

Time Series Econometrics: Notes

Rob Wlodarski

March 6, 2024

Contents

1	Univariate Models	4
1.1	Preliminaries	4
1.1.1	Probability & Sample Spaces	4
1.1.2	σ -Algebra	4
1.1.3	Probability Measure	5
1.1.4	Random Variables, Stochastic Processes, & Their Classification	6
1.1.5	Characterisation of Univariate Time Series	7
1.2	Stationarity	7
1.2.1	Autocorrelation Function	8
1.2.2	Strict Stationarity & Ergodicity	8
1.3	Prediction	9
1.3.1	Loss Function(s)	10
1.3.2	BLP & LS Prediction Rule(s)	11
1.3.3	Properties of Projection	12
1.3.4	Time Series Context	12
1.4	The Wold Representation Theorem	14
1.4.1	White Noise (WN)	14
1.4.2	Uses of the Wold Representation Theorem	15
1.5	Frequency Domain	16
1.5.1	A Digression: Cyclic Functions of Constant Period	16
1.5.2	Spectral Analysis	17
1.6	ARMA Models	21
1.6.1	MA(1) Process	22
1.6.2	MA(2) Process	23
1.6.3	MA(k) Process	24
1.6.4	AR(1) Process	25
1.6.5	AR(2) Process	26
1.6.6	AR(k) Process	27
1.6.7	ARMA Process	27
1.7	Partial Autocorrelation Function	28
1.8	Long Memory Process	29
1.9	Temporal Aggregation	29
1.9.1	Relationship between Discrete and Continuous Time Processes	31

2	Multivariate Time Series	33
2.1	Introduction	33
2.1.1	Vector Stochastic Processes	33
2.1.2	Stationarity	35
2.2	Prediction	35
2.3	Multivariate Spectral Analysis	37
2.3.1	Multivariate Harmonic Process	37
2.3.2	Spectral Density Matrix	38
2.4	VARMA Models	39
2.4.1	VMA(1) Process	39
2.4.2	VMA(k) Process	40
2.4.3	VAR(1) Process	41
2.4.4	VAR(h) Process	43
2.4.5	VARMA(h,k) Process	43
2.5	Granger Causality	43
2.6	Linear Transformations, Marginalisations, and Contemporaneous Aggregation	45
2.6.1	Linear transformations	45
2.6.2	Marginalisation	45
2.6.3	Contemporaneous Aggregation	46
2.7	Time-Invariant Linear Filters	47
2.8	Structural VARs and Impulse Responses	48
2.9	Dynamic Models with Latent Variables	49
2.9.1	Linear, Time-Invariant, State Space Models	49
2.9.2	Structural and Reduced Forms: UCARIMA Model	50
2.9.3	Structural and Reduced Forms: Dynamic Factor Model	51
2.9.4	Identification of UCARIMA Models	51
2.9.5	Identification of dynamic single factor models	51
2.9.6	The Wiener-Kolmogorov Filter	52
2.9.7	Dynamic Models with Latent Variables	54
2.9.8	Kalman Filter	54
2.9.9	Kalman Smoother	55
3	Integration and Non-Linear Models	55
3.1	Models for Non-Stationary Series	55
3.2	Cointegration	57
3.3	Non-Linear Models	59
3.3.1	Readons for Using Non-Linear Models	59
3.3.2	TAR & STAR Models	59
3.3.3	ARCH & SV Models	60
3.3.4	Compound Autoregressive Processes (CAR)	62
3.3.5	Binary Markov Chains	63
3.3.6	k-ary Markov chains	64
3.3.7	Absorbing Markov Chains	64
3.3.8	Regime switching (RS) Models	66
4	Inference with Dependent Observations	66
4.1	DGPs, Models and Parameters	66
4.1.1	Introduction	66
4.2	Statistical Properties	72
4.2.1	Static Gaussian PMLE of a Homogenous Binary Markov Chain	74
4.2.2	MA(1) estimated as AR(1)	75
4.2.3	Identification	78

4.2.4	Consistency	79
4.2.5	Limiting distribution & Sandwich Formula	80
4.2.6	Gaussian Pseudo Maximum Likelihood Estimation	82
4.2.7	Correctly Specified 1st and 2nd Conditional Moments	83
4.3	Spectral Maximum Likelihood	85
4.3.1	Univariate Models	85
4.3.2	Multivariate Case	87
4.3.3	Spectral Maximum Likelihood: Pros and Cons	87
4.4	Sequential Estimation	88
4.5	Hypothesis Testing	89
4.5.1	Classical Hypothesis Testing	89
4.6	Specification Tests	91
4.6.1	Serial Correlation Tests in Frequency Domain	91
4.6.2	(Durbin-Wu-)Hausman Tests	92
4.6.3	The Matryoshka Dolls' Relative Efficiency of the Estimators	94
4.6.4	Information Matrix Tests	95
5	Estimating Time Series Regression Models	97
5.1	Autoregressive Models	97
5.1.1	Case I	97
5.1.2	Case II	98
5.2	LLNs and CLTs with Dependent Observations	99
5.3	Autoregressive Models Cont'd	100
5.3.1	Case III	100
5.3.2	Case IV	100
5.4	Unit Roots	102
5.4.1	Driftless Random Walk	102
5.4.2	Unit Root Tests	103
5.5	Cointegration Tests	104
5.6	Dynamic Regression Models in Practice: Exchange Rate Example	105

1 Univariate Models

1.1 Preliminaries

1.1.1 Probability & Sample Spaces

A space is a set with some added structure. Here, we focus on **probability spaces, a special case of measurable spaces**. A probability space is constructed with a specific random experiment in mind. By random experiment, we mean that the outcome is not certain, but each time the experiment is repeated, the set of possible outcomes is the same. The probability of all its elements is the same.

