have

$$x_t = x_{t-1}+u_t=\sum_{s=1}^{t}u_t. \tag{2.9}$$

This is called an *integrated* or partial sum process. Under the other stated assumptions,

$$E(x_{t-1}^2)=(t-1)\sigma_u^2. \tag{2.10}$$

The sample average of these terms is therefore likely to be diverging, not tending to a constant. The obvious thing is to consider instead

$$T^{-2}\sum_{t=2}^{T}x_{t-1}^2. \tag{2.11}$$

Since it is well-known that

$$\sum_{t=2}^{T}(t-1)=\frac{T(T-1)}{2}$$

it is evident from (2.10) that (2.11) has a finite mean in the limit, equal to $\sigma_u^2/2$. So far, so good. However, is there a law of large numbers with this normalization, such that (2.11) converges to a constant? This is where the conventional arguments start to come unstuck, for the sequence $\{x_1^2,...,x_T^2\}$ does not satisfy a LLN. Since the mean (and also variance) of these terms grows with T, at most a finite number of them at the end of the sequence must dominate the average. The variable represented in (2.11) is accordingly random even in the limit. Moreover, while we can show that $T^{-1}\sum_{t=2}^{T}x_{t-1}u_t$ converges to a limiting distribution, this is both non-normal and also correlated with (2.11). Therefore, the limiting distribution of the appropriately normalized error of estimate $T(\hat{\lambda}-\lambda)$ is not normal. The 'naïve' distribution theory for least squares estimates therefore fails, even as an approximation.

The strategy for tackling the problem involves two steps. First, we have to recognize that the integrated time series $\{x_1,...,x_T\}$ has a limiting distribution itself, after suitable normalization. Second we make use of this distribution, and of the CMT and some related tools, to derive the limiting distribution of the sample statistic. To illustrate these developments we continue with the simple partial-sum process (2.9). Since the time series is a cumulation of independent shocks, it is immediate from the CLT that

$$\frac{1}{\sqrt{T}}x_T\xrightarrow{\mathrm{d}}\mathrm{N}(0,\sigma_u^2). \tag{2.12}$$

However, if T is an even number then

$$\frac{1}{\sqrt{T/2}}x_{T/2}\xrightarrow{\mathrm{d}}\mathrm{N}(0,\sigma_u^2) \tag{2.13}$$

is equally true. Indeed, any fixed fraction of the sample, such that its size increases with T, can be held to obey the corresponding rule. Therefore, let r denote any point of the interval $[0, 1]$, and use this to select a part of the sample by defining the notation $[Tr]$ to mean the largest integer not exceeding Tr. Then (2.12) can be generalized to

$$\frac{1}{\sqrt{T}} x_{[Tr]} \xrightarrow{\mathrm{d}} \mathrm{N}(0, r\sigma_u^2).$$

The next key step is to observe that (for any particular realization $\{u_1, ..., u_T\}$) the equations

$$X_T(r) = \frac{1}{\sigma_u \sqrt{T}} x_{[Tr]}, \quad 0 \leq r \leq 1 \tag{2.14}$$

define a function on the real line. This function is discontinuous, having little jumps at the points where $Tr = [Tr]$, but it has the interesting property that, regarded a drawing from the underlying joint distribution of shocks, its values in the intervals $(t-1)/T \leq r < t/T$, for $t = 1, ..., T$, are tending to become Gaussian with variance r. Of course, at the same time the intervals are getting narrower as T increases. The variance of the random "jumps" shrinks by a factor $1/T$ as the intervals shrink by a factor of $1/T$. It appears that the limiting case is a continuous 'Gaussian function'. In the following section we review the properties of this limiting function in detail, and then go on to consider the manner of convergence to the limit.

3 Distributions on a Function Space

The first major hurdle is to extend distribution theory from comparatively simply objects like random variables to very complex objects such as random functions. When we say 'complex', the point to bear in mind is that a function on a real-valued domain, say $x(r)$ for $0 \leq r \leq 1$, represents an uncountable infinity of real numbers. To each of these points, in principle, a distribution has to be assigned. We are familiar with the idea of a joint distribution of two, three, or several random variables, that has to specify not just the distribution of each one standing alone (the marginal distributions), but also the nature of their interactions and the dependence between them. The extension of these ideas to the case of functions is evidently fairly heroic, and as we shall see, a number of interesting mathematical problems arise along the way. The acknowledged primary source for many of the results is Billingsley (1968). Although a fundamental underpinning of modern asymptotic theory, the material of this section has yet to make its way into mainstream econometrics texts. The most accessible reference is probably Davidson (1994). Other useful sources include Pollard (1984) and Jacod and Shiryaev (1987), and many probability texts devote sections to stochastic processes, see for example Feller (1971), Billingsley (1986) or Dudley (1989).

3.1 Random Sequences

The sequences considered in this section are not, as a rule, assumed to be converging like the sequences discussed in Section 2.1. They are typically collections of sequential observations, with equally spaced dates attached (the indices t). Being an ordered, countably infinite collection of real numbers, one may also conveniently think of a sequence as a point in the space $\mathbb{R}^\infty$.

Given the probability space $(\Omega, \mathcal{F}, P)$, a random sequence may be defined as a measurable mapping from Ω to $\mathbb{R}^\infty$. Note the significance of this definition; *one* drawing $\omega \in \Omega$ maps into an *infinite* sequence

$$x(\omega) = \{X_1(\omega), X_2(\omega), \ldots, X_t(\omega), \ldots\}.$$

Our next task is to construct this derived probability space in which the random elements are infinite sequences. For obvious reasons, this is not a trivial undertaking.

Let $\mathcal{C}$ denote the collection of *finite-dimensional cylinder sets* of $\mathbb{R}^\infty$, that is, the sets of $\mathbb{R}^\infty$ of which at most a finite number of sequence coordinates are restricted to sets of $\mathcal{B}$ (the Borel field of of $\mathbb{R}$). Thus, the elements of $\mathcal{C}$ can be thought of as mapping one-for-one into random vectors, points in $\mathbb{R}^k$ for some k, representing the number of restricted sequence coordinates. Recall that a Borel field of any space is the σ-field generated by the open sets. In this case, it is possible to show that $\mathcal{B}^\infty$, the Borel field of sets of $\mathbb{R}^\infty$, is identical with $\sigma(\mathcal{C})$, the smallest σ-field containing $\mathcal{C}$. It remains to show that distributions can be constructed on the pair $(\mathbb{R}^\infty, \mathcal{B}^\infty)$.

The fundamental result is *Kolmogorov's consistency theorem*. It is difficult to give a formal statement of the theorem without resorting to technical language, but an informal explanation goes as follows. First, the distribution of any k coordinates of the sequence can be represented as a distribution of a random k-vector in the space $(\mathbb{R}^k, \mathcal{B}^k, \mu_k)$, which is a reasonably straightforward extension of the one-dimensional case. Even with an infinite sequence, marginalizing with respect to all but k coordinates yields such a distribution. These are called the *finite dimensional distributions* (or *fidis*) of the sequence. Note that, for any choice of k coordinates, the largest of them is still a finite number, and so the fidis can always be represented in this way. The consistency theorem then states that, given a family of distributions $\{\mu_k\}$ specified for every finite k, subject to the consistency condition

$$\mu_k(E) = \mu_m(E \times \mathbb{R}^{m-k}) \text{ for } E \in B^k \text{ and } m > k > 0$$

there exists a unique infinite random sequence distributed on $(\mathbb{R}^\infty, \mathcal{B}^\infty, \mu_\infty)$ such that the collections of the μ_k, for $k = 1, 2, 3, \ldots$, are the fidis of the sequence. In other words, to specify a unique distribution for x it is sufficient to specify the fidis in a suitable manner. This overcomes the practical problem of specifying how an infinite sequence of coordinates is distributed. The consistency condition says that, if μ_m is a fidi, then the distributions μ_k for $k < m$ are obtained in the usual way, by marginalizing with respect to the $m - k$ extra coordinates. Another way to express the same result is to say that the cylinder sets $\mathcal{C}$ form a determining class for x. If probabilities are assigned to each member of $\mathcal{C}$ (a feasible project by standard methods) the whole distribution of x is given uniquely.

3.2 Spaces of Functions

The number of coordinates of a sequence is countably infinite. The number of coordinates of a function is uncountable, which is to say, equipotent with the real continuum, and not capable of being labelled using the integers as an index set.

Let $R_{[0,1]}$, also denoted as just R when the context is clear, denote the set of all possible real-valued functions $x : [0, 1] \mapsto \mathbb{R}$, an element of which associates every point $r \in [0, 1]$ with a unique value $x(r)$. If the domain represents time, as it usually does, we call this a *process*, and if x is a random drawing from a specified distribution, a *stochastic* process. To construct such a distribution, the first question to be considered is whether a Borel field can be constructed for R, by analogy with $\mathbb{R}$ and $\mathbb{R}^\infty$. Since a Borel field is generated by the open sets of a space, this means being able to define an open set, or equivalently, to define a topology on the space.[1] This is usually done by defining a metric, a measure of 'closeness' of two elements of the space, and so making the space into a metric space. The usual metric adopted for $\mathbb{R}$ is, of course, the *Euclidean*

[1] A *topology* on a space is a collection of subsets that includes the whole space and the empty set, and is closed under arbitrary unions and finite intersections. Such sets are called *open*. This defines an open set, but the usual characterization of openness in $\mathbb{R}$ and other metric spaces derives from this fundamental definition.

distance $|x-y|$, for real numbers x and y, although the metric

$$d_0(x,y) = \frac{|x-y|}{1+|x-y|}$$

is topologically equivalent to $|x-y|$, while being bounded by 1.[2] Constructing a topology for $\mathbb{R}^\infty$ is a formalization that could be avoided, and so was left implicit in the discussion of random sequences, although note that a valid metric for this purpose (also bounded) is provided by

$$d_\infty(x,y) = \sum_{k=1}^{\infty} 2^{-k} d_0(x_k, y_k).$$

In the case of a space of functions, however, the topological properties of the space are a central issue. Since we can no longer enumerate the coordinates, a natural alternative is to adopt the *uniform metric*, defining the distance between elements $x, y \in R$ as

$$d_U(x,y) = \sup_{0\le t\le 1} |x(t) - y(t)|.$$

Armed with this definition, we can define the Borel field $\mathcal{B}_R$ of the metric space (R, d_U) as the smallest σ-field containing the *open spheres* of the space, which are defined as

$$B(x,r) = \{y \in R : d_U(x,y) < r\} \text{ for all } x \in R,\ r > 0.$$

How, then, to assign probabilities to sets of $(R, \mathcal{B}_R)$? One place we might start is with the finite dimensional cylinder sets; in other words, sets of functions that are unrestricted except at a finite number of coordinates, $t_1, \ldots, t_k$. For example, consider a case with $k = 2$; the open set $A = \{x \in R : x(1/2) < 0, 0 < x(3/4) < 2\}$. Let $\mathcal{H}$ denote the set of all such finite-dimensional sets, and let $\mathcal{P} = \sigma(\mathcal{H})$. $\mathcal{P}$ is called the *projection* σ-field, since we can think of $\mathcal{H}$ are the set of projections of the function onto finite sets of coordinates. If this case were to be comparable to the sequence case, then we should have $\mathcal{P} = \mathcal{B}_R$. Alas, it turns out that $\mathcal{P} \subset \mathcal{B}_R$. In other words, $\mathcal{B}_R$ contains sets that are not countable unions of $\mathcal{H}$ sets. The non-countability of the domain presents a problem.

As a first approach to the measurability problem, we might attempt to construct a probability measure on $(R, \mathcal{P})$. Assuming the fundamental space $(\Omega, \mathcal{F}, P)$, let $x : \Omega \mapsto R$ denote an $\mathcal{F}/\mathcal{P}$-measurable mapping, and so generate a probability measure (p.m.) derived from P. The fidis associated with this distribution are the joint distributions of finite sets of coordinates $(x_{t_1}, \ldots, x_{t_k})$. The consistency theorem can be extended to identify this p.m. uniquely with the fidis, provided a second consistency condition is satisfied. This is as follows:

> Permuting the coordinates $t_1, \ldots t_k$ changes the p.m. according to the same rule as permuting the integers $1, \ldots, k$.

[2] A metric on a space $\mathbb{S}$ is a distance measure for pairs of points of the space, having the following properties for all $x, y \in \mathbb{S}$:

1. $d(x,y) = d(y,x)$
2. $d(x,y) = 0$ if and only if $x = y$
3. $d(x,y) \le d(x,z) + d(y,z)$, $z \in \mathbb{S}$

There is no way to represent this distribution by a simple device such as the c.d.f. or characteristic function. However, the expected values $E(f(x))$, for all bounded continuous functionals[3] $f : R \mapsto \mathbb{R}$, are always defined uniquely and these values can be used to fingerprint the distribution.

So far, so good, and if this approach serves, it is the simplest available. However, the problem is that $\mathcal{P}$ may not contain all the cases we are interested in. Our goal should be to assign probabilities to all the Borel sets. This has proved impossible because the space (R, d_U) is really too big to handle. It is not a *separable* space, which means, as we shall see, that it contains Borel sets that cannot have measures assigned to them without running into inconsistencies. There are two ways we can attempt to overcome this difficulty, both of which, separately or in combination, have been exploited in the literature. The first is to consider a suitable subset of R containing most of the functions of interest. The second is to adopt a different topology, so that opens sets can be defined in a more tractable way. The first of these methods involves fewer technicalities, but is mathematically rather clumsy. The second – or strictly, a combination of both approaches – has been the most favoured technique in recent research

3.3 The Space $C_{[0,1]}$

One way to view the problem of applying distribution theory to functions is to find generalizations of the commonplace relations between real numbers that we use, often unconsciously, to construct distributions of random variables on the line. One way to achieve this is by working in the space of *continuous* functions on the unit interval, equipped with the uniform metric. This is denoted $C_{[0,1]}$, properly $(C_{[0,1]}, d_U)$ when it is necessary to specify the choice of metric, but also sometimes just C, when the context is clear.

The main virtue of $C_{[0,1]}$ is that it is a separable space, which means that it contains a countable, dense subset.[4] The space $\mathbb{R}$ is separable, since the rational numbers are countable, and also dense in the space. To show that $C_{[0,1]}$ is separable, one can exhibit the set of *piecewise-linear* functions, which consists of functions constructed from a countable set of points of the domain joined up with straight lines – the type of construction commonly used to plot discrete time series on a graph. If the points are assigned rational ordinates, then a countable collection of numbers defines each member of the set, and accordingly the set itself is countable – yet we can show that every continuous function is arbitrarily close, in d_U, to a member of this set.