Definition 1.1 (Probability Space). A **probability space** consists of three parts:

1. the **sample space**, which is the set of all possible outcomes;
2. the **set of events**, where each event is a set containing zero or more outcomes; and
3. the **assignment of probabilities** to the events

Definition 1.2. A **sample space**, Ω , (or *state space* or *set of states of nature*) is the set of all possible outcomes.

An **outcome is the result of a single execution of the experiment**. Since individual outcomes might be of little practical use, more complex events are used to characterise groups of outcomes.

Example 1.1. When drawing a card from a standard deck of 52 playing cards, one possibility for the sample space could be the rank (*Ace* through *King*), while another could be the suit (e.g. diamonds, etc.). A complete description of outcomes would specify both, and a sample space describing each card can be constructed as the Cartesian product of the two sample spaces noted above.

Any **subset of the sample space is called an event**. However, this gives rise to problems when the sample space is infinite, so we need to be more precise.

1.1.2 σ -Algebra

To turn a set of outcomes into a measurable space one endows it with a σ -algebra.

Definition 1.3. Let Ω be some set, and 2^Ω its power set (i.e., the set of all subsets of Ω). Then a subset $\mathcal{G} \subset 2^\Omega$ is called **σ -algebra** if it satisfies the following three properties:

1. \mathcal{G} is **non-empty**: there is at least one $B \subset \Omega$ in \mathcal{G} ;
2. \mathcal{G} is **closed under complementation**: if $B \in \mathcal{G}$, then $\Omega \setminus B \in \mathcal{G}$; and
3. \mathcal{G} is **closed under countable unions**: if $B_1, B_2, \dots \in \mathcal{G}$, then $B_1 \cup B_2 \cup \dots \in \mathcal{G}$.

Elements of the σ -algebra are called **measurable sets**. An ordered pair (Ω, \mathcal{G}) , where Ω is a set of outcomes and \mathcal{G} is a σ -algebra over Ω , is called a **measurable space**.

Theorem 1.1. Let G be an arbitrary family of subsets of Ω . Then, **there exists a unique smallest σ -algebra which contains every set in G** (even though G may or may not itself be a σ -algebra).

This σ -algebra is denoted $\sigma(G)$ and called the σ -algebra generated by G .

Definition 1.4. A **partition** of Ω is a collection of sets $\mathcal{B} = \{B_1, \dots, B_n\}$ such that $B_i \cap B_j = \emptyset$ (**disjoint**) and $\bigcup_{i=1}^n B_i = \Omega$ (**exhaustive**).

Proposition 1.1.1. *If \mathcal{B} is a partition of Ω , then:*

$$\mathcal{G} = \left\{ G = \bigcup_{i \in I} B_i : I \subset \{1, \dots, n\} \right\} \quad (1.1)$$

is a σ -algebra.

Every σ -algebra can be generated by a partition using the previous result.

Proposition 1.1.2. $\mathcal{G} \subset \mathcal{G}^T$ if and only if the corresponding partitions \mathcal{B} and \mathcal{B}^T satisfy that for each $B \in \mathcal{B}$, there exists a $B^T \in \mathcal{B}^T$ such that $B^T \subset B$.

We denote by \mathcal{F} **the collection of all the subsets of Ω .**

Example 1.2. Generate the smallest and largest σ -algebras over $\Omega = \{\omega_1, \omega_2, \omega_3\}$, as well as $\sigma(\omega_1)$.

The smallest σ -algebra over Ω is $\mathcal{G}_0 = \{\emptyset, \Omega\}$.

The largest σ -algebra over Ω is:

$$\mathcal{F} = \{\emptyset, \omega_1, \omega_2, \omega_3, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \{\omega_2, \omega_3\}, \Omega\}. \quad (1.2)$$

Lastly, we have that:

$$\sigma(\omega_1) = \{\emptyset, \omega_1, \{\omega_2, \omega_3\}, \Omega\}. \quad (1.3)$$

We study this because, in time series, information accumulates. This means that we have nesting.

1.1.3 Probability Measure

Definition 1.5 (Measure). Let \mathcal{G} be a σ -algebra over a set Ω . A **function μ from Ω to the extended real line is called a measure** if it satisfies the following properties:

1. **non-negativity:**

$$\mu(B) \geq 0 \quad \forall B \in \mathcal{F} \quad (1.4)$$

2. **countable additivity** (or σ -additivity): for all countable collections $\{B_i\}_{i \in I}$ of pairwise disjoint sets in Ω :

$$\mu\left(\bigcup_{i \in I} B_i\right) = \sum_{i \in I} \mu(B_i); \text{ and} \quad (1.5)$$

3. **null empty set:** $\mu(\emptyset) = 0$

To summarise, the pair **(Ω, \mathcal{G}) is called a measurable space**, the members of \mathcal{G} are called measurable sets, and the triple $(\Omega, \mathcal{G}, \mu)$ is called a measure space.

Definition 1.6 (Probability Measure). A **probability measure** is a measure with total measure one (i.e., $\mu(\Omega) = 1$); a probability space is a measure with a probability measure.

Example 1.3 (French Roulette). The French (or European) roulette has 37 cells with equal odds of hitting. They are numbered $0, 1, \dots, 36$. American roulette has an additional cell labelled “00.”

📌 The sample space is $\Omega = \{0, 1, \dots, 36\}$

📌 The largest σ -algebra \mathcal{F} of events is the set of all the 2^{37} different bets that one could conceive, such as ‘Impair’, ‘Rouge’, ‘Manque’, ‘Troisième colonne’, ‘Dernière douzaine’, their combinations, and includes Ω itself (a bet on every single number), as well as the empty set ϕ , which is the no outcome event (the ball jumps out of the roulette!).

📌 The probability measure Π is such that $\mathbb{P}(\omega = j) = \frac{1}{37}$ where $j \in \{0, 1, \dots, 36\}$.

1.1.4 Random Variables, Stochastic Processes, & Their Classification

Consider a probability space (Ω, \mathcal{F}, P) .

Definition 1.7 (Random Variable/ Vector). A **random variable** is function that maps Ω to $\mathcal{X} \subseteq \mathbb{R}$ i.e. $\omega \rightarrow x(\omega)$. Ω and \mathcal{X} are the domain and range of the random variable. A **random vector**: function that maps Ω to \mathbb{R}^N .

Definition 1.8 (Univariate Stochastic Process). A **univariate stochastic process** with range \mathcal{X} is a collection of \mathcal{X} -valued random variables indexed by a set \mathcal{T} (e.g. "time") $\{x_t : t \in \mathcal{T}\}$ where each x_t is an \mathcal{X} -valued random variable.

If \mathcal{T} had a finite number of elements, we could represent the stochastic process as a random vector! But this is often impractical, and in fact, it is impossible if \mathcal{T} contains an uncountable number of elements. For that reason, it is **better to regard a stochastic process as a random function**.

Definition 1.9 (Random Function). A **random function** is a collection of random variables defined over the same Ω and indexed with parameter $t \in \mathcal{T}$ i.e. $x : \mathcal{T} \times \Omega \rightarrow \mathbb{R}$, $x(t, \omega)$.

$\mathcal{T} \times \Omega$ is the Cartesian product (i.e., ordered pairs) of sample space and index space. Randomness comes from ω .

Definition 1.10 (Realisation). A **realisation** (or path) of the stochastic process is the value that this random function takes for all values of $t \in \mathcal{T}$ for a fixed ω .

In contrast, x_t is an RV for any given $t \in \mathcal{T}$. Real experiments which generate a stochastic process are not, in general, replicable, so it is like having a sample of size 1.

Definition 1.11 (Index Set-Based Classification of Stochastic Process). We can **classify them according to cardinality**:

1. for the **discrete-time parameter**:
 - ☑ **regular**: the index difference is always the same; e.g., sample VIX at 4pm every day,
 - ☑ **irregular**: e.g., ultra-high-frequency financial data
2. for the **continuous time parameter**: the index parameter is a time interval.

We can classify them according to the **dimension of the index parameter**:

1. **unidimensional**: t is scalar; e.g., time series; and
2. **multidimensional**: t is a vector with more than one index; e.g., meteorology (latitude, longitude, altitude, time) random field.

Definition 1.12 (Random Variable-Based Classification of Stochastic Process). We can classify it according to the **range**:

1. **real**: RVs take values on \mathbb{R} ; and
2. **complex**: $a_t + b_t i$, with $a_t, b_t \in \mathbb{R}$.

We can classify it according to the **size**:

1. **univariate**: RV is a scalar f over $\mathbb{R}(\mathbb{C})$; and
2. **multivariate**: RV is a vector f over $\mathbb{R}^N (\mathbb{C}^N)$.

We can classify it according to the **state space**:

1. **discrete**: RV is discrete, e.g., a binary RV reflecting business cycle; and
2. **continuous**: obvious, e.g. a *Gaussian* RV.

Definition 1.13 (Nature-Based Classification of Stochastic Process). We can classify it as:

1. **stock**: time series that can be measured at any instant of time; e.g., exchange rates, stock prices and inventories; and
2. **flow**: it accounts for those RVss that cannot be measured at any instant because the flow is infinitesimal; e.g., consumption, income and many other

1.1.5 Characterisation of Univariate Time Series

RVs are fully characterised by their joint distribution or characteristic function, which is a one-to-one transformation of the joint distribution.

For TS, the complete characterisation is done by the distribution (or characteristic) function of sequences of their elements. For example:

$$F_{X_{t_1}, X_{t_2}, \dots, X_{t_j}}(x_{t_1}, x_{t_2}, \dots, x_{t_j}) = \mathbb{P}(X_{t_1}(\omega) \leq x_{t_1}, \dots, X_{t_j}(\omega) \leq x_{t_j}) \quad (1.6)$$

In practice this may be inconvenient (i.e. dealing with very large vectors); we could instead focus on their moments. Recall that joint distribution gives full characterisation of a random vector; but is not easy for random functions. For that reason, we will **often focus on a set of moments, even though this only provides a partial characterisation**. Even so, we need to exploit the concept of stochastic equilibrium (a.k.a. steady state). At the end of the day, we only observe one realisation. How can we make inferences? We need additional assumptions.

1.2 Stationarity

Definition 1.14 (Trend Function).

$$\mu_t = \mathbb{E}(X_t) = \int_{-\infty}^{\infty} x dF_{X_t}(x) \quad (1.7)$$

Definition 1.15 (Mean-Stationarity). A **mean-stationary stochastic process** is such that the mean (across paths) is the same regardless of the instant of time t .

Define the following:

$$\gamma_{t,s} \equiv \text{cov}(X_t, X_s) = \mathbb{E}[(X_t - \mu_t)(X_s - \mu_s)]; \text{ and} \quad (1.8a)$$

$$\gamma_{t,t} = \mathbb{V}(X_t) = \sigma_t^2 \quad (1.8b)$$

The **autocovariance function is positive semi-definite**.