The second important property of $C_{[0,1]}$ is *completeness.* This means that every Cauchy sequence of elements has a limit lying in the set. Recall that a Cauchy sequence is one in which the successive points are getting arbitrarily close together as we move down it. Although it is not hard to define sequences in $C_{[0,1]}$ having discontinuous limits, in the uniform metric such sequences cannot be Cauchy, because going from continuous to discontinuous must involve a positive jump at some point – and remember that all the points of the sequence must lie in $C_{[0,1]}$. Completeness is another property shared with $\mathbb{R}$, for Cauchy sequences of real numbers all have a real number as the limit.

The cash value of these properties, from the present point of view, is very simply that they imply $\mathcal{P}_C = \mathcal{B}_C$, where these are the restrictions to $C_{[0,1]}$ of the projection σ-field and Borel field of functions respectively.[5] The space $C_{[0,1]}$ equipped with d_U is from the topological viewpoint sufficiently like $\mathbb{R}$ that the construction of a probability space on $(C_{[0,1]}, \mathcal{B}_C)$ can follow similar lines. The actual procedure is the same as was described in the last section, constructing fidis on

[3] A functional is a function whose argument is another function. Integrals are well-known examples.

[4] A set A is dense in a space $\mathbb{S}$ if every point of $\mathbb{S}$ lies arbitrarily close to a point of A.

[5] Thus, for example, $\mathcal{B}_C = \{A \cap C : A \in \mathcal{B}_R\}$

the cylinder sets, and then extending, using the consistency theorem. The difference is that the extension takes us to $\mathcal{B}_C$, while it failed to take us to $\mathcal{B}_R$.

The chief difficulty with working in $C_{[0,1]}$ to derive FCLTs is that the functions of chief interest do not belong to it! These are, of course, of the form X_T as defined in (2.14). Note that this function is constant except at points r that satisfy $r = [Tr]/T$, at which it jumps a distance $u_{[Tr]}$. A slightly clumsy fix to this problem is provided by the following trick. Define

$$X_T^*(r) = X_T(r) + \frac{u_{[Tr]+1}(Tr - [Tr])}{\sigma_u \sqrt{T}}. \tag{3.1}$$

This is a piecewise linear function of the type just described. X_T^* is an element of $C_{[0,1]}$ for any T, and as we shall see subsequently, the extra term can be shown to be negligible, and hence ignored in convergence arguments. This is how the simplest proofs of the FCLT are set up, and is the approach we adopt below. However, recent research has tended to adopt a different approach, somewhat more technically advanced, but also more flexible and easy to generalize. We next look briefly at the arguments invoked in this approach.

3.4 The Space $D_{[0,1]}$

We found in Section 3.2 that the space R was 'too large', when endowed with the uniform metric, to allow construction of a probability distribution without running into problems. Then we showed that the space $C_{[0,1]}$ was a feasible case, but unfortunately too small for our needs, without invoking awkward constructions. A compromise is provided by the space $D_{[0,1]}$ of *cadlag* functions on [0,1]. Cadlag is a French acronym standing for 'continue à droit, limites à gauche', in other words, functions that may contain jumps, but not isolated points, such as to be discontinuous in both directions. Cadlag functions are right-continuous, and every point has a limit-point to its left. X_T in (2.14) is a good example. Its value at any point of discontinuity can be represented as the limit of a decreasing sequence of points to the right of it.

$D_{[0,1]}$ turns out to contain all the functions we are interested in, and there is no loss in excluding those cases featuring isolated discontinuities. However, $D_{[0,1]}$ is not separable under the uniform metric. The problem is that functions with discontinuities can have positive *uniform* distance from each other in spite of being equal at all but a single point. Consider, for example, the set of functions $\{x_\theta : \theta \in [0,1]\}$, where

$$x_\theta(t) = \begin{cases} 0, & t < \theta \\ 1, & t \geq \theta \end{cases}. \tag{3.2}$$

There are an uncountable number of these functions, one for each point of the unit interval, and yet in the uniform metric they are all at a distance $d_U = 1$ from each other. Thus, no subset of $(D_{[0,1]}, d_U)$ can be dense in $(D_{[0,1]}, d_U)$ yet also countable.

What this means in practice is that $\mathcal{B}_D$ with the uniform metric still contains too many sets to define a probability space. Some of its elements are nonmeasurable. As already mentioned, one possible solution is to work with the projection σ-field[6]. However, probably the most commonly adopted solution in practice is to work with a different metric, such that $D_{[0,1]}$ becomes separable. What is popularly called the *Skorokhod metric* (actually, what Skorokhod (1956) dubbed the J1 metric) is

$$d_S = \inf_{\lambda \in \Lambda} \left\{ \varepsilon > 0 : \sup_t |\lambda(t) - t| \leq \varepsilon, \sup_t |x(t) - y(\lambda(t))| \leq \varepsilon \right\}$$

[6] See Pollard (1984) for an account of this approach.

where Λ denotes the set of all increasing continuous functions $\lambda : [0,1] \mapsto [0,1]$. When functions have discontinuities, this is a more natural distance measure than d_U, because it allows functions to be compared by moving them 'sideways' as well as vertically. The functions Λ can be thought of as representing a choice of little distortions of the time domain, and we choose the one that makes the distance between x and y as small as possible in both directions. Functions, x_θ and $x_{\theta+\delta}$, jumping a distance 1 at times θ at $\theta + \delta$ respectively and otherwise constant, can now be considered at a distance δ from each other, not 1, which is what the uniform metric would give.

The Skorokhod metric defines a topology on $D_{[0,1]}$, and it is this that matters from the point of view of defining the Borel field. The key property is that the metric space $(D_{[0,1]}, d_S)$ is separable. Unfortunately, though, it is not a complete space. Consider, for given $\delta > 0$, the function $z_{\theta\delta} = x_\theta - x_{\theta+\delta}$. This function is equal to 0 on the intervals $[0,\theta)$ and $[\theta+\delta, 1]$, and to 1 on the interval $[\theta, \theta+\delta)$. Thus, consider the Cauchy sequence of functions $z_{\theta,1/n}$, for $n = 1, 2, 3, \ldots$ The limit of this sequence is equal to 1 at point θ and zero everywhere else. It has an isolated discontinuity, and so is not an element of $D_{[0,1]}$.

However, this problem can be remedied by a modification that preserves the same topology. Billingsley (1968) shows that $(D_{[0,1]}, d_B)$ is a separable complete metric space, where

$$d_B = \inf_{\lambda\in\Lambda} \left\{ \varepsilon > 0 : \|\lambda\| \le \varepsilon, \sup_t |x(t) - y(\lambda(t))| \le \varepsilon \right\}$$

with

$$\|\lambda\| = \sup_{t\neq s} \left| \log \frac{\lambda(t) - \lambda(s)}{t - s} \right|$$

and Λ is the set of increasing continuous functions λ such that $\|\lambda\| < \infty$. Note how Billingsley's definition imposes some smoothness on the choice of possible transformations of the time domain, since its slope must be kept as near 1 as possible at each point, to prevent $\|\lambda\|$ becoming large. d_B also generates the Skorokhod topology. The key consequence is, of course, that the Borel field $\mathcal{B}_D$ is equal to the projection σ-field of $(D_{[0,1]}, d_B)$. Finite dimensional projections are a determining class for distributions on $(D_{[0,1]}, \mathcal{B}_D)$ with this metric.

3.5 Brownian Motion

To conclude this discussion of function spaces and distributions defined on them, we exhibit the most important and best known example. As is well-known, the botanist Robert Brown first noted the irregular motion of pollen particles suspended in water in 1827. As we now also know, thanks to Einstein's famous 1905 paper,[7] these are due to thermal agitation of the water molecules. The mathematical model of Brownian motion was developed by Norbert Wiener (1923) and it is sometimes called *Wiener measure.*

Formally, the standard real-valued Brownian motion (BM) process on the unit interval, denoted B, is defined by the following three properties:

1. $B(r) \sim \mathrm{N}(0, r)$, $0 \le r \le 1$.

2. Increments $B(r_1) - B(0), B(r_2) - B(r_1), \ldots, B(r_N) - B(r_{N-1})$ are totally independent of each other, for any collection $0 < r_1 < \cdots < r_N$.

3. Realizations of B are continuous, with $B(0) = 0$, with probability 1.

[7] This is not the 1905 paper on special relativity, nor the Nobel-prize winning contribution of the same year on the photoelectric effect, but the third original product of Einstein's *annus mirabilis.*

This is the natural extension to function spaces of the ordinary Gaussian distribution specified in the CLT. According to property 3, we may consider it as an element of $C_{[0,1]}$ almost surely. However, since exceptions of probability zero should not be ruled out in a formal statement, even if they have no practical consequences, we should strictly treat it as an element of $D_{[0,1]}$.

BM has a number of well-known attributes following from the definition. Independence of the increments means that $E(B(t)B(s)) = \min(t,s)$ and $E[(B(t) - B(s))B(s)] = 0$ for $t > s$. BM is *self-similar*, meaning that B has the same distribution as B^* defined by

$$B^*(t) = k^{-1/2}(B(s+kt) - B(s)), \quad 0 \le t \le 1$$

for any s and k such that $0 \le s < 1$ and $0 < k \le 1 - s$. As such, realizations of BM belong to the class of curves known as *fractals* (Mandelbrot, 1983). While it is almost surely continuous, observe that for any $t \in [0,1)$ and $0 < h < 1 - t$, $B(t+h) - B(t) \sim \mathrm{N}(0,h)$ implying

$$\frac{B(t+h) - B(t)}{h} \sim \mathrm{N}(0, h^{-1}). \tag{3.3}$$

Letting $h \downarrow 0$ results in distributions with infinite variance. In other words, BM is nowhere differentiable, with probability 1. Moreover, realizations are almost surely of unbounded variation. Observe that $E|B(t+h) - B(t)| = O(T^{-1/2})$, and hence we can show

$$\sum_{j=0}^{T-1} |B((j+1)/T) - B(j/T)| \to \infty \text{ as } T \to \infty, \text{ with probability 1.}$$

4 The Functional Central Limit Theorem

The object of this section is to show that empirical processes such as X_T in (2.14) converge 'weakly' to BM, under suitable regularity conditions. The result to be ultimately proved, under rather simple assumptions, is the following, originally due to Donsker (1951).

Theorem 4.1 *Suppose that $u_t \sim iid(0, \sigma_u^2)$, and the stochastic process X_T is defined by*

$$X_T(r) = \frac{1}{\sigma_u \sqrt{T}} \sum_{t=1}^{[Tr]} u_t, \quad 0 \le r \le 1$$

Then $X_T \xrightarrow{d} B$.

Before proceeding to an actual demonstration, we must first define some terms and deal with various technical preliminaries.

4.1 Weak Convergence

The term 'weak convergence of distributions' is probably widely misunderstood. The term 'weak' here has a wholly different connotation from that in (say) 'weak law of large numbers', and derives from concepts from topology.

The fundamental problem, worth digressing into briefly at this point, is to consider what it means to say that one p.m. is close to another. In the preceding section, we have been addressing this question in respect of functions on the unit interval. The problem we now consider, while obviously related, is distinct and somewhat more abstract. We must be careful not to confuse them. Since a p.m. is, in essence, just a set of rules for assigning probabilities to sets of random objects (in the present case functions on [0,1], but they could be completely arbitrary), it is not at all obvious how to answer this question, and impose a topological structure on the problem. This is essential, however, if we are to know what we mean by 'convergence'.

For a random variable, it turns out that the characteristic function (see (2.3)) acts as a reliable 'fingerprint' for the distribution, and has a central role in proofs of the CLT. We watch what happens to the ch.f. as T increases, and see if it approaches the known normal case as in (2.5). For more general distributions, this neat trick unfortunately does not work. Instead we can consider a *class* of functions, for example, the set $\mathbb{U}$ of bounded, uniformly continuous functions whose domain is the set of random objects in question, and whose range is the real line. Note that, regardless of the underlying probability space, the expectations $E(f) = \int f d\mu$, for $f \in \mathbb{U}$, are always just finite real numbers.

Now, let $\mathbb{M}$ denote the space (i.e., the collection) of all the p.m.s μ under consideration. Given $f \in \mathbb{U}$, note that $\int f d\mu$, $\mu \in \mathbb{M}$, defines a real-valued function with domain $\mathbb{M}$. We can work with the idea that two p.m.s are close to each other if these expectations are close (in the usual Euclidean norm on $\mathbb{R}$) for a number of different $f \in \mathbb{U}$. They are considered closer, as the distances are smaller, and the number of different f for which they are close is greater. This trick can be used to show that the p.m.s inhabit a metric space on which standard notions of convergence are defined. The 'weak topology' on $\mathbb{M}$, defined by these functions, is simply the minimal collection of subsets of $\mathbb{M}$ whose images under the mappings $\int f d\mu : \mathbb{M} \mapsto \mathbb{R}$, for all $f \in \mathbb{U}$, are open sets of $\mathbb{R}$. When we speak of weak convergence, we simply mean that the definition of closeness implied by the weak topology is used as a criterion for the convergence of a sequence of p.m.s, such as that defined by increasing the sample size for partial sums, in the standard application.

In practice, weak convergence of a sequence of probability measures $\{\mu_T, T = 1, 2, 3, \cdots\}$ to a limit μ, written $\mu_T \Rightarrow \mu$, is equivalent to the condition that $\mu_T(A) \to \mu(A)$ for every set A of random objects in the probability space, such that $\mu(\delta A) = 0$, where δA denotes the boundary points of A. This is the definition already given in Section 2.1 for the case of real random variables.

4.2 Tightness

Now, it might appear that, in an informal way, we have already proved the FCLT. That parts 1 and 2 of the definition of Brownian motion are satisfied is immediate from the CLT, and the construction of the process, and we have even made a good informal case for Part 3. In fact, the CLT alone is sufficient to establish *pointwise* convergence to B. However, this by itself is not sufficient for all the applications of these results. For example, it is not sufficient to establish such results as

$$\sup_{0 \leq r \leq 1} |X_T(r)| \overset{\mathrm{d}}{\to} \sup_{0 \leq r \leq 1} |B(r)| \tag{4.1}$$

whereas this does follow from the FCLT, which establishes convergence of the function as a whole, not just finite-dimensional projections of it.

Thus, the question to be considered is whether a given sequence of p.m.s on the spaces $C_{[0,1]}$ or $D_{[0,1]}$ converges weakly to a limit in the same space. This is not a foregone conclusion. There are familiar examples of sequences within a certain space whose limits lie outside the space. Consider a sequence of rational numbers (terminating decimals) whose limit is a real number (a non-terminating decimal). In the present case, the question is whether a sequence of p.m.s defined on a space of functions ($C_{[0,1]}$ or $D_{[0,1]}$, as the case may be) has as a limit a p.m. on the same space. This is, in its essence, the issue of uniform tightness of the sequence.