Definition 1.16 (Covariance Stationarity). A **covariance stationary stochastic process** is such that the following conditions are met:

1. **mean-stationarity**;
2. **constant variance across time**; and
3. covariance structure **depend only on $|t - s|$** , that is $\gamma_{t,s} = \gamma_{|t-s|}$.

1.2.1 Autocorrelation Function

Even though it entails a loss of information concerning γ_s , it is very popular precisely because it is unitless.

Definition 1.17 (Autocorrelation Function). The covariance function is given by:

$$\rho_{t,s} = \frac{\text{Cov}(X_t, X_s)}{\sqrt{\mathbb{V}(X_t)\mathbb{V}(X_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} \in [-1, 1]. \quad (1.9)$$

Again, this function is **symmetric and positive (semi-)definite**. Examples include:

› *White noise* (WN):

$$u_t : t \in \{T_{-i}, \dots, 0, \dots, T_i\} \text{ such that} \quad (1.10a)$$

$$\mathbb{E} u_t = 0; \quad (1.10b)$$

$$\mathbb{V}(u_t) = \sigma_u^2 < \infty; \text{ and} \quad (1.10c)$$

$$\text{cov}(u_t, u_s) = \mathbb{I}\{t = s\} \sigma_u^2. \quad (1.10d)$$

› *Harmonic series*;

› *Wiener process* (Brownian motion):

$$t \in [0, T_f] \quad \wedge \quad W(0) = 0 : \quad (1.11a)$$

$$\mathbb{E}[W(t) - W(s)] = 0 \quad \forall t, s; \quad (1.11b)$$

$$\mathbb{V}[W(t) - W(s)] = \sigma^2(t - s) \quad \forall t \geq s; \quad (1.11c)$$

$$\text{cov}[W(t) - W(s), W(v) - W(w)] = 0 \quad \forall t \geq s \geq v \geq w; \quad (1.11d)$$

$$\mathbb{E} W(t) = \mathbb{E}[W(t) - W(0) + W(0)] = 0 \quad \forall t; \quad (1.11e)$$

$$\mathbb{V}[W(t)] = t\sigma^2; \text{ and} \quad (1.11f)$$

$$\underbrace{\text{cov}[W(t), W(s)] = \text{cov}[W(t) - W(s) + W(s), W(s)] = s\sigma^2.}_{\text{Not Covariance Stationary.}} \quad (1.11g)$$

› *Factor process*:

$$x_t = a + u_t, \quad u_t \text{ is WN}; \quad (1.12a)$$

$$a \text{ is a RV, } a \perp \{u_t\}; \quad (1.12b)$$

$$\mathbb{E} a = 0 \quad \wedge \quad \mathbb{V}(a) = \sigma_a^2 < \infty; \quad (1.12c)$$

$$\mathbb{E} x_t = \mathbb{E} a + \mathbb{E} u_t = 0; \quad (1.12d)$$

$$\mathbb{V}(x_t) = \mathbb{V}(a) + \mathbb{V}(u_t) = \sigma_a^2 + \sigma_u^2 < \infty; \quad (1.12e)$$

$$\text{cov}(x_t, x_s) = \text{cov}(a + u_t, a + u_s) = \mathbb{V}(a) = \sigma_a^2; \text{ and} \quad (1.12f)$$

$$\underbrace{\text{corr}(x_t, x_s) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2}}_{\text{Covariance Stationarity.}} \quad (1.12g)$$

› *Binary Markov Chain*.

1.2.2 Strict Stationarity & Ergodicity

Definition 1.18 (Strict Stationarity). A process is **strictly stationary**^{1.1} if:

^{1.1} It is a **much stronger process than covariance stationarity**!

1. marginals are the same for all t ; and
2. the joint distribution of two, three, or more elements of the stochastic process no matter which two, three, or more values of the index one chooses only depends on the distance between the values of the index.

Definition 1.19 (Ergodicity). **Ergodicity** essentially requires that any path is fully representative of the stochastic process.

This condition is important for making inferences about population characteristics from a single path.

Example 1.4 (Factor Process, Again). Consider:

$$x_t = a + u_t \implies \quad (1.13a)$$

$$\mathbb{E} x_t = 0 \implies \quad (1.13b)$$

$$\frac{1}{T} \sum_t x_t = \underbrace{\frac{1}{T} \sum_t a}_{\rightarrow a} + \underbrace{\frac{1}{T} \sum_t u_t}_{\rightarrow 0} = a. \quad (1.13c)$$

Use Chebyshev's LLN.

This implies that the process is not ergodic! You **cannot learn about any of the moments from one particular realisation**. This is an example of a **covariance stationary process without ergodicity**! The root of the problem is the fact that the autocorrelations do not “die down” as we move away between two processes.^{1,2}

1.3 Prediction

Point forecasts are easy to understand but in the continuous state world, they are wrong with probability 1. **Interval forecasts** are better. Suppose we are interested in forecasting y , we can set $[\hat{y}_L, \hat{y}_U]$ s.t. $\mathbb{P}(y \in [\hat{y}_L, \hat{y}_U]) = 1 - \alpha$. An example includes the *Value at Risk* measure (= quantile of financial losses). **Distribution forecasts** are more informative.

Example 1.5 (Bank of England's Distribution Forecasts). The BoE provides inflation forecasts every month for the following two years with their corresponding quantiles. The “Old Lady's” density forecast takes the form:

$$g(\pi_{t+h} | I_t) = \begin{cases} k_t \exp \left[0.5 \times \frac{(\pi_{t+h} - \mu_{t+h|t})^2}{\sigma_{lt+h|t}^2} \right] & \text{if } \pi_{t+h} < \mu_{t+h|t} \\ k_t \exp \left[0.5 \times \frac{(\pi_{t+h} - \mu_{t+h|h})^2}{\sigma_{rt+h|t}^2} \right] & \text{if } \pi_{t+h} \geq \mu_{t+h|t} \end{cases} \quad \text{where} \quad (1.14a)$$

$$k_t = \frac{2}{\sqrt{2\pi} (\sigma_{lt+h|t}^2 + \sigma_{rt+h|t}^2)}. \quad (1.14b)$$

The **density forecast** is completed by some choice for:

1. the mode/median, $\mu_{t+h|t}$;
2. the “downside” variance, $\sigma_{lt+h|t}^2$; and
3. the “upside” variance, $\sigma_{rt+h|t}^2$.

The members of the *Monetary Policy Committee* devote multiple sessions to those three choices. The resulting density is leptokurtic unless the two variances are equal. It is also asymmetric under the same circumstances because: UNTIL SLIDE 29

^{1,2} The process of ergodicity was introduced by Markov to show that LLN and CLT could apply without i.i.d. data, but with ergodicity (under certain conditions).

1.3.1 Loss Function(s)

$$e \equiv y - \hat{y}. \quad (1.15)$$

We consider a few types of **ex-post loss functions**.

Quadratic [$C(e) = e^2$] The **quadratic loss function** has the following qualities:

- 🏠 affine transformations don't change the result;
- 🏠 overestimating \sim underestimating;
- 🏠 small forecast errors are “less costly”;
- 🏠 it worries too much about making very big mistakes; and
- 🏠 the solution is:

$$\hat{y}^* = \mathbb{E}(y \mid x = x_0) \quad (1.16)$$

because the cond. mean minimises the quadratic loss.

Absolute [$C(e) = |e|$] The **absolute loss function** has the following qualities:

- 🚗 overestimating \sim underestimating.
- 🚗 the concern is proportional to the size of forecast errors.
- 🚗 the solution:

$$\hat{y}^* = \text{median}(y \mid x = x_0). \quad (1.17)$$

Lin-lin [$C(e) = \tau|e|I(e < 0) + (1 - \tau)|e|I(e > 0)$] ^{1.3} The **lin-lin loss function** has the following qualities:

- 👤 it worries more about $+$ ($-$) errors and less about $-$ ($+$) mistakes; e.g., $\tau < 1/2$ [$> 1/2$]; and
- 👤 the solution:

$$\hat{y}^* = Q_\tau(y \mid x = x_0). \quad (1.18)$$

Quad-quad [$C(e) = \tau e^2 I(e < 0) + (1 - \tau)e^2 I(e > 0)$] The **quad-quad loss function** has the following qualities:

- 🍏 it worries more about $+$ ($-$) **big** errors; e.g., $\tau < \frac{1}{2}$ [$> \frac{1}{2}$]; ^{1.4} and
- 🍏 the solution:

$$\hat{y}^* = \text{expectile}_\tau(y \mid x = x_0). \quad (1.19)$$

Now, consider **ex-ante forecasts** (with a little bit more detail). Let $h(x)$ denote the forecast rule. Notice the difference with respect to the **interim forecast** (just a standard optimisation problem); here instead, one has to choose among all possible rules:

$$h^*(x) = \arg \underbrace{\left\langle \min_{h(x) \in H} \{ \mathbb{E}_{X,Y} [c[y - h(x)]] \} \right\rangle}_{=\text{Minimisation Over Functions in Space } H}. \quad (1.20)$$

This is ex-ante; i.e., don't know x , so I average over joint (x, y) distribution. If no restrictions on H , just solve the interim problem for each x . In practice, often, it is convenient to restrict the set of rules; e.g., a linear $h(x) = \alpha^T x$. There is an **important distinction between linear and affine**. The affine rules include an intercept.

Definition 1.20 (Affine Transformation). An **affine** transformation contains a **constant**.

^{1.3} This is the so-called **check function**.

^{1.4} The loss function is an **asymmetric parabola**.

1.3.2 BLP & LS Prediction Rule(s)

Quadratic BLP

$$h^*(x) = \mathbb{E}(yx^T) \mathbb{E}^{-1}(xx^T) x \quad (1.21)$$

If we have an intercept then substitute expectations of cross-products by covariances (the so-called **Best Affine Predictor**).

Lin-Lin BLP We get a linear quantile projection which has no closed form. The special case is the linear median regression:

$$\min_{\alpha^T x} \{ \mathbb{E} [|y - \alpha^T x|] \} \quad (1.22)$$

Quad-Quad BLP This is solved by a linear expectile projection which has no closed form.

The message is that **prediction under linear rules is not better than an unrestricted prediction but it is easier to implement.**

Least Squares Prediction Rules Suppose you wish to predict y over a set of admissible rules H . E.g. in a linear world, let $H = \{h(x) = \alpha^T x : \alpha \in \mathbb{R}^N\}$ be the linear span of x . Also let L^2 be the set of RVs defined under (Ω, \mathcal{F}, P) whose second moments are finite; i.e. $\mathbb{E}(X^2) < \infty$.