In the CLT, the random elements under consideration (normalized sums) are real-valued random variables with probability 1, and the fact that the limit random variable is also distributed on the real line is something that we take as implicit. However, it is possible to construct sequences of distributions on the real line, depending on an index T, that are well defined for every finite T, yet break down in the limit. A simple example is where X_T is uniformly distributed on the interval $[-T, T]$. The c.d.f. of this distribution is

$$F_T(x) = \begin{cases} 0 & x < -T \\ (1 + x/T)/2 & -T \leq x \leq T \\ 1 & x > T \end{cases}$$

However, as $T \to \infty$, $F_T(x) \to \frac{1}{2}$ for every $x \in \mathbb{R}$, which does not define a distribution. This is a distribution that is getting "smeared out" over the whole line and, in the limit, appearing to assign positive probability mass to infinite values. Such distributions lack the attribute known as *tightness*.[8] A distribution on the line is tight if there exists, for every $\varepsilon > 0$, a finite interval having a probability exceeding $1 - \varepsilon$. Non-tight distributions are regarded as 'improper', and are not well-defined distributions according to the mathematical criteria. Here we have a case of sequences of distributions that are not *uniformly* tight, even though tight for all finite T.

While the only examples of non-tight distributions on the line are obviously pathological examples like the above, in the world of continuous or cadlag functions on the unit interval, the uniform tightness property is a real concern. In $C_{[0,1]}$ the issue becomes, in effect, one of whether the limit distribution assigns a probability arbitrarily close to 1 to some set in $C_{[0,1]}$, such that discontinuities arise with probability zero. It is not at all difficult to construct examples where the sample processes are a.s. continuous for all finite T (like X_T^* in (3.1)), yet are almost surely discontinuous in the limit. Even if the sequences we consider lie in $D_{[0,1]}$ for all finite T, like X_T in (2.14), we still need to establish that the limit lies in the same space, and is also continuous a.s., such as to correspond to BM. Proving uniform tightness of the sequence of p.m.s is what converts the pointwise convergence implied by the CLT to the FCLT proper, allowing conclusions such as (4.1).

4.3 Stochastic Equicontinuity

To set conditions ensuring that a sequence of distributions on $C_{[0,1]}$, or $D_{[0,1]}$, is uniformly tight, the natural step is to characterize compact sets of functions in these spaces, the analogues of the finite interval of the line. One can then impose tightness by ensuring a high enough probability is assigned to these compact sets in the limit. In $\mathbb{R}$, a compact set is one that is closed and bounded, and a finite interval can contain any such set. Compactness in a general topological space is a more primitive and abstract concept, but the basic idea is the same; a point cannot be the limit of a sequence of points of a compact space without itself belonging to the space. Compactness is not the same thing as completeness, which is the property that Cauchy sequences always converge in the space. Compactness implies, rather, that all sequences contain convergent subsequences with limits contained in the space.

To characterize compactness, we conventionally appeal to a well-known result on the topology of function spaces, the *Arzelà-Ascoli theorem*. For a function $x : [0, 1] \mapsto \mathbb{R}$, the *modulus of continuity* is defined as

$$w_x(\delta) = \sup_{|s-t|<\delta} |x(s) - x(t)|. \tag{4.2}$$

In other words, it is the largest change in the function over an interval of width less than δ. In a uniformly continuous function, we must have that $w_x(\delta) \to 0$ as $\delta \to 0$. According to the Arzelà-Ascoli theorem, a set of functions $A \subset C_{[0,1]}$ is relatively compact (i.e., its closure is compact) if the following two conditions hold:

[8] For another example, consider (3.3).

(a) $\sup\limits_{x \in A}|x(0)| < \infty$

(b) $\lim\limits_{\delta \to 0} \sup\limits_{x \in A} w_x(\delta) = 0.$

The space $C_{[0,1]}$ of continuous functions inevitably contains elements that are arbitrarily close to discontinuous functions, which would therefore violate condition (b), but if we confine attention to a set A satisfying the Arzelà-Ascoli conditions, we know that it is relatively compact, and therefore that its closure contains the cluster points of all sequences in the set.

In the case of $D_{[0,1]}$, the modulus of continuity has to be defined differently. It can be shown that a cadlag function on $[0,1]$ can have at most a countable set of discontinuity points. Let Π_δ denote a partition $\{t_1, \dots, t_r\}$ of the unit interval with $r \leq [1/\delta]$ and $\min_i\{t_i - t_{i-1}\} > \delta$. Then define

$$w'_x(\delta) = \inf_{\Pi_\delta}\left\{\max_{1 \leq i \leq r}\left\{\sup_{s,t \in [t_{i-1},t_i)} |x(t) - x(s)|\right\}\right\}$$

This definition modifies (4.2) by allowing the function to jump at up to r points $t_1, \dots, t_r$, so that $w'_x(\delta) \to 0$ as $\delta \to 0$ when x is cadlag. However, note that right continuity is required for this to happen, so arbitrary discontinuities are ruled out.

A sequence of functions in $C_{[0,1]}$ or $D_{[0,1]}$ that is contained in a relatively compact set is said to be *uniformly equicontinuous*, where 'uniformly' refers to uniformity with respect to the domain of the functions, and 'equicontinuous' means continuous in the limit. The Arzelà-Ascoli conditions ensure functions are uniformly equicontinuous. *Stochastic equicontinuity* (the 'uniform' qualifier being generally taken as implicit here) is a concept applying to random functions, and refers to the probability that a particular sequence of functions is relatively compact. There are several possible definitions, but it is typically sufficient to specify that this probability approaches 1 'in the tail', beyond some finite point in the sequence.

4.4 Convergence of Probability Measures

The next step in the argument is to make the formal link between compactness of sets of continuous (or cadlag) functions, and uniform tightness of probability measures on $C_{[0,1]}$ (or $D_{[0,1]}$). For reasons of space we examine these arguments, and then go on to give a proof of the FCLT, only for the case $C_{[0,1]}$. This means working with version (3.1) of the partial sum function, rather than (2.14). With certain technical modifications, the parallel arguments for $D_{[0,1]}$ are broadly similar, once stochastic equicontinuity has been defined appropriately. For details see Billingsley (1968) and also Davidson (1994), Chapter 28.

Think of a discontinuous function as having a relationship to $C_{[0,1]}$ analogous to that of the points $\pm\infty$ in relation to $\mathbb{R}$. A p.m. on $C_{[0,1]}$ that assigned positive probability to functions with discontinuities would fail to be tight, in just the same way as measures on the line that assigned probabilities exceeding $\varepsilon > 0$ to sets outside some specified finite interval, depending on ε. It will be sufficient for tightness if we can show, using the Arzelà-Ascoli conditions, that the p.m.s in question assign probabilities arbitrarily close to 1 to a compact set in $C_{[0,1]}$. This is the object of a collection of theorems due to Billingsley (1968). Letting $\{\mu_T, T = 1, 2, 3, \dots\}$ denote a sequence of probability measures on $C_{[0,1]}$, Billingsley's results for this case can be summarized as follows:

Theorem 4.2 *The sequence $\{\mu_T\}$ is uniformly tight if there exists $T^* \geq 1$ such that, for all $T > T^*$ and all $\eta > 0$,*

(a) $\exists\ M > 0$, *such that* $\mu_T(\{x : |x(0)| > M\}) \leq \eta$,

(b) *for each $\varepsilon > 0$, $\exists\ \delta \in (0,1)$ such that*

$$\mu_T(\{x : w_x(\delta) \geq \varepsilon\}) \leq \eta. \tag{4.3}$$

Moreover, (4.3) holds if

$$\sup_{0\leq t\leq 1-\delta} \mu_T\big(\{x : \sup_{t\leq s\leq t+\delta} |x(s) - x(t)|\ \geq \frac{\varepsilon}{2}\}\big)\leq \frac{\eta\delta}{2}. \tag{4.4}$$

This result formalizes, in the 'ε-δ' style of analysis, the finiteness and equicontinuity requirements needed to ensure that the limiting measure assigns probabilities at most arbitrarily close to zero, to sets containing discontinuous functions.

The fact that this is a non-trivial requirement can be appreciated by returning to the example of X_T^* in (3.1). We are depending on the random variables $u_{[Tr]+1}/\sqrt{T}$, for every $r \in [0,1]$, becoming negligible with arbitrarily high probability as T increases. The problem is that even if the probability of an extreme value of order T is very small, the number of potential 'jumps' increases with the sample size. While it may be the case that $u_{[Tr]+1}/\sqrt{T} \to 0$ in probability, for any r, there can still be circumstances in which $\sup_{0\leq r\leq 1}(u_{[Tr]+1}/\sqrt{T})$ vanishes with probability less than 1. This would lead to a failure of condition (b) in Billingsley's theorem. However, as we now show, a condition sufficient to avoid this outcome is the existence of the variance or, in other words, that $\sigma^2 < \infty$. Since this is required in any case for the pointwise convergence via the CLT, no additional restrictions on the setup are necessary.

4.5 Proof of the FCLT

To prove Theorem 4.1, there are broadly two steps involved. The first, which we take as already done, is to prove the CLT for $X_T(1)$, and hence also for $X_T(r)$ for each r. Since the increments are independent by assumption, it is an elementary application of the CLT to show that the fidis of the process – the joint distributions of all the finite collections of process coordinates $r = r_1, r_2, \dots r_M$, for any finite M – are multivariate Gaussian, with covariance matrix defined according to

$$E(X(r_i)X(r_j)) = \min(r_i, r_j).$$

If the limiting process has these finite dimensional distributions, *and* is almost surely continuous, it fulfils the definition of BM. Note that there is no issue of uniqueness of the limit to worry about here. If the sequence converges in the space at all, it can only be to BM given the facts noted above, so it suffices to ensure this happens with probability 1.

Hence, the second step is just to prove uniform tightness of the sequence of p.m.s in $C_{[0,1]}$. It is sufficient that the processes X_T^* satisfy the conditions of Theorem 4.2, being finite at the origin (trivial in this case since $X_T^*(0) = 0$ a.s.) and uniformly equicontinuous. This result depends on a well known maximal inequality for partial sums, which for sums of independent processes is known as *Kolmogorov's inequality*:

Lemma 4.1 *If $S_T = x_1 + \cdots + x_T$, where $x_1, \dots, x_T$ are i.i.d. random variables, then for $\lambda > 0$ and $p \geq 1$,*

$$P\big(\max_{1\leq k\leq T}|S_k| > \lambda\big) \leq \frac{E|S_T|^p}{\lambda^p}$$

Clearly, we may use this result to deduce that

$$P\big(\sup_{r\leq s\leq r+\delta}|X_T(s) - X_T(r)|\ > \lambda\big) \leq \frac{1}{\lambda^p}E|X_T(r+\delta) - X_T(r)|^p$$

for any chosen r and δ. Note that, since the number of terms in this partial sum is increasing with T, the pointwise CLT allows us to deduce that, for any $r \in [0, 1-\delta)$,

$$E|X_T(r+\delta) - X_T(r)|^p \to \delta^{p/2}\mu_p \tag{4.5}$$

as $T \to \infty$, where $\mu_p < \infty$ is the pth absolute moment of the standard normal distribution. Given Lemma 4.1, there is therefore a sample size T large enough (say T_1^*) that

$$P\big(\sup_{r \le s \le r+\delta} |X_T(s) - X_T(r)| > \lambda \big) \le \frac{\delta^{p/2}\mu_p}{\lambda^p},\ T > T_1^*. \tag{4.6}$$

Now, given arbitrary $\varepsilon > 0$ and $\eta > 0$, choose λ, $p > 2$ and $0 < \delta < 1$ to satisfy the inequalities

$$0 < \lambda \le \varepsilon/4,\ \frac{\delta^{p/2}\mu_p}{\lambda^p} \le \frac{\eta\delta}{4}. \tag{4.7}$$

With $p = 3$, for example, the requirement is fulfilled by setting $\lambda = \varepsilon/4$ and

$$\delta < \min\left\{ 1, \left(\frac{\varepsilon^3\eta}{64\mu_3} \right)^2 \right\}.$$

Since the same argument holds for every $r \in [0, 1-\delta]$, (4.6) and (4.7) imply that

$$\sup_{0 \le r \le 1-\delta} P\big(\sup_{r \le s \le r+\delta} |X_T(s) - X_T(r)|\ > \frac{\varepsilon}{4} \big) \le \frac{\eta\delta}{4},\ T > T_1^*. \tag{4.8}$$

Comparing with (4.4), it can be seen that we are on the way to fulfilling the equicontinuity requirement for X_T^*.

To complete the argument, refer to (3.1) and note that, for all $0 \le r < s \le 1$,

$$\begin{aligned} |X_T^*(s) - X_T(s) - X_T^*(r) + X_T(r)| &\le \frac{|u_{[Tr]+1}| + |u_{[Ts]+1}|}{\sigma\sqrt{T}} \\ &= O_p(1/\sqrt{T}). \end{aligned}$$

Therefore, it is certainly true that there exists a T large enough (say T_2^*) that

$$P\left(|X_T^*(s) - X_T(s) - X_T^*(r) + X_T(r)| \ge \frac{\varepsilon}{4} \right) \le \frac{\eta\delta}{4},\ T > T_2^*. \tag{4.9}$$

For any random variables x and y, we have

$$\begin{aligned} P(|x+y|\ > \varepsilon/2) &\le P(\{|x|\ > \varepsilon/4\} \cup \{|y|\ > \varepsilon/4\}) \\ &\le P(|x|\ > \varepsilon/4) + P(\ |y|\ > \varepsilon/4) \end{aligned}$$

where the first inequality holds because the one event is implied by the other, and the second is the sub-additive property of probabilities. Therefore, from (4.8) and (4.9) there exists $T^* = \max(T_1^*, T_2^*)$ such that, for $T > T^*$,

$$\sup_{0 \le r \le 1-\delta} P\big(\sup_{r \le s \le r+\delta} |X_T^*(s) - X_T^*(r)|\ > \frac{\varepsilon}{2} \big) \le \frac{\eta\delta}{2}.$$

Here, P implicitly denotes the probability measure relating to X_T^*, and hence inequality (4.4) is established. The sequence of distributions of the X_T^* processes is uniformly tight, and the functional central limit theorem follows.