The **LS prediction problem** is as follows. Given y and H , both in L^2 the idea is to minimise the distance of the prediction, $y - h(x)$, in terms of MSE:

$$h^* \in H : \mathbb{E} [(y - h^*)^2] \leq \mathbb{E} [(y - h)^2] \quad \forall h \in H. \quad (1.23)$$

For a solution to exist, H should be closed and linear subspace of L^2 (otherwise, the solution on the boundary might not belong to H). Since H is a set of functions, we need a **metric (mean sq.)**:

$$\underbrace{h_n \xrightarrow{m.s.} h \quad \text{if} \quad \lim_{n \rightarrow \infty} \mathbb{E} [(h_n - h)^2]}_{\text{Limit Definition in This Metric Space.}} \quad (1.24)$$

Define:

$$h^* \equiv \mathbb{P}(y | H), \quad (1.25)$$

i.e., the projection of y onto H (minimises the quadratic loss function). The **solution is characterised by the “orthogonality condition:”**

$$\mathbb{E} [(y - h^*) h] = 0, \quad \forall h \in H \quad (1.26)$$

In other words, the **prediction error using the optimal rule is going to be orthogonal to all possible rules.**

Example 1.6. Consider the following H s.

■ $H = \{c\}$ **only constant rules** (constant prediction; then H is the real line). Then:

$$\mathbb{P}(y | \{c\}) = \mathbb{E} y \quad (1.27)$$

which is the unconditional mean.

□ $H = L^2$ (i.e., **any prediction rule with bounded second moment**) implies:

$$\mathbb{P}(y | L^2) = y. \quad (1.28)$$

■ $H = \{h = \alpha^T x : \alpha \in \mathbb{R}^k, k = \dim(x)\}$ (**linear projections**), then

$$h^*(x) = \mathbb{E}(yx^T) \mathbb{E}^{-1}(xx^T)x \quad (1.29)$$

assuming x has a full-rank second-moment matrix. If not h^* is still defined uniquely but a different formula should be used instead).

□ Let \mathcal{G} denote the set of all measurable functions of x with bounded second moments; i.e., $\mathcal{G} \subset \mathcal{L}$. Then:

$$\mathbb{P}(y | \mathcal{G}) = \mathbb{E}(y | x) \quad (1.30)$$

which is the regression function. **Conditional expectations are a particular case of projections.**

1.3.3 Properties of Projection

Theorem 1.2 (Properties of Projections). *Projections have the following properties:*

1. **linearity** - for $a_1, a_2 \in \mathbb{R}$ and $y_1, y_2 \in L^2$:

$$\mathbb{P}(a_1 y_1 + a_2 y_2 | H) = a_1 \mathbb{P}(y_1 | H) + a_2 \mathbb{P}(y_2 | H), \quad (1.31)$$

which **holds for any h (linear or not)**;

2. the **updating rule** - let $H, I \subset L^2$ with $H \subseteq I$, then:

$$\mathbb{P}(y | I) = \mathbb{P}(y | H) + \mathbb{P}(y | U) = \mathbb{P}(y | H \oplus U), \quad (1.32)$$

where U denotes info in I that was unpredictable only knowing H , formally, $U = \{u \in I : \mathbb{P}(u | H) = 0\}$ so $u \perp H$ and $u = i - \mathbb{P}(i | H)$ for $i \in I$,^{1.5}

3. the **Law of Iterated Projections** - let $H, I \subset L^2$ with $H \subseteq I$, then:

$$\mathbb{P}(y | H) = \mathbb{P}[\mathbb{P}(y | I) | H], \quad (1.33)$$

where **Law of Iterated Expectations** is a particular case and it also works for linear projections or any other kind of projections.

1.3.4 Time Series Context

Let $y = x_t$ and we have the history of $x_t (x_{t-1}, x_{t-2}, \dots, x_0)$. In practice, $x_{t-1}, x_{t-2}, \dots, x_0$ could be potentially very large. Instead, we first study a solution based on $x_{t-1}, x_{t-2}, \dots, x_0, \dots, x_{-\infty}$.

Linear Case Denote:

$$H_{t-1, \tau} = \left\{ \sum_{j=0}^{\tau} \alpha_j x_{t-1-j} \right\}, \quad (1.34)$$

which is the space of linear functions that consider $x_{t-1}, x_{t-2}, \dots, x_{t-\tau-1}$. Hence, the projection is:

$$\mathbb{P}(x_t | H_{t-1, \tau}) = \mathbb{E}[x_t (x_{t-1}, \dots, x_{t-\tau-1})] \times \mathbb{E}^{-1} \left[\begin{pmatrix} x_{t-1}^2 & \cdots & x_{t-1}x_{t-\tau-1} \\ \vdots & \ddots & \vdots \\ x_{t-1}x_{t-\tau-1} & \cdots & x_{t-\tau-1}^2 \end{pmatrix} \right] \begin{pmatrix} x_{t-1} \\ \vdots \\ x_{t-\tau-1} \end{pmatrix}. \quad (1.35)$$

This is too big! Define

$$H_{t-1} \equiv \overline{\bigcup_{\tau=0}^{\infty} H_{t-1, \tau}}. \quad (1.36)$$

^{1.5} Intuitively, U is the set of surprises in the prediction rules in I , given the prediction rules in H .

We close it because a **countable union of closed sets is not necessarily closed, and we need it for existence**. Hence:

$$\mathbb{P}(x_t | H_{t-1}) = MS \lim_{\tau \rightarrow \infty} \mathbb{P}(x_t | H_{t-1, \tau}), \quad (1.37)$$

where $MS \lim$ is the mean-squared limit!

Non-Linear Case Define:

$$G_{t-1, \tau} = \{ \text{measurable functions } g \in L^2 : g = f(x_{t-1}, \dots, x_{t-\tau-1}) \}, \quad (1.38)$$

Then:

$$\mathbb{P}(x_t | G_{t-1, \tau}) = \mathbb{E}[x_t | x_{t-1}, \dots, x_{t-\tau-1}] \quad (1.39)$$

Define:

$$G_{t-1} \equiv \overline{\bigcup_{\tau=0}^{\infty} G_{t-1, \tau}}. \quad (1.40)$$

I close it because a countable union of closed sets is not necessarily closed, and we need it for existence). Hence, we have

$$\mathbb{P}(x_t | G_{t-1}) = MS \lim_{\tau \rightarrow \infty} \mathbb{P}(x_t | G_{t-1, \tau}) = \mathbb{E}(x_t | x_{t-1}, x_{t-2}, \dots) = \mathbb{E}_{t-1}(x_t) \quad (1.41)$$

Let:

$$V_t \equiv \{v : v = g - \mathbb{P}(g | G_{t-1}), g \in G_t\} \quad (1.42)$$

which is the orthogonal space to G_{t-1} in G_t . So $G_t = G_{t-1} \oplus V_t$ and

$$x_t = \mathbb{P}(x_t | G_t) \implies \quad (1.43a)$$

$$x_t = \mathbb{P}(x_t | G_{t-1}) + \mathbb{P}(x_t | V_t) \implies \quad (1.43b)$$

$$x_t = \mathbb{P}(x_t | G_{t-2}) + \mathbb{P}(x_t | V_{t-1}) + \mathbb{P}(x_t | V_t) \implies \quad (1.43c)$$

$$x_t = \sum_{k=0}^{\infty} \mathbb{P}(x_t | V_{t-k}) + \mathbb{P}(x_t | G_{-\infty}), \quad (1.43d)$$

where:

$$G_{-\infty} \equiv \bigcap_{k=0}^{\infty} G_{t-k}. \quad (1.44)$$

Any time series can be written as a sum of updates plus the projection at the beginning of times. **Successive updates in prediction constitute a martingale difference**.

Definition 1.21 (Martingale Sequence). y_t, y_{t-1}, \dots is a **martingale sequence** if $\mathbb{E}[y_t | y_{t-1}, y_{t-2}, \dots] = y_{t-1}$. That is, each incremental gain has zero mean given the history of the game up to that point.

Example 1.7 (Martingale Examples). The following gives some examples.

 Consider a fair game of chance, in which net gains have zero mean. Then, its cumulative gain is a Martingale.

 Forecast sequence for fixed forecast variable $y_t = \mathbb{E}[x_{300} | x_t, x_{t-1}, \dots]$.

Definition 1.22 (Martingale Difference Sequence). A **martingale difference sequence** meets the following condition.

$$\varepsilon_t, \varepsilon_{t-1}, \dots \quad \text{where } \varepsilon_t = y_t - y_{t-1} \quad (1.45)$$

is the **first increment in a martingale sequence**.

Example 1.8 (Martingale Difference Examples). Below, you can find examples of the martingale difference sequences:

✘ incremental gains in a fair game;

✘ revision in the sequence of forecasts in the previous example; and

✘ one-period ahead forecast errors.

1.4 The Wold Representation Theorem

1.4.1 White Noise (WN)

In a linear world, the analogue to a **martingale difference sequence** is a **WN sequence**:

$$u_t = x_t - \mathbb{P}(x_t | H_{t-1}) \quad (1.46)$$

which is one period ahead of linear prediction error. u_t is then **linearly unpredictable** (uncorrelated with x_{t-1}, \dots), but it could be non-linearly predictable.

Definition 1.23 (Orthogonal Sum). $X_t \oplus Y_t$ represents the elements of Y_t that are orthogonal to all of the elements of X_t .

Since $H_t = H_{t-1} \oplus U_t$, we can conduct the following decomposition:

$$x_t = \mathbb{P}(x_t) \implies \quad (1.47a)$$

$$x_t = \mathbb{P}(x_t | H_{t-1}) + \mathbb{P}(x_t | U_t) \implies \quad (1.47b)$$

$$x_t = \mathbb{P}(x_t | H_{t-2}) + \mathbb{P}(x_t | U_{t-1}) + \mathbb{P}(x_t | U_t) \implies \quad (1.47c)$$

$$x_t = \sum_{k=0}^{\infty} \mathbb{P}(x_t | U_{t-k}) + \mathbb{P}(x_t | H_{-\infty}), \quad (1.47d)$$

where:

$$H_{-\infty} \equiv \bigcap_{k=0}^{\infty} H_{t-k} \quad (1.48)$$

However, there's an **important difference to what we see in the previous sub-chapter**:

$$\mathbb{P}(x_t | U_t) = \frac{\mathbb{E}(x_t u_t)}{\mathbb{E}(u_t^2)} u_t \quad (1.49)$$

because U_t is a **unidimensional linear space spanned by u_t** . It can be shown that:

$$x_t = \sum_{j=0}^{\infty} \delta_{t,t-j} u_{t-j} + \mathbb{P}(x_t | H_{-\infty}) \quad (1.50)$$

The problem is that $\delta_{t,t-j}$ depend on t . The solution is to **apply the equilibrium concept that says that the stochastic behaviour is the same irrespective of t** . If we have covariance stationarity:

$$\frac{\mathbb{E}(x_t u_{t-j})}{\mathbb{E}(u_{t-j}^2)} = \delta_j \quad \forall t. \quad (1.51)$$

Theorem 1.3 (Wold Representation Theorem). *This way, we get the Wold representation theorem that states:*

$$x_t = \underbrace{\sum_{j=0}^{\infty} \delta_j u_{t-j}}_{=\text{Prediction Errors}} + \mathbb{P}(x_t | H_{-\infty}) \quad (1.52)$$

This decomposes the process into a linearly deterministic part and a linearly regular part. Importantly, the WN driving the linearly regular part is the one-period ahead linear prediction errors in x_t .