Let's reiterate that this is only one of several possible approaches to proving Donsker's theorem. The alternative of working directly in $D_{[0,1]}$ is illustrated by the approach of Theorem 8.2 below, among others. Billingsley (1968), Section 16, gives details.

4.6 The Multivariate Case

For practical applications in econometrics, the basic FCLT for scalar processes will generally need to be extended to cover convergence of the joint distributions of sequences of vectors $\boldsymbol{x_t} = (x_{1t}, \dots, x_{mt})'$. For this purpose, it is necessary to consider the topology of the space $C^m_{[0,1]}$, which can be thought of as the Cartesian product of m copies of $C_{[0,1]}$. $C^m_{[0,1]}$ can be endowed with a metric such as

$$d^m_U(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \le j \le m} \{d_U(x_j, y_j)\}$$

and it can be shown that d^m_U induces the *product topology* – that is to say, the weak topology induced by the coordinate projections. Under the product topology, the coordinate projections are continuous. For any set $A \in C^m_{[0,1]}$, let $\pi_j(A) \in C_{[0,1]}$ be the set containing the jth coordinates of the elements of A, for $1 \le j \le m$. If $\pi_j(A)$ is open, continuity implies that A is open. The important implication of this for the present purpose is that $(C^m_{[0,1]}, d^m_U)$ is a separable space, inheriting the property from $(C_{[0,1]}, d_U)$.

It follows that the arguments deployed above for distributions on $C_{[0,1]}$ can be generalized in quite a straightforward way to deal with the vector case. The Cramér-Wold Theorem from Section 2.1 is applied to generate the fidis of the multivariate distributions, wherever required. Under suitably generalized regularity conditions for the FCLT, the limit processes are vector Brownian motions $\boldsymbol{B}$. These constitute a family of distributions with multivariate Gaussian fidis, having covariance matrices $\boldsymbol{\Omega}$ such that $E(\boldsymbol{B}(r)\boldsymbol{B}(r)') = r\boldsymbol{\Omega}$. Essentially, the approach is to show that $\boldsymbol{X}_T \to_d \boldsymbol{B}$ if $\boldsymbol{\lambda}'\boldsymbol{X}_T \to_d \boldsymbol{\lambda}'\boldsymbol{B}$ for each fixed $\boldsymbol{\lambda}$ ($m \times 1$) of unit length. Note that if each of the elements of the vector $\boldsymbol{X}_T$ is in $C_{[0,1]}$, then $\boldsymbol{\lambda}'\boldsymbol{X}_T \in C_{[0,1]}$ too, and under the FCLT the limits $\boldsymbol{\lambda}'\boldsymbol{B}$ are scalar BMs with variances $\boldsymbol{\lambda}'\boldsymbol{\Omega}\boldsymbol{\lambda}$.

One very important feature of the multivariate FCLT is the fact that, because the limit process is Gaussian, dependence between the coordinates is completely represented by $\boldsymbol{\Omega}$ in the limit. This means that even if the distribution of $\boldsymbol{x_t}$ features arbitrary forms of dependence between the coordinates, linear projections will nonetheless suffice to induce independence in large samples. Thus, if $\boldsymbol{x_t} = (y_t, \boldsymbol{z}'_t)'$ with corresponding limit process $\boldsymbol{B}_x = (B_y, \boldsymbol{B}'_z)'$, then $\boldsymbol{B}_z$ ($m-1 \times 1$) and $B_{y|z} = B_y - \boldsymbol{B}'_z\boldsymbol{\Omega}^{-1}_{zz}\boldsymbol{\Omega}_{zy}$ (defining the obvious partition of $\boldsymbol{\Omega}$) are independent BMs, a property invariant to the joint distribution of y_t and $\boldsymbol{z}_t$ in finite samples.

In principle, the same type of arguments can be adapted to vectors of cadlag processes. Referring to the definitions in Section 3.4, consider the metric space $(D^m_{[0,1]}, d^m_B)$ where

$$d^m_B(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \le j \le m} \{d_B(x_j, y_j)\}.$$

This metric induces the Skorokhod topology in suitably generalized form, and the space is separable and complete. There is just one potential pitfall deserving mention in this case. In the case of continuous functions, we noted that linear combinations of continuous functions are continuous, and further that a sequence $\{\boldsymbol{\lambda}'\boldsymbol{X}_T\}$ has its limit in C provided this is true of each coordinate of $\boldsymbol{X}_T$. However, the analogous property does not hold for elements of $(D^m_{[0,1]}, d^m_B)$. Consider for example a vector $(X_{1T}, X_{2T})'$ where $X_{1T} = x_\theta$ as defined in (3.2), for every T, whereas $X_{2T} = x_{\theta+1/T}$. Both $\{X_{1T}\}$ and $\{X_{2T}\}$ converge in $(D_{[0,1]}, d_B)$ (the former, trivially) but the limit of the sequence $\{X_{2T} - X_{1T}\}$ is not in $D_{[0,1]}$, featuring an isolated discontinuity at θ. Under the d_B metric, this is not a Cauchy sequence. Hence, convergence of the marginal distributions of individual functions in $D_{[0,1]}$ does not imply their joint convergence, without further restrictions. However, it suffices for the limit vector to lie in $C^m_{[0,1]}$ with probability 1. Therefore, proofs of the multivariate FCLT can be set in the cadlag framework, just like their scalar counterparts. More details on all these issues can be found in Davidson (1994) Chapters 6.5, 27.7 and 29.5

4.7 Functionals of B

The functional central limit theorem has to be supplemented by the continuous mapping theorem (CMT) to have useful applications in statistical inference. The CMT was stated for real r.v.s in Section 2.1, but this is just a case of a much more general result. For present purposes, the required variant is as follows.

> *Continuous mapping theorem for random functions.*
> Let $h : D_{[0,1]} \mapsto \mathbb{R}$ be a measurable mapping from elements of $D_{[0,1]}$ to points of the real line. If the mapping is continuous except at points of the domain with zero probability under the distribution of B, and $X_T \xrightarrow{\mathrm{d}} B$, then $h(X_T) \xrightarrow{\mathrm{d}} h(B)$.

For example, consider the integral

$$\int_0^1 X_T(r)dr = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sigma_u\sqrt{T}}x_t = \frac{1}{\sigma_u T^{3/2}}\sum_{t=1}^{T}\sum_{s=1}^{t}u_s \tag{4.10}$$

The FCLT and CMT allow us to conclude that this random variable has a Gaussian limit. One can easily show (apply the ordinary CLT for heterogeneous sequences after re-arranging the double sum) that it is distributed as $\mathrm{N}(0, 1/3)$. However, the foregoing argument shows that it can also be written as $\int_0^1 B(r)dr$.

The same approach allows us to establish that

$$\int_0^1 X_T(r)^2 dr = \frac{1}{\sigma_u^2 T^2}\sum_{t=1}^{T}x_t^2 \xrightarrow{\mathrm{d}} \int_0^1 B(r)^2 dr$$

which is, however, a random variable without a representation in terms of known distributions. The FCLT and CMT simply assure us that all squared partial sum processes with increments satisfying the regularity conditions converge in distribution to the same limit. This is what we call an *invariance principle*, because the limit is invariant to the distribution of the statistic in finite samples.

Returning to the original problem of model (2.9), we can now consider the distribution of $T(\hat{\lambda} - 1)$, which is better known as the Dickey-Fuller statistic of the first type. A further result required is the distribution of

$$\frac{1}{T}\sum_{t=1}^{T}x_{t-1}u_t \tag{4.11}$$

but this can be obtained by a neat trick. Note that

$$2x_{t-1}u_t = x_t^2 - x_{t-1}^2 - u_t^2 \tag{4.12}$$

and hence, if $u_t \sim iid(0, \sigma_u^2)$ and $x_0 = 0$, then

$$\begin{aligned}\frac{1}{\sigma_u^2 T}\sum_{t=1}^{T}x_{t-1}u_t &= \frac{1}{2\sigma_u^2 T}\left(x_T^2 - \sum_{t=1}^{T}u_t^2\right)\\ &\xrightarrow{\mathrm{d}} \tfrac{1}{2}(B(1)^2 - 1)\\ &\sim \tfrac{1}{2}(\chi^2(1) - 1).\end{aligned} \tag{4.13}$$

Since the variance cancels in the ratio, the CMT now allows us to conclude that

$$T(\hat{\lambda} - 1) \xrightarrow{\mathrm{d}} \frac{B(1)^2 - 1}{2\int_0^1 B(r)^2 dr}$$

which is the well known formula for the limiting distribution of the Dickey-Fuller statistic without mean or trend correction. However, it is important to note that the steps in (4.13) do *not* generalize to other models, in particular where the u_t are not identical with the increments of x_t. There is no useful vector generalization of (4.12). The problem of general stochastic integrals is treated in Section 7.

Using the FCLT and CMT, a number of other important processes related to BM are easily shown to be the weak limits of sample processes. We consider just two simple examples. Defining

$$S_t = \sum_{t=1}^{t} u_t$$

consider the mean deviation process

$$x_t = S_t - \bar{S}. \tag{4.14}$$

where $\bar{S} = T^{-1}\sum_{s=1}^{T} S_s$. If $X_T(r) = x_{[Tr]}/(\sigma\sqrt{T})$, it is easily shown using (4.10) that

$$X_T \xrightarrow{\text{d}} B - \int_0^1 Bds$$

where the limit process is called a *de-meaned* BM.

However, be careful to note that de-meaning is not the appropriate way to account for an intercept in the generating equation. Suppose that, instead of (2.9), we have

$$\begin{aligned} x_t &= \alpha + x_{t-1} + u_t. \\ &= S_t + t\alpha. \end{aligned} \tag{4.15}$$

The intercept induces a deterministic trend that dominates the stochastic trend in the limit, and the normalization of $T^{-1/2}$ is not appropriate. Instead, divide by T to obtain the limit

$$X_T(r) = T^{-1}x_{[Tr]} \xrightarrow{\text{pr}} \alpha r.$$

However, the stochastic component can be isolated by regressing x_t onto the trend dummy. Let the residuals from this regression (including an intercept) be represented as

$$\begin{aligned} x_t^* &= x_t - \bar{x} - (t - \bar{t})\frac{\sum_{s=1}^{T}(s-\bar{t})x_s}{\sum_{s=1}^{T}(s-\bar{t})^2} \\ &= S_t - \bar{S} - (t - \bar{t})\frac{\sum_{s=1}^{T}(s-\bar{t})S_s}{\sum_{s=1}^{T}(s-\bar{t})^2} \end{aligned}$$

where $\bar{t} = T^{-1}\sum_{t=1}^{T} t = (T+1)/2$, and the second equality follows directly on substitution from (4.15). Noting that $([Tr] - \bar{t})/T \to r - \frac{1}{2}$ and that $T^{-3}\sum_{t=1}^{T}(t-\bar{t})^2 \to 1/12$, we can therefore use the CMT to show that

$$\begin{aligned} X_T(r) &= x^*_{[Tr]}/(\sigma_u\sqrt{T}) \\ &\xrightarrow{\text{d}} B(r) - \int_0^1 B(s)ds - 12(r - \tfrac{1}{2})\int_0^1 (s - \tfrac{1}{2})B(s)ds \end{aligned}$$

This process is called a de-trended BM.

Whereas (4.14) puts the partial sum process into mean deviation form, putting the increments of the partial sums into mean deviation form yields a quite different limit. Consider the array

$$x_{Tt} = \sum_{s=1}^{t}(u_s - \bar{u}), \quad t = 1, \dots, T$$

where $\bar{u} = T^{-1}\sum_{s=1}^{T} u_t$. This process has the property that $x_{TT} = 0$ identically. If $X_T(r) = x_{T,[Tr]}/(\sigma_u\sqrt{T})$, it is again easy to see using the FCLT and CMT that $X_T \xrightarrow{\mathrm{d}} B^o$, where

$$B^o(r) = B(r) - rB(1), \quad 0 \leq r \leq 1$$

B^o is called a *Brownian bridge*, having the property $B^o(1) = B^0(0) = 0$. It does not have independent increments, and $E(B^o(t)B^o(s)) = \min\{t,s\} - ts$.

5 Dependent Increments

5.1 Martingale Differences

We have assumed for simplicity up to now that the driving sequence $\{u_t\}$ is independent and identically distributed. In practice, the same results can be obtained under a variety of weaker assumptions about the dependence. One easy extension is to allow $\{u_t\}$ to be a stationary martingale difference (m.d.) sequence; in other words, an integrable process having the property

$$E(u_t|\mathcal{F}_{t-1}) = 0 \text{ a.s.}$$

where $\{\mathcal{F}_t\}$ is a nested sequence of σ-fields such that u_t is $\mathcal{F}_t$-measurable (the pairs $\{u_t, \mathcal{F}_t\}$ are said to form an *adapted* sequence). Intuitively, an m.d. is a process that is unpredictable in mean, one step ahead.

The two main planks in our FCLT proof that used the independence assumption, the CLT and the Kolmogorov maximal inequality, can both be extended to the m.d. case without additional assumptions.[9] To verify that the fidis are converging to those of BM, we use the martingale CLT and the fact that m.d.s are uncorrelated sequences. The key assumption that $E(X_T(r)^2) = r$ for $0 \leq r \leq 1$ follows from this property, and the increments of X_T are uncorrelated. They are not independent in finite samples, in general, but because uncorrelated *Gaussian* increments are independent of each other, the increments are *asymptotically* independent. Hence, the fidis match all the requirements. With the modified maximal inequality for m.d.s, the FCLT proof given above goes through otherwise unchanged.

This result is of particular importance in econometrics, since the assumption of independence is strong, and much less likely to hold in economic data than the m.d. assumption. For example, considering the case where $u_t = \boldsymbol{y}_t - \boldsymbol{\beta}'\boldsymbol{x}_t$, the disturbances in a regression model, the only assumption needed to make u_t an m.d. is that $E(y_t|\mathcal{I}_t) = \boldsymbol{\beta}'\boldsymbol{x}_t$ where $\mathcal{F}_{t-1} \subset \mathcal{I}_t \subset \mathcal{F}_t$ and $\boldsymbol{x}_t$ is $\mathcal{I}_t$-measurable. These assumptions are plausibly fulfilled in a correctly specified regression model, and they do not rule out such phenomena as stationary ARCH or GARCH (predictability in variance) provided there is no predictability in mean. By contrast, there are rarely firm grounds to assume that the disturbances in a time series regression are truly independent. Moreover, the m.d. property is inherited by $\boldsymbol{x}_t u_t$ under the assumption given, which is very useful for establishing asymptotic normality in the general setup described in Section 2.2.

5.2 General Dependence

Of course, we would like the FCLT to hold in more general cases still, where the increments are autocorrelated, and generally dependent. The reason this is especially important is that the FCLT is used to construct the distributions of statistics involving observed integrated processes. In stationary data, it is sufficient for the asymptotic normality of regression coefficients if the

[9] See Davidson (1994), respectively Theorem 24.3 and Theorem 15.14, for these results.

disturbances satisfy the m.d. assumption, as detailed in the previous section, since only a LLN is needed to hold for the regressor variables. By contrast, the FCLT must hold for all the regressors in an I(1) modelling setup. To illustrate with a familiar case, the Dickey-Fuller test is customarily used to test whether measured time series are I(1) or I(0). While there is an 'augmented' variant of the test for data with correlated increments, the sole function of the augmenting correction is to estimate the increment variance consistently. The FCLT is required to hold with correlated increments, in this case.

In essence the sufficient conditions are twofold. First, it is necessary that the increments have at least finite variances, with a slightly stronger moment condition in cases when the marginal distributions are heterogeneous. Second, the processes must satisfy a short-memory condition. Memory conditions are, first and foremost, conditions on the autocorrelation function of the process. When the increments are covariance stationary, note that

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}E\left(\sum_{t=1}^{T}u_t\right)^2 &= \sum_{j=-\infty}^{\infty}E(u_1u_{1-j}) \\ &= \sigma_u^2 + 2\lambda_{uu} = \omega_{uu} \qquad (5.1)\end{aligned}$$

where $\lambda_{uu} = \sum_{j=1}^{\infty}E(u_1u_{1-j})$. Accordingly, $\omega_{uu}^{1/2}$ has to replace σ_u as the normalization to make the limit process a standard BM, and ω_{uu} must be both finite and positive. As is well known, this is equivalent to the condition that the spectral density is bounded away from both infinity and zero at the origin. Summability of the autocovariances is the property sometimes called weak dependence.[10]

However, limited autocorrelations cannot supply a sufficient condition for the CLT and FCLT to hold unless the process in question is jointly Gaussian, such that the dependence is fully represented by the autocorrelation function. In non-Gaussian and nonlinear processes (the best-known examples of the latter are probably ARCH/GARCH processes), there has to be some limitation on the nonlinear dependence as well. The term 'I(0)'[11] is sometimes defined loosely to mean a stationary process (such that $E(u_tu_{t-j})$ is both finite, and independent of t, for every j) and also, more properly, to mean a stationary weakly dependent process (such that the covariances are summable). It would both be more consistent, and avoid complications arising with nonlinear dependence, to use 'I(0)' to refer simply to processes whose normalized partial sums converge weakly to BM. Then, we should note that non-stationarity (of a local variety) is not necessary, but also that 'weak dependence' is a sufficient condition only in Gaussian and linear cases.

5.3 Linear Processes

Linear processes, including ARMA processes of finite or infinite order, are a common assumption in time series modelling. The one-sided moving average form, which includes (possibly after solving out) all the cases of interest, is

$$u_t = \sum_{j=0}^{\infty}\theta_j\varepsilon_{t-j} = \theta(L)\varepsilon_t \qquad (5.2)$$

[10] This is (unfortunately) the third distinct usage of the word 'weak' in this literature, with a different connotation from both 'weak LLN' and 'weak convergence'.

[11] This is in the familiar context in which I(1) denotes an integrated process, where the argument represents the order of difference necessary to get I(0).

where $\{\varepsilon_t\}$ is the driving process, often assumed i.i.d.$(0, \sigma^2)$, $\{\theta_j\}_{j=0}^{\infty}$ is a sequence of constant coefficients with $\theta_0 = 1$, and $\theta(L) = \sum_{j=0}^{\infty} \theta_j L^j$, where L denotes the lag operator. A weaker assumption that will often serve as well as serial independence is that $\{\varepsilon_t\}$ is a stationary m.d. The convenience of linear models is that stationarity of u_t then holds by construction and, in effect, the dependence is solely a property of the coefficients θ_j. The scheme is sometimes justified by an appeal to Wold's Theorem (Wold 1938), which states that every stationary nondeterministic process can be put in the form of (5.2), where the process $\{\varepsilon_t\}$ is stationary and *uncorrelated*. However, this uncorrelatedness is equivalent to neither the i.i.d. nor the stationary m.d. assumptions, unless the process is also Gaussian. This is itself a rather strong assumption that fails in many economic and financial data sets.

If the dependence is linear, there is a technically easy way to modify the application of the FCLT using the *Beveridge-Nelson* (1981) *decomposition.* This idea is examined in Phillips and Solo (1992). Considering again the model $x_t = x_{t-1} + u_t$, where now u_t is given by (5.2), the easily verified identity

$$\theta(L) = \theta(1) + \theta^*(L)(1 - L)$$

where $\theta_j^* = -\sum_{i=j+1}^{\infty} \theta_i$, may be used to re-order the summation as

$$u_t = \theta(1)\varepsilon_t + \theta^*(L)\Delta\varepsilon_t.$$

Hence, letting $z_t = \sum_{s=1}^{t} \varepsilon_s$,

$$x_t = \sum_{s=1}^{t} u_t = \theta(1)z_t + \theta^*(L)(\varepsilon_t - \varepsilon_0).$$

Now write the normalized partial sum process as

$$X_T(r) = \theta(1)Z_T(r) + \frac{\theta^*(L)(\varepsilon_{[Tr]} - \varepsilon_0)}{\sigma_u \sqrt{T}}.$$

Provided the second right-hand side term is of small order, it can be neglected in calculations involving the limit distribution. A sufficient condition is that the sequence $\{\theta_j^*\}$ be absolutely summable, implying that the sequence $\{\theta^*(L)\varepsilon_t\}$ is 'short memory'. Using elementary results on summability,[12] this is true if $\theta_j^* = O(j^{-1-\mu})$ for $\mu > 0$, and hence if $\theta_j = O(j^{-2-\mu})$. Sequences with the latter property are sometimes said to be '1-summable', since $\sum_{j=1}^{\infty} j|\theta_j| < \infty$.

5.4 Mixing

When the linear process assumption does not serve, mixing conditions are probably the best-known type of restriction for limiting serial dependence. There are several variants, of which the most often cited are probably strong mixing (α-mixing) and uniform mixing (ϕ-mixing). To define the mixing concept, begin by introducing the notation $\mathcal{F}_{t_1}^{t_2} \subset \mathcal{F}$ for $t_1 \leq t_2$, for a σ-field representing 'information dated $s \in [t_1, t_2]$'. If $\mathcal{F}$ represents the collection of events relating to the complete history of a stochastic sequence $\{x_t, \, -\infty < t < +\infty\}$, we also sometimes write $\mathcal{F}_{t_1}^{t_2} = \sigma(x_{t_1}, ..., x_{t_2})$ to show that this is the σ-field 'generated by' this segment of the sequence, although it is also possible for the set to contain additional information relating to these dates. Then, $\mathcal{F}_{-\infty}^{t}$ represents 'history up to date t', and $\mathcal{F}_{t+m}^{+\infty}$ 'events from date $t + m$ onwards'. With this notation, the mixing coefficients are defined respectively as

$$\alpha_m = \sup_t \sup_{A \in \mathcal{F}_{-\infty}^{t}, B \in \mathcal{F}_{t+m}^{+\infty}} |P(A \cap B) - P(A)P(B)|$$

[12] If $\theta_j = O(j^{-1-\mu})$ for $\mu > 0$, then $\sum_{j=0}^{\infty} |\theta_i| < \infty$ and $\sum_{i=j+1}^{\infty} |\theta_i| = O(j^{-\mu})$.

$$\phi_m = \sup_t \sup_{A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{t+m}^{+\infty}, P(B)>0} |P(A|B) - P(A)|$$

for $m = 1, 2, 3, ...$ The sequences $\{\alpha_m\}_0^\infty$ and $\{\phi_m\}_0^\infty$ are alternative measures of the rate at which remote parts of the sequence are becoming independent of each other as the gap increases. A random sequence is said to be strong (resp. uniform) mixing of size $-\lambda_0$ if $\alpha_m = O(m^{-\lambda})$ (resp. $\phi_m = O(m^{-\lambda})$) for $\lambda > \lambda_0$. It is not difficult to show that $\phi_m \geq \alpha_m$, so that 'strong' mixing is actually the weaker of the two concepts.

This approach to quantifying dependence has the advantage of being completely nonparametric, and well defined whatever the actual dynamic generation mechanism of the data. However, it has some drawbacks. An obvious one is to be able to actually verify from the structure of a model that the condition is satisfied, rather than just assuming it. However, the chief problem is that, because we are 'sup'ing over all possible pairs of remote events, this necessarily embraces any odd and pathological cases that may exist. It is well-known (Andrews 1984) that even a stable autoregression (which depends on the whole history of past shocks, albeit with weights converging to zero exponentially fast) can violate the strong mixing condition. Andrews' counter-example involves an AR(1) with a discrete (Bernoulli) innovation sequence, but to derive mild sufficient conditions for mixing in linear processes is nonetheless surprisingly difficult. (See Davidson 1994, Section 14.5). Not merely a continuous shock distribution, but some kind of smoothness condition on the density, appears unavoidable.

5.5 Near-Epoch Dependence

Because of the limitations of mixing assumptions they can, with advantage, be combined with or substituted by another condition called near-epoch dependence. This one has slightly more structure, in that it explicitly represents a random sequence as depending on past and/or future values of some underlying sequence (that could be mixing, or possibly independent) and also calls for the existence of moments up to some order such as 2. For a chosen $p > 0$, the condition takes the general form

$$\sqrt[p]{E|u_t - E(u_t|\mathcal{F}_{t-m}^{t+m})|^p} \leq d_t \nu_m \tag{5.3}$$

where d_t is a sequence representing scale factors in a heterogeneous sequence (e.g. $d_t = (E|u_t|^p)^{1/p}$). We say that the sequence is L_p-NED of size $-\mu_0$ if $\nu_m = O(m^{-\mu})$ for $\mu > \mu_0$.

$E(u_t|\mathcal{F}_{t-m}^{t+m})$ represents the best prediction of u_t based on information from the 'near epoch'. If u_t is a short-memory process, then it is mainly determined by the near epoch, and ν_m should diminish rapidly with m. Note that NED is not a restriction on the process memory as such, but rather a condition on the mapping from the underlying driving process generating $\{\mathcal{F}_s^t\}$ to the observed process. In practice, it is applied by specifying that the underlying process is mixing, or possibly independent. Its advantage is that it avoids the odd counter-examples that make the mixing condition over-sensitive in many situations. For example, the AR(1) case with Bernoulli innovations cited in Section 5.4 is certainly L_2-NED on the independent shock process. For additional background, see Gallant and White (1988) and Davidson (1994).

The way these conditions can be applied is illustrated by the following FCLT for processes with dependent increments, taken from Davidson (2002). This result also departs from the stationarity assumption to show how, in particular, processes with heterogeneous variances might be handled. We might, for example, have a situation where the variances changed according to a seasonal pattern. Most situations except trending variances (dealt with in Section 6.1) can be handled in this way. Similarly, it is not assumed that $E(u_t) = 0$ for every t, only that a suitable mean correction is made to each observation

Theorem 5.1 *Let $X_T : [0,1] \mapsto \mathbb{R}$ be defined by*

$$X_T(r) = \omega_T^{-1/2} \sum_{t=1}^{[Tr]} (u_t - Eu_t) \quad 0 < r \leq 1 \tag{5.4}$$

where $\omega_T = \mathrm{Var}(\sum_{t=1}^T u_t)$. Let the following assumptions hold:

(a) *u_t is L_2-NED of size $-\frac{1}{2}$ on a process $\{\boldsymbol{v}_s\}$ with respect to constants $d_t \leq E(|u_t|^r)^{1/r}$, where $\boldsymbol{v}_s$ is either α-mixing of size $-r/(r-2)$ for $r > 2$ or ϕ-mixing of size $-r/(2r-2)$, for $r \geq 2$.*

(b) *$\sup_t E|u_t - Eu_t|^r < \infty$ for r defined in (a), and if $r = 2$ then $\{(u_t - Eu_t)^2\}$ is uniformly integrable.*

(c) *$\dfrac{\omega_T}{T} \to \omega_u > 0$, as $T \to \infty$.*

Then, $X_T \xrightarrow{\mathrm{d}} B$.

Note that being able to assert ϕ-mixing allows more relaxed size and moment conditions than otherwise. The special feature of this result, which distinguishes it from sufficient conditions for the CLT, is condition (c). This is sufficient to ensure that $E(X_T(r)^2) \to r$ as $T \to \infty$ for $0 < r \leq 1$, as is required for BM to be the limit process. By contrast, the ordinary CLT specifies only the conditions for convergence for $r = 1$. Various kinds of heterogeneity of the variance sequence are compatible with this case, including trending variances as in (2.6), but not with condition (c).

5.6 More on Linear Forms

It is of interest to contrast linearity with the property of near-epoch dependence on an independent shock process. It is easy to show direct from (5.3) (see Davidson 1994, Example17.3) that, if (5.2) describes the relationship between u_t and a driving process $\{\varepsilon_t\}$, then

$$\nu_m = O\big(\textstyle\sum_{j=m+1}^\infty |\theta_j|\big)$$

in (5.3). The usual summability arguments yield the requirement $\theta_j = O(j^{-3/2-\mu})$ for $\mu > 0$, in order for the L_p-NED size of a linear process to be $-\frac{1}{2}$, and so satisfy the condition of Theorem 5.1. Note that this is a weaker condition than that obtained for the FCLT using the Beveridge-Nelson decomposition, without even taking into account that the driving process can itself be generally dependent in this case. The i.i.d. or m.d. assumptions imposed in (5.2) can be replaced by a mixing assumption. For example, with Gaussian increments (letting $r \to \infty$) the α-mixing size of $-\frac{1}{2}$ is permitted, and ϕ-mixing size of -1 is compatible with no moments beyond the variance.

However, the virtue of linear forms, in a more general context, is that the amount of dependence compatible with weak convergence to Brownian motion can be significantly greater than in the purely nonparametric setting. Moving averages of processes satisfying the conditions of Theorem 5.1 yield the same limit result, under quite weak restrictions on the coefficients. This result, from Davidson (2002), goes as follows.

Theorem 5.2 *Let $X_T : [0,1] \mapsto \mathbb{R}$ be defined by*

$$X_T(\xi) = \sigma_T^{-1} \sum_{t=1}^{[T\xi]} (u_t - Eu_t) \quad 0 < \xi \leq 1$$

where $\sigma_T^2 = \mathrm{Var}(\sum_{t=1}^T u_t)$*, and let*

$$u_t = \sum_{j=0}^{\infty} \theta_j v_{t-j}$$

where

(a) v_t *satisfies conditions (a), (b) and (c) of Theorem 5.1,*

(b) *the sequence* $\{\theta_j\}$ *is regularly varying at* ∞ *and satisfies the conditions*

$$0 < \left| \sum_{j=0}^{\infty} \theta_j \right| < \infty, \tag{5.5}$$

$$\sum_{k=0}^{\infty} \left(\sum_{j=1+k}^{T+k} \theta_j \right)^2 = o(T). \tag{5.6}$$

Then, $X_T \xrightarrow{\mathrm{d}} B$*.*

The first striking thing about this result is that the summability specified in condition (b) is weaker than absolute summability. If the coefficients change sign, they can take substantially longer to decay absolutely than if they are all (say) positive. In fact, the only restriction placed on their absolute rate of decay is (5.6), which enforces square summability. The second striking feature, as we show in Section 6.3, is that these come close to being necessary conditions. We exhibit a case where (b) is violated, and show (under supplementary assumptions) that a limit distribution different from B is obtained.

6 Processes Related to Brownian Motion

6.1 Variance-Transformed BM

The results given in Sections 5.5 and 5.6 allowed some nonstationarity of the distributions of the $\{u_t\}$ sequence, provided this was of a local sort that would be averaged out in the limit. Cases where the nonstationarity is not averaged out, and therefore affects the limit distribution, are said to be globally nonstationary.

Thus, suppose the variance of the driving process is growing, or shrinking, with time. Specifically, suppose that $E(u_t^2) = (1+\alpha)t^\alpha \sigma_u^2$ as in (2.6), where $\alpha > -1$. Assume serial independence, so that $\omega_T = \mathrm{Var}(\sum_{t=1}^T u_t) \approx T^{1+\alpha}\sigma_u^2$. Then, defining

$$X_T(r) = \frac{1}{\omega_T^{1/2}} \sum_{t=1}^{[Tr]} u_t$$

note that

$$E(X_T(r)^2) \to r^{1+\alpha}. \tag{6.1}$$

Using a version of the CLT for trending-variance processes, as in (2.6), it is possible to show that the fidis of the limit process are those of

$$B_\alpha(r) = B(r^{1+\alpha}).$$

Stochastic equicontinuity holds for processes of this type, leading to an FCLT. More generally, if $\omega_{[Tr]}/\omega_T \to \eta(r)$, where $\eta(\cdot)$ is any increasing homeomorphism on $[0,1]$ and the other FCLT assumptions hold, the limit process takes the form

$$B_\eta(r) = B(\eta(r)).$$

For these results see White and Wooldridge (1988), Davidson (1994) and de Jong and Davidson (2000). Note that these processes are not BM, although they are still a.s. continuous Gaussian processes having independent increments. They can be thought of as BMs that have been subjected to some lateral distortion, through stretching or squeezing of the time domain.

6.2 Ornstein-Uhlenbeck Process

Consider a process X defined by

$$X(r) = \frac{B(e^{2\beta r})}{\sqrt{2\beta} e^{\beta r}}, \quad 0 \le r \le 1 \tag{6.2}$$

for $\beta > 0$, where if the domain of X is $[0,1]$ as we have assumed, then the domain of B must in this case be $[0, e^{2\beta}]$. It can be verified that $E(X(r)X(s)) = 1/(2\beta e^{\beta|r-s|})$, depending only on $|r-s|$, and hence the process is stationary. It is also easily verified that the increments are negatively correlated.

Considering a time interval $[r, r+\delta]$, observe that

$$\begin{aligned}\frac{E[B(e^{2\beta(r+\delta)}) - B(e^{2\beta r})]^2}{2\beta e^{2\beta r}} &= \frac{e^{2\beta\delta}-1}{2\beta} \\ &= E[B(r+\delta) - B(r)]^2 + o(\delta).\end{aligned}$$

Letting $B(r+d) - B(r)$ approach $dB(r)$ as $\delta \to 0$, $X(r)$ can evidently be viewed, for r large enough, as a solution of the stochastic differential equation

$$dX(r) = -\beta X(r)dr + dB(r).$$

The conventional solution of this equation yields the alternative representations

$$X(r) = \int_0^r e^{-\beta(r-s)} dB = B(r) - \beta \int_0^r e^{-\beta(r-s)} B(s) ds \tag{6.3}$$

where the second version of the formula is obtained by integration by parts. This is the Ornstein-Uhlenbeck (OU) process. Letting β tend to 0 yields ordinary BM in the limit according to the last formula, although note that the stationary representation of (6.2) is accordingly breaking down as $\beta \downarrow 0$.

The OU process is the weak limit of an autoregressive process with a root local to unity, that is,

$$x_t = (1 - \beta/T)x_{t-1} + u_t.$$

To show this, write for brevity $\lambda_T = 1 - \beta/T$, and using the identity

$$\lambda_T^k - 1 = (\lambda_T - 1)\sum_{j=0}^{k-1} \lambda_T^j$$

note that

$$x_t = \sum_{j=1}^{t} \lambda_T^{t-j} u_j$$

$$
\begin{aligned}
&= \sum_{j=1}^{t-1} [\lambda_T^{t-j} - 1] u_j + \sum_{j=1}^{t} u_j \\
&= (\lambda_T - 1) \sum_{j=1}^{t-1} \sum_{k=0}^{t-j-1} \lambda_T^k u_j + \sum_{j=1}^{t} u_j \\
&= (\lambda_T - 1) \sum_{k=1}^{t-1} \lambda_T^{t-k-1} \sum_{j=1}^{k} u_j + \sum_{j=1}^{t} u_j .
\end{aligned}
$$

Substituting for λ_T and also noting that $(1 - \beta/T)^{[T\delta]} = e^{-\beta\delta} + o(1)$, the Brownian FCLT and CMT then yield

$$
\begin{aligned}
\frac{1}{\sqrt{T}} x_{[Tr]} &= \frac{1}{\sqrt{T}} \sum_{j=1}^{[Tr]} u_j - \beta \frac{1}{T} \sum_{j=1}^{[Tr]-1} e^{-\beta([Tr]-j)} \frac{1}{\sqrt{T}} \sum_{k=1}^{j} u_j \\
&\xrightarrow{d} B(r) - \beta \int_0^r e^{-\beta(r-s)} B(s) ds
\end{aligned}
$$

6.3 Fractional Brownian Motion

Consider the *fractionally integrated* process defined by $x_t = x_{t-1} + (1-L)^{-d} u_t$ for $-\frac{1}{2} < d < \frac{1}{2}$, where

$$
(1-L)^{-d} = \sum_{j=0}^{\infty} b_j L^j \tag{6.4}
$$

in which

$$
b_j = \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)}.
$$

The justification for the notation $(1-L)^{-d}$ follows from calculating the infinite order binomial expansion and verifying that the coefficients match the b_j. For the case $d > 0$, these increment processes are called 'long memory' because the MA lag weights (and hence the autocorrelations) are nonsummable. In the case $d < 0$, the increments are called 'anti-persistent' and feature the special property that the MA lag weights sum to 0, indicating a generalized form of over-differencing.

The summations may be rearranged to obtain

$$
x_t = \sum_{k=1}^{t} \sum_{j=0}^{\infty} b_j u_{k-j} = \sum_{s=-\infty}^{t} a_{ts} u_s \tag{6.5}
$$

where

$$
a_{ts} = \sum_{j=\max\{0,1-s\}}^{t-s} b_j .
$$

From this representation, it can be shown by direct calculation that $T^{-1-2d} E(x_T^2) \to \omega_{uu} V_d$ where V_d is a scale constant, to be defined. This suggests defining the empirical process

$$
X_T(r) = \frac{x_{[Tr]}}{\omega_{uu}^{1/2} \sqrt{V_d} T^{1/2+d}} .
$$

It can then be further verified that

$$
E(X_T(r))^2 \to r^{1+2d}
$$

and

$$E(X_T(r+\delta) - X_T(r))(X_T(r)) \to \tfrac{1}{2}[(r+\delta)^{1+2d} - r^{1+2d} - \delta^{1+2d}].$$

Thus, this process has a different convergence rate, similar to the trending variance case in (6.1), but unlike that case, it has dependent increments, even in the limit. It can be verified that the serial correlation is positive when $d > 0$ and negative when $d < 0$.

To establish a weak limit for X_T, the first step is to identify a candidate limit process. One sharing the same correlation structure is *fractional* Brownian motion (fBM) defined by Mandelbrot and Van Ness (1968). Denote fBM by X, where

$$X(r) = \frac{1}{\Gamma(d+1)}\left(\int_{-\infty}^{r}(r-s)^d dB - \int_{-\infty}^{0}(-s)^d dB\right),\ r \geq 0 \qquad (6.6)$$

and B is regular BM. It can be verified that

$$\begin{aligned} E(X(1)^2) &= \frac{1}{\Gamma(d+1)^2}\left(\frac{1}{2d+1} + \int_0^\infty [(1+\tau)^d - \tau^d)]^2 d\tau\right) \\ &= \frac{\Gamma(1-2d)}{(2d+1)\Gamma(1-d)\Gamma(1+d)} \end{aligned}$$

and this constant[13] is equated with V_d defined above. To complete the proof that fBM is the weak limit of the fractionally integrated process, it remains to show that the fidis of X_T are Gaussian, and that the distributions are uniformly tight. These results can be established using the approach of Davydov (1970). The basic step is to note that the fidis are those of weighted sums of the driving process $\{u_t\}$ as in (6.5). The re-ordered summation converts an issue of long memory into one of heteroscedasticity. The trending variances of the terms of the form $a_{ts}u_s$ in (6.5) can be handled using variants of the techniques cited in Section 6.1. The stochastic equicontinuity of the increments is established similarly. Essentially, a result is available for the case where $\{u_t\}$ satisfies the assumptions of Theorem 5.1; see Davidson and de Jong (2000) for the details. However, it may be noted that the conditions on $\{\theta_j\}$ specified in Theorem 5.2 are violated by the coefficients in (6.4), and the same argument can be used to show that the limit is an fBM in that case. Theorem 5.2 can be viewed as a giving a necessary memory condition for the linear case.

Sometimes, (6.6) is referred to as Type 1 fBM, terminology coined by Marinucci and Robinson (1999) . Type 2 fBM is the case

$$X(r) = \frac{1}{\Gamma(d+1)}\int_0^t (r-s)^d dB \qquad (6.7)$$

with the integral over the range $(-\infty, 0]$ omitted. This formula is obtained as the limit of the process derived from

$$x_t = \sum_{k=1}^{t}\sum_{j=0}^{k-1} b_j u_{k-j} = \sum_{s=1}^{t} a_{ts}u_s$$

(compare (6.5)) or, in other words, the process generated from the sequence $\{I(t \geq 1)u_t\}$, where $I(.)$ is the indicator function of its argument). A number of studies have used the Type 2 representation of fBM, although arguably the setup is somewhat artificial, and it is perhaps worrying that this model yields different asymptotics. However, one virtue of formula (6.7) is that it may be compared directly with the OU process (6.3), in which the hyperbolic weight function is replaced by an exponential weight function. The two models represent alternative ways of introducing autocorrelation into a stochastic process.

[13] The first formula appears in Mandelbrot and van Ness (1968) and other references. I am grateful to a referee for pointing out that the second formula is equivalent.

7 Stochastic Integrals

There is one result required for asymptotic inference for which the FCLT and CMT do not provide a general solution. This is the limiting distribution of normalized sums of the form

$$G_T = \frac{1}{T\sqrt{\omega_{uu}\omega_{ww}}}\sum_{t=1}^{T-1}\sum_{s=1}^{t} u_s w_{t+1} \tag{7.1}$$

where w_t and u_t are I(0) processes. The weak limits of expressions of this type, involving both integrated processes and their increments, are known as Itô integrals.

7.1 Derivation of the Itô Integral

Let f be a random element of the space $D_{[0,1]}$, and let B be a BM. The Itô integral of f with respect to B (respectively, integrand and integrator functions) is a random process written

$$I(t) = \int_0^t f dB$$

and satisfying the property

$$E(I(t)^2) = E\left(\int_0^t f^2 ds\right). \tag{7.2}$$

Itô's rule is the generalization of the integration-by-parts formula to objects of this kind. It states that, if g is a twice-differentiable function of a real variable and $g(0) = 0$, then

$$g(B(t)) = \int_0^t g'(B)dB + \tfrac{1}{2}\int_0^t g''(B)ds, \text{ a.s.}$$

For example, the case $g(B) = B^2$ yields

$$\int_0^t BdB = \tfrac{1}{2}\left(B(t)^2 - t\right) \tag{7.3}$$

which may be noted to correspond to (4.13) with $t = 1$.

We sketch here the derivation of these processes by a limit argument, omitting many details and technical excursions. For a fuller account see, for example, Karatzas and Shreve (1988) among numerous good texts on stochastic calculus, many with applications to financial modelling. A *filtration* $\{\mathcal{F}(r), r \in [0,1]\}$ is a collection of nested σ-fields indexed by points of the line, and we assume that f and B are adapted processes, such that $f(r)$ and $B(r)$ are $\mathcal{F}(r)$-measurable. Choose a partition of $[0,1]$, $\Pi_k = \{r_0, r_1, \ldots r_k\}$, where $r_0 = 0$, $r_k = 1$ and $r_j < r_{j+1}$, and consider

$$I_k = \sum_{j=0}^{k-1} f(r_j)(B(r_{j+1}) - B(r_j)). \tag{7.4}$$

Notwithstanding that B is of unbounded variation, we may show that I_k converges to the limit $\int_0^1 fdB$ in mean square, as $k \to \infty$, such that $\max_j |r_{j+1} - r_j| \to 0$ (for example, let $r_j = j/k$).

It is easiest to show this result initially for the case of *simple* functions, having the form

$$f_k(r) = f(r_j), \quad r_j \le r < r_{j+1},\ j = 0, \ldots, k-1. \tag{7.5}$$

In these cases, the Itô integral with respect to BM is defined by (7.4), since

$$I_k = \int_0^1 f_k dB = \sum_{j=0}^{k-1} f(r_j)\int_{r_j}^{r_{j+1}} dB(r).$$

Note the special property that the increments of the integrator function are dated to lead the integrand, and hence are unpredictable with respect to $\mathcal{F}(r_j)$, by definition of BM.

Now define an increasing sequence $\{k(n), n = 1, 2, 3, \ldots\}$ with a corresponding sequence of partitions satisfying $\Pi_{k(1)} \subset \Pi_{k(2)} \subset \Pi_{k(3)} \subset \cdots$. For example, take $k(n) = 2^n$. Note that for $m > 0$,

$$I_{k(n+m)} - I_{k(n)} = \sum_{j=0}^{k(n+m)-1} [f_{k(n+m)}(r_j) - f_{k(n)}(r_j)](B(r_{j+1}) - B(r_j))$$

Also, applying the law of iterated expectations and the fact that $f(r_j)$ is measurable with respect to $\mathcal{F}(r_j)$, we have the properties

$$\begin{aligned} E[f(r_j)(B(r_{j+1}) - B(r_j))]^2 &= E[f(r_j)^2 E[(B(r_{j+1}) - B(r_j))^2 | \mathcal{F}(r_j)] \\ &= E[f(r_j)^2](r_{j+1} - r_j) \end{aligned}$$

and

$$E[f(r_j)(B(r_{j+1}) - B(r_j))f(r_{j'})(B(r_{j'+1}) - B(r_{j'}))] = 0$$

whenever $r_j \neq r_{j'}$. Hence,

$$E(I_k^2) = \sum_{j=0}^{k-1} Ef(r_j)^2(r_{j+1} - r_j) = E\left(\int_0^1 f(r)^2 dr\right).$$

Now, every function f on $[0,1]$ can be associated with a sequence $\{f_{k(n)}\}$ of simple functions, by applying definition (7.5) for a suitable class of partitions. For the class of square-integrable, 'progressively measurable' functions f on $[0,1]$,[14] it can be shown that there exists a sequence of simple functions such that, for each $m > 0$,

$$E(I_{k(m+n)} - I_{k(n)})^2 = E\left(\int_0^1 (f_{k(n+m)}(r) - f_{k(n)}(r))^2 dr\right) \to 0$$

as $n \to \infty$. It follows that

$$E(I_{k(n)} - I)^2 \to 0$$

where $I = \int_0^1 f dB$ is defined in this mean-square limiting sense. BM is, of course, a candidate case of f under this definition, leading to the class of random variables with the property in (7.3).

7.2 Weak Convergence of Covariances

In functional limit theory, the role of the Itô integral is to characterize the weak limits of random sequences of type (7.1). The arguments needing to be deployed to show convergence are again quite involved and technical,[15] and it is not possible to do more than just sketch an outline of the main steps. Let $X_T(r)$ be defined as in (2.14) and $Y_T(r)$ defined similarly with respect to the partial sum process $y_t = \sum_{s=1}^t w_s$, where $w_t \sim iid(0, \sigma_w^2)$. We assume that $(X_T, Y_T) \xrightarrow{\mathrm{d}} (B_X, B_Y)$, where the notation is intended to indicate *joint* convergence of the sequence of pairs, and we don't rule out the possibility that x_t and y_t are contemporaneously correlated or, indeed, identical. Let

$$G_T^* = \sum_{j=0}^{k(T)-1} X_T(r_j)(Y_T(r_{j+1}) - Y_T(r_j))$$

[14] See e.g. Davidson (1994) Chapter 30.2 for the definition of progressive measurability.

[15] Chan and Wei (1988) and Kurtz and Protter (1991) are important references here, and see also Hansen (1992), Davidson (1994) Section 30.4 and de Jong and Davidson (2000).

where $\{r_0, \dots, r_{k(T)}\} = \Pi_T$ is a nested sequence of partitions as defined previously, and $k(T) \to \infty$ as $T \to \infty$, although more slowly, such that $k(T)/T \to 0$, and

$$\min_{0 \leq j < k(T)} |Tr_{j+1} - Tr_j| \to \infty$$

whereas the condition

$$\max_{0 \leq j < k(T)} |r_{j+1} - r_j| \to 0 \tag{7.6}$$

still holds. G_T^* is an approximation to G_T, constructed from time-aggregated components that approach independent segments of Brownian motions as T increases, by the FCLT. We accordingly attempt to show that

$$G_T^* \xrightarrow{\mathrm{d}} \int_0^1 B_X dB_Y. \tag{7.7}$$

The FCLT is not sufficient alone to let us deduce (7.7). However, a clever result known as the *Skorokhod representation theorem* supplies the required step. Roughly speaking, this asserts that (in the present case), whenever the joint distribution of a sequence $(X_T,\, Y_T)$ converges weakly to (B_X, B_Y), there exists a sequence of random processes (X^T, Y^T) converging almost surely to limit processes that are jointly distributed as (B_X, B_Y). (Since these processes are elements of $D_{[0,1]}$, the relevant concept of convergence has, in practice, to be carefully specified in terms of the Skorokhod topology.) It therefore suffices to show that, if a random variable G^{*T} is constructed like G_T^* except in terms of the Skorokhod processes (X^T, Y^T), then

$$\left| G^{*T} - \int_0^1 B_X dB_Y \right| \xrightarrow{\mathrm{pr}} 0.$$

Since convergence in probability implies convergence in distribution, and G^{*T} and G_T^* have the same distribution by construction, this line of argument suffices to establish (7.7).

It then remains to consider $G_T - G_T^*$. Defining $T_j = [Tr_j]$, we have

$$Y_T(r_{j+1}) - Y_T(r_j) = \frac{1}{\sigma_w T^{1/2}} \sum_{t=T_j}^{T_{j+1}-1} w_{t+1}$$

and

$$X_T(r_j) = \frac{1}{\sigma_u T^{1/2}} \sum_{m=0}^{T_j - 1} u_{T_j - m}.$$

Therefore,

$$\begin{aligned} G_T - G_T^* &= \frac{1}{\sigma_u \sigma_w T} \sum_{j=1}^{k(T)} \left(\sum_{t=T_{j-1}}^{T_j-1} \sum_{m=0}^{T_j-1} u_{t-m} w_{t+1} - \sum_{m=0}^{T_j-1} u_{T_j-m} \sum_{t=T_{j-1}}^{T_j-1} w_{t+1} \right) \\ &= \frac{1}{\sigma_u \sigma_w T} \sum_{j=1}^{k(T)} \sum_{t=T_{j-1}+1}^{T_j-1} \left(\sum_{m=0}^{t-T_{j-1}-1} u_{t-m} \right) w_{t+1} \end{aligned}$$

Recall our assumption that the processes $\{u_t, w_t\}$ are serially independent. In this case we may show convergence in mean square, as follows:

$$E(G_T - G_T^*)^2 = \frac{1}{\sigma_u^2 \sigma_w^2 T^2} \sum_{j=1}^{k(T)} \sum_{t=T_{j-1}+1}^{T_j-1} E \left(\sum_{m=0}^{t-T_{j-1}-1} u_{t-m} \right)^2 E(w_{t+1})^2$$

$$
\begin{aligned}
&= \frac{1}{T^2}\sum_{j=1}^{k(T)}\sum_{t=T_{j-1}+1}^{T_j-1}(t-T_{j-1}) \\
&\leq \frac{1}{T^2}\sum_{j=1}^{k(T)}(T_j-T_{j-1})^2 \\
&= O\left(\max_{1\leq j\leq k(T)}|r_j-r_{j-1}|\right) = o(1)
\end{aligned}
$$

where the final order of magnitude holds in view of (7.6). This implies convergence in probability, as noted.

Showing this convergence in probability for the general case where $\{u_t, w_t\}$ are serially correlated is generally more difficult, and will not be attempted here, although the assumptions required are comparable with those needed for the FCLT to hold. The main thing to note is that in these cases the probability limit is not zero, and

$$\text{plim}(G_T - G_T^*) = \lambda_{uw} = \frac{1}{\sqrt{\omega_{uu}\omega_{ww}}}\sum_{m=1}^{\infty}E(u_{1-m}w_1).$$

It goes without saying that the cross-autocorrelations need to form a summable series, under the assumptions validating these results. Hence, the full convergence result that is commonly cited in arguments relating to convergence of regression statistics, for example, is

$$G_T \xrightarrow{\text{d}} \int_0^1 B_X dB_Y + \lambda_{uw}.$$

For reference, let us state here a fairly general stochastic integral convergence theorem (SICT), adapted from De Jong and Davidson (2000). This is a companion result to Theorem 5.1 with essentially the same set of assumptions.

Theorem 7.1 *Let (X_T, Y_T) be defined as in (5.4) with respect to increment processes $\{u_t, w_t\}$. Let these processes satisfy conditions (a)-(c) of Theorem 5.1, with long-run variances ω_{uu} and ω_{ww} and long-run covariance*

$$\omega_{uw} = \sigma_{vw}^2 + \lambda_{uw} + \lambda_{wu}'$$

where $\lambda_{uw} = \sum_{m=1}^{\infty}E(u_{1-m}w_1)$ and $\lambda_{wu}' = \sum_{m=1}^{\infty}E(w_{1-m}u_1)$. If

$$G_T = \sum_{t=1}^{T}X_T(t/T)(Y_T((t+1)/T) - Y_T(t/T))$$

then

$$(X_T, Y_T, G_T) \xrightarrow{\text{d}} (B_X, B_Y, \int_0^1 B_X dB_Y + \lambda_{uw})$$

Note the importance of specifying the *joint* convergence of the three components, in this result. This is mainly a formality as far as the proof is concerned, but it is required to be able to apply the CMT to functionals of the components.

7.3 Application to Regression

A standard case is a cointegrating regression. In a simplified setup in which there are no deterministic components such as intercepts, consider the model

$$y_t = \beta x_t + \varepsilon_t$$

$$\Delta x_t = u_t$$

where $\{u_t, \varepsilon_t\}$ are (at worst) stationary short memory processes, such that their normalised partial sums converge to BM. Let $\hat{\beta}$ represent the OLS estimator. Combining the FCLT and SICT with the CMT, it can be shown that

$$T(\hat{\beta} - \beta) = \frac{T^{-1}\sum_{t=1}^{T} x_t\varepsilon_t}{T^{-2}\sum_{t=1}^{T} x_t^2} \xrightarrow{\text{d}} \frac{\sqrt{\omega_{uu}\omega_{\varepsilon\varepsilon}}\int_0^1 B_x dB_\varepsilon + \sigma_{u\varepsilon} + \lambda_{u\varepsilon}}{\omega_{uu}\int_0^1 B_x^2 dr} \tag{7.8}$$

where the quantities $\omega_{\varepsilon\varepsilon}$, $\sigma_{u\varepsilon}$ and $\lambda_{u\varepsilon}$ are defined by analogy with Theorem 7.1, although be careful to note that the increments ε_t are not those of the y_t process.

This result embodies all the known asymptotic properties of the cointegrating regression. First, 'superconsistency', or convergence at the rate T^{-1}. Second, median bias of order $O(T^{-1})$.[16] Third, the limiting distribution is certainly not Gaussian, noting that the numerator and denominator in (7.8) are dependent in general.

However, it is also possible to appreciate the special properties of OLS in the case where the regressor is *strongly exogenous*. This is the case when the pairs (u_{t-j}, ε_t) are independent, for all j and all t. This is not an easy assumption to validate in most economic contexts, but if it holds, much more useful results emerge. Firstly, we can put $\sigma_{u\varepsilon} = \lambda_{u\varepsilon} = 0$, removing this source of bias. More importantly, however, it is now valid to consider the limiting distribution of $T(\hat{\beta} - \beta)$ *conditional* on B_X, which in the context of the conditional distribution can be treated as a deterministic process - effectively, as though it were a type of trend function. It is straightforward to show that

$$\sqrt{\omega_{uu}\omega_{\varepsilon\varepsilon}}\int_0^1 B_x dB_\varepsilon | B_x \sim N\left(0, \omega_{uu}\omega_{\varepsilon\varepsilon}\int_0^1 B_x^2 dr\right)$$

and hence that

$$T(\hat{\beta} - \beta)|B_x \sim N\left(0, \omega_{\varepsilon\varepsilon}\left(\omega_{uu}\int_0^1 B_x^2 dr\right)^{-1}\right)$$

Here the unconditional distribution is *mixed Gaussian*. This means that it is distributed like a drawing from a two-stage sampling procedure in which first a positive random variable V is drawn from the distribution of $\omega_{\varepsilon\varepsilon}\left(\omega_{uu}\int_0^1 B_x^2 dr\right)^{-1}$, and then a drawing from the $N(0, V)$ distribution yields the observed variable. This distribution is distinguished by excess kurtosis, but this is less important than the fact that the t ratio for the regression, being normalized by an "estimate" of the random variance, is asymptotically standard normal. Hence, standard asymptotic inference is available in this model, and the same arguments generalize in a natural way to the multiple regression case. However, in the case of non-strongly endogenous regressors, different procedures are necessary to achieve standard inference. See, for example, Phillips and Hansen (1990) and Phillips and Loretan (1988) for further discussion of these methods.

8 The Infinite Variance Case

8.1 $\alpha-$Stable Distributions

An essential property of the driving process in all the foregoing results has been a finite variance. Without this, there is no CLT in the usual sense. To understand what might happen instead,

[16] It is evident that the numerator does not have a mean of zero, but since it is not straightforward to compute the moments of the limiting ratio, it is better to comment on its distribution directly without invoking integrability.

we must define the class of *stable* distributions. If X has a stable distribution, then for any T, there exist identically distributed r.v.s $U_1, \dots, U_T$ and U, and real sequences $\{a_T, b_T\}$, where $a_T > 0$, such that $S_T = U_1 + \dots + U_T$ has the same distribution as $a_T U + b_T$. The distribution is sometimes called *strictly* stable when this is true for the case $b_T = 0$. A stable distribution is, by construction, *infinitely divisible*, meaning that it can be represented as the distribution of a T-fold sum of i.i.d. r.v.s, for any T. The leading example is the $N(\mu, \sigma^2)$ distribution, being the stable case in which S_T has the same distribution as $\sqrt{T}U + (T - \sqrt{T})\mu$. It is clear that, if it is to act as an attractor for the distribution of a normalized partial sum, a distribution must be stable.

The normal is the unique stable distribution possessing a variance. However, there exists an extensive class of stable distributions with infinite variance, identified by the form of their characteristic functions. (Except for the normal, the density functions are not generally available in closed form.) In general, members of the stable class can be positively or negatively skewed around a point of central tendency. Confining attention for simplicity to symmetric distributions centered on zero, the ch.f.s have the form.

$$\phi_U(\lambda) = e^{-\delta|\lambda|^\alpha} \tag{8.1}$$

for $0 < \alpha \leq 2$ (stability parameter) and $\delta > 0$ (scale parameter). The $N(0, \sigma^2)$ distribution is the case where $\alpha = 2$ and $\delta = \sigma^2/2$. When $\alpha < 2$, distributions with ch.f. shown in (8.1), having no variance, are characterized by 'fat tails' with a relatively high probability of outlying values.

Considering again identity (2.4), we find that, in this case,

$$[\phi_U(\lambda)]^T = \phi_U(\lambda T^{1/\alpha})$$

and the stability property is obtained with $b_T = 0$ and $a_T = T^{1/\alpha}$. Distributions having this property for a_T (not necessarily with $b_T = 0$) are called stable with exponent α, or α-stable. If the ch.f. takes the special form of (8.1) they are called *symmetric α-stable* or $S\alpha S$. Another leading $S\alpha S$ case is the Cauchy distribution, corresponding to $\alpha = 1$. Since $e^{-|\lambda|^\alpha} = E(e^{i\lambda\delta^{-1/\alpha}U})$, one might think by analogy with the standard normal of the 'standard' $S\alpha S$, having $\delta = 1$, obtained by dividing the r.v. by $\delta^{-1/\alpha}$. Useful references on the properties of stable distributions include Breiman (1968), Feller (1971), Ibragimov and Linnik (1971) and Samorodnitsky and Taqqu (1994).

Every distribution with finite variance lies in the 'domain of attraction of the normal law'. However, there evidently exists a whole family of stable convergence laws of which the CLT is only the leading case. The most important question, of course, is what distributions may fall within the domain of attraction of a stable law associated with given α? The answer to this more general question relates to the tail behaviour of the distributions in question, and can be stated in terms of the distribution function F. Breiman (1968, Th. 9.34) gives the necessary and sufficient conditions as follows: that there exist nonnegative constants M^+ and M^-, of which at least one is positive, such that

$$\frac{F(-x)}{1 - F(x)} \to \frac{M^-}{M^+} \text{ as } x \to \infty$$

and for every $\xi > 0$, as $x \to \infty$,

$$\text{if } M^+ > 0 \text{ then } \frac{1 - F(\xi x)}{1 - F(x)} \to \xi^{-\alpha} \tag{8.2}$$

$$\text{if } M^- > 0 \text{ then } \frac{F(-\xi x)}{F(-x)} \to \xi^{-\alpha}. \tag{8.3}$$

Note that stable laws can be skewed although, of course, $M^+ = M^-$ in symmetric cases. This is one point of difference from the normal CLT, in which normalized sums of skewed r.v.s are nonetheless symmetric in the limit.

In the symmetric case, properties (8.2) and (8.3) are satisfied if

$$P(|U| > x) = x^{-\alpha} L(x)$$

where L is a *slowly varying* function, having the property that $L(ax)/L(x) \to 1$ as $x \to \infty$ for any $a > 0$. For example, $\log x$ is slowly varying. The key property is that there exists $B > 0$ such that $x^{-\alpha} < L(x) < x^{\alpha}$ for all $x > B$. We say in this case that the tails of the distribution obey a power law with exponent α. Note the implication, if the distribution is continuous and f_U denotes the p.d.f., that

$$f_U(x) = O(|x|^{-\alpha-1} L(x))$$

as $x \to \pm\infty$, such that the tail integrals are of the specified order of magnitude. By contrast, to ensure that

$$E(U^2) = \int_{-\infty}^{\infty} x^2 f_U(x) dx < \infty$$

note that the tails of the p.d.f. must be of $O(|x|^{-3-\varepsilon})$ for $\varepsilon > 0$, and hence need $\alpha > 2$. In this case, the attractor p.d.f. has exponentially vanishing tails according to the CLT, regardless of α, whereas for $\alpha < 2$ the tail behaviour is shared by the limit distribution, under the stable law.

We summarise these points by noting that there exists a generalization of the CLT of the following sort.

Theorem 8.1 *If $S_T = \sum_{t=1}^{T} U_t$, and $U_1, \ldots, U_T$ are i.i.d. random variables in the domain of attraction of a symmetric stable law with parameter α, there exists a slowly varying function L such that $S_T/(T^{1/\alpha} L(T))$ converges weakly to $S\alpha S$.*

Little research appears to have been done to date on the case of dependent increments, one obvious difficulty being that the autocorrelation function is undefined. However, it appears a reasonable conjecture that suitable variants of the arguments deployed for the dependent CLT, such as Theorem 5.1, might yield generalizations for L_p-NED functions of mixing processes with $p < \alpha$.

8.2 Lévy Motion

If convergence to an α-stable law takes the place of the usual CLT, the existence of a corresponding FCLT must be our next concern. Consider, as before, the normalized partial sum process $X_T \in D_{[0,1]}$ such that

$$X_T(r) = \frac{1}{T^{1/\alpha} L(T)} \sum_{t=1}^{[Tr]} U_t. \tag{8.4}$$

Consider the behaviour of this process as T increases. Note first of all that

$$P(T^{-1/\alpha} L(T)^{-1} |U_t| > x) = O(T^{-1} L(T))$$

and, therefore, the normalization ensures that the probability that the process jumps by a positive amount is converging to zero. However, note too how the proof of tightness in $C_{[0,1]}$, given in Section 4.5, now fails. With $p < 2$, the inequality in (4.7) cannot be satisfied for arbitrary $\varepsilon, \eta > 0$ under the stated conditions. In fact, the limit process does *not* lie in $C_{[0,1]}$ almost surely.

A more general class of processes in $D_{[0,1]}$ containing these limits is defined as follows. A random function $X : [0,1] \mapsto \mathbb{R}$ is called a *Lévy process* if

(a) it is cadlag, a.s.

(b) it has independent increments,

(c) it is continuous in probability: $P(X(r+s) - X(r)) \to 0$ as $s \to 0$,

(d) it is time-homogeneous: the distribution of $X(s+r) - X(r)$ does not depend on r,

(e) $X(0) = 0$ a.s.

BM is a Lévy process on this definition, although one enjoying additional continuity properties. On the other hand, if the fidis of a Lévy process are strictly stable distributions with parameter α, such that $X(s+r) - X(r)$ is distributed as $s^{1/\alpha}X(1)$ for all $0 < s < 1$ and $0 \leq r < 1 - s$, it is called an α-stable motion, or a *Lévy motion*, denoted Λ_a.

In addition to the fidis converging to those of Lévy motion, the counterpart of the FCLT for these processes must establish that the sequence of p.m.s associated with X_T is uniformly tight, and the limit process lies in $D_{[0,1]}$ almost surely. The tightness conditions are less stringent than for the Gaussian case, since it is only necessary to rule out realizations with isolated discontinuities arising with positive probability. Billingsley (1968, Theorem 15.6) gives a sufficient condition for tightness for processes in $D_{[0,1]}$. The weak convergence specified in the following result is defined as before with respect to the Skorokhod topology on $D_{[0,1]}$.

Theorem 8.2 *If X and $\{X_T, T = 1, 2, 3, \ldots\}$ are random elements of $D_{[0,1]}$ and*

(a) *the fidis of X_T converge to those of X;*

(b) *there exist $\gamma \geq 0$, $\mu > 1$ and a continuous nondecreasing function F on $[0,1]$ such that*

$$P(\min\{|X_T(r) - X_T(r_1)|, |X_T(r_2) - X_T(r)|\} \geq \lambda) \leq \frac{[F(r_2) - F(r_1)]^\mu}{\lambda^{2\gamma}}$$

for all $T \geq 1$, and $r_1 \leq r \leq r_2$ whenever $r_1, r_2 \in K_X$, where $K_X \subseteq [0,1]$ is the set of points r at which $P(X(r) \neq X(r-)) = 0$, including 0 and 1;

then $X_T \xrightarrow{\text{d}} X$.

For the case of a Lévy motion we can take $K_X = [0,1]$. Condition (b) sets the probability of isolated discontinuities sufficiently low to ensure uniform tightness. A sufficient condition for (b) is

$$E(|X_T(r) - X_T(r_1)|^\gamma |X_T(r_2) - X_T(r)|^\gamma) \leq [F(r_2) - F(r_1)]^\mu. \tag{8.5}$$

Suppose the fidis are tending to $S\alpha S$ limits, and that the increments are independent. Then, choosing $\gamma < \alpha$ such that the moments exist, note that the left-hand side of (8.5) equals

$$E|X_T(r) - X_T(r_1)|^\gamma E|X_T(r_2) - X_T(r)|^\gamma \leq M\frac{([Tr_2] - [Tr])^{\gamma/\alpha}([Tr] - [Tr_1])^{\gamma/\alpha}}{T^{2\gamma/\alpha}} \tag{8.6}$$

for some $M < \infty$. If $\min(r_2 - r, r - r_1) < T^{-1}$ then note that the left-hand side of (8.6) vanishes. Otherwise, the right-hand side is bounded by

$$M\frac{([Tr_2] - [Tr_1])^{2\gamma/\alpha}}{T^{2\gamma/\alpha}} \leq 4M(r_2 - r_1)^{2\gamma/\alpha}.$$

Choosing $\gamma > 0$ to satisfy $\gamma < \alpha < 2\gamma$ shows that condition (b) of Theorem 8.2 is satisfied, and the weak limit of such a process is a Lévy motion.

More research remains to be done before the infinite variance case is understood as thoroughly as convergence to Brownian motion, especially under dependence. There are various new issues to be treated, such as the restrictions on joint convergence to limits in D^m, as described in Section 4.6. However, the prevalence of fat tailed distributions in economic data, especially in finance, is not in doubt. See Barndorff-Nielsen and Shephard (2001), among many recent references, on the importance of these processes in econometric inference.

References

Andrews, D. W.K. 1984. Non-strong mixing autoregressive processes, Journal of Applied Probability 21, 930-4.

Barndorff-Nielsen, O. E., and N. Shephard 2001. “Modelling by Lévy Processess for Financial Econometrics,” in Ole E. Barndorff-Nielsen, T. Mikosch and S. Resnick (eds.), Lévy Processes – Theory and Application New York: Birkhauser, 283–318.

Beveridge, S., and Nelson, C.R. 1981. A new approach to the decomposition of economic time series into permanent and transitory components with particular attention to measurement of the ‘business cycle’. Journal of Monetary Economics 7, 151-174.

Billingsley, P., 1968. Convergence of Probability Measures, John Wiley, New York.

Billingsley, P., 1986. Probability and Measure 2nd Edn., John Wiley, New York.

Breiman, L. 1992. Probability. Philadelphia: Society for Industrial and Applied Mathematics.

Chan, N.H. and Wei, C.Z. 1988. Limiting distributions of least squares estimates of unstable autoregressive processes, Annals of Statistics, 16, 367-401.

Davidson, J. 1994. Stochastic Limit Theory, Oxford: Oxford University Press.

Davidson, J. 2000. Establishing conditions for the functional central limit theorem in nonlinear and semiparametric time series processes. Journal of Econometrics 106 243-269

Davidson, J. 2004. Convergence to stochastic integrals with fractionally integrated integrator processes: theory, and applications to fractional cointegration analysis". Working paper at http://www.ex.ac.uk/~jehd201

Davidson, J. and R. M. de Jong 2000. The functional central limit theorem and weak convergence to stochastic integrals II: fractionally integrated processes Econometric Theory 16, 5, 643-66.

De Jong, R. M. and J. Davidson 2000. The functional central limit theorem and weak convergence to stochastic integrals I: weakly dependent processes. Econometric Theory 16, 5, 621-42.

Donsker, M. D. 1951. An invariance principle for certain probability limit theorems, Memoirs of the American Mathematical Society, 6, 1-12.

Dudley, R. M. 1989. Real Analysis and Probability, Wadsworth and Brooks/Cole, Pacific Grove, Calif.

Embrechts, P., C. Kluppelberg and T. Mikosch 1997. Modelling Extremal Events, Springer, Berlin.

Feller, W. 1971. An Introduction to Probability Theory and its Applications, ii, John Wiley, New York.

Gallant, A. R. 1997. An Introduction to Econometric Theory. Princeton: Princeton University Press.

Gallant, A. R. and White H. 1988. A Unified Theory of Estimation and Inference for Nonlinerar Dynamic Models, Basil Blackwell, Oxford.

Hansen, B. E. 1992. Converence to Stochastic integrals for dependent heterogeneous processes, Econometric Theory 8, 489-500.

Ibragimov, I.A., and Linnik, Yu. V. 1971. Independent and Stationary Sequences of Random Variables, Wolters-Noordhoff, Groningen.

Jacod, J and A. N. Shiryaev 1987. Limit Theorems for Stochastic Processes, Springer, Berlin

Karatzas, I. and Shreve, S. E. 1988. Brownian Motion and Stochastic Calculus, Springer-Verlag, New York.

Kurtz, T. G. and Protter, P. 1991. Weak limit theorems for stochastic integrals and stochastic differential equations, Annals of Probability 19, 1035-70.

Mandelbrot, B. B. 1983. The Fractal Geometry of Nature, W.H. Freeman, New York.

Mandelbrot, B. B. and J. W. van Ness 1968. Fractional Brownian motions, fractional noises and applications. SIAM Review 10, 4, 422–437.

Marinucci, D. and P. M. Robinson, 1999. Alternative forms of fractional Brownian motion. Journal of Statistical Inference and Planning 80, 111–122.

McCabe, B. and A. Tremayne 1993. Elements of modern asymptotic theory with statistical applications. Manchester: Manchester University Press.

Phillips, P.C.B. 1988. Weak convergence of sample covariance matrices to stochastic integrals via martingale approximations, Econometric Theory 4, 528-33

Phillips, P. C. B. and B. E. Hansen 1990. Statistical inference in instrumental variables regression with I(1) processes, Review of Economic Studies 57, 99–125.

Phillips, P. C. B. and M. Loretan 1991. Estimating long-run economic equilibria. Review of Economic Studies 58, 407–37.P.

Phillips, P. C. B. and V. Solo 1992. Asymptotics for linear processes, Annals of Statistics 20, pp. 971–1001.

Pollard, D. 1984. Convergence of Stochastic Processes,Springer-Verlag, New York.

Pollard, D. 2002. A User's Guide to Measure Theoretic Probability, Cambridge: Cambridge University Press.

Samorodnitsky, G. and M. Taqqu 1994. Stable Non-Gaussian Random Processes. Boca Raton: Chapman and Hall.

Skorokhod, A.V. 1956. Limit theorems for stochastic processes, Theory of Probability and its Applications 1, 261-90.

White, H. 1984: Asymptotic Theory for Econometricians. Orlando: Academic Press.

Wiener, Norbert 1923. Differential space, Journal of Mathematical Physics 2, 131-74.

Wold, H. 1938. A Study in the Analysis of Stationary Time Series. Uppsala: Almqvist and Wiksell.

Woolridge, J. M. and White, H. 1988. Some invariance principles and central limit theorems for dependent heterogeneous processes, Econometric Theory 4, 210-30.