Let

$$A \equiv \sum_{j=0}^{\infty} \delta_j u_{t-j} \quad (1.53)$$

denote the **linearly regular part of the stochastic process**.

Similarly, let

$$B \equiv \mathbb{P}(x_t \mid H_{-\infty}) \quad (1.54)$$

denote its **linearly deterministic part**. Since **both parts are orthogonal**, $V(x_t) = V(A) + V(B)$

$$\mathbb{V}(x_t) = \mathbb{V}(A) + \mathbb{V}(B) \quad (1.55)$$

Since the process is **covariance stationary**, $V(x_t)$ **must be bounded**. Hence:

$$\mathbb{V}(A) = \mathbb{V}\left[\sum_{j=0}^{\infty} \delta_j u_{t-j}\right] = \sigma_u^2 \sum_{j=0}^{\infty} \delta_j^2 \quad (1.56)$$

We must **have square summable Wold coefficients** $\sum_{j=0}^{\infty} \delta_j^2 < \infty$!

1.4.2 Uses of the Wold Representation Theorem

Forecasts $\mathbb{P}(x_{t+k} \mid H_{t-1})$, replace x_{t+k} by Wold representation theorem, and then apply linearity:

$$\mathbb{P}(x_{t+k} \mid H_{t-1}) = \sum_{j=1}^{\infty} \delta_{k+j} u_{t-j} + \mathbb{P}(x_{t+k} \mid H_{-\infty}). \quad (1.57)$$

Forecast Updates Consider:

$$\mathbb{P}(x_{t+k} \mid H_t) - \mathbb{P}(x_{t+k} \mid H_{t-1}) = \delta_k u_t, \quad (1.58)$$

which is **the new info that has arrived today**.

Autocovariance Structure of the Process

$$\text{cov}(x_t, x_{t-k}) = \text{cov}\left[\sum_{j=0}^{\infty} \delta_j u_{t-j}, \sum_{j=0}^{\infty} \delta_j u_{t-k-j}\right] + \text{cov}[\mathbb{P}(x_t \mid H_{-\infty}), \mathbb{P}(x_{t-k} \mid H_{-\infty})] \quad (1.59)$$

We would to find this for any k , so:

$$\gamma_k = \sigma_u^2 \sum_{j=0}^{\infty} \delta_j \delta_{j-k} \quad (1.60)$$

because $\text{cov}(u_{t-j}, u_{t-k-j}) = 0$ whenever $j \neq 0$.

Autocovariance Generating Function

$$\psi_{xx}(z) = \sum_{j=-\infty}^{\infty} \gamma_{xx}(j) z^j; \text{ where} \quad (1.61a)$$

$$\gamma_{xx}(k) = \text{cov}(x_t, x_{t-k}). \quad (1.61b)$$

This only works for covariance stationary processes. Assume $\delta_j = 0$ if $j < 0$ and take $\gamma_k = \sigma_u^2 \sum_{j=0}^{\infty} \delta_j \delta_{j-k}$, some algebra shows that

$$\psi_{xx}(z) = D(z)D(z^{-1}) \cdot \sigma_u^2 \quad (1.62)$$

with:

$$D(z) \equiv \delta_0 + \delta_1 z + \cdots = \sum_{j=0}^{\infty} \delta_j z^j. \quad (1.63)$$

1.5 Frequency Domain

1.5.1 A Digression: Cyclic Functions of Constant Period

Although cyclic processes of the constant period are rare in economics, their study is **useful to understand time series in the frequency domain**. The sine and cosine functions are the best-known examples of constant cycle processes. Their values (measured in radians) are repeated every 2π periods since :

$$\sin t = \sin(t + k \cdot 2\pi); \text{ and} \quad (1.64a)$$

$$\cos t = \cos(t + k \cdot 2\pi) \quad \forall k \in \mathbb{Z}. \quad (1.64b)$$

Furthermore, they are bounded between -1 and 1 , and their position along the horizontal axis is known. For example:

$$\sin(0) = 0; \quad (1.65a)$$

$$\cos(0) = 1; \quad (1.65b)$$

$$\sin\left(\frac{\pi}{2}\right) = 1; \quad (1.65c)$$

$$\cos\left(\frac{\pi}{2}\right) = 0; \quad (1.65d)$$

$$\sin(\pi) = 0; \quad (1.65e)$$

$$\cos(\pi) = -1; \quad (1.65f)$$

$$\sin\left(\frac{3\pi}{2}\right) = -1; \text{ and} \quad (1.65g)$$

$$\cos\left(\frac{3\pi}{2}\right) = 0. \quad (1.65h)$$

Therefore, they have many limitations to represent the general constant cycle process. However, they can be easily generalised.

First of all, we can achieve a **contraction or expansion along the time axis if we introduce a parameter λ multiplying the argument**, namely $\sin \lambda t$ and $\cos \lambda t$. As a result, the **cycle period (or wavelength), which is the time it takes for the function to complete one cycle**, goes from 2π to $2\pi/\lambda$. The parameter λ is known as the (angular) frequency of the process, and divided by 2π gives us the number of cycles per unit of time, which is nothing other than the reciprocal of the wavelength. For example, if we worked with monthly data, the frequency that corresponds to cycles with a period of one year would be $\frac{\pi}{6}$.

Likewise, if we **subtract from the argument of the sine and cosine functions a parameter $\theta \in [-\pi, \pi]$ known as the phase, we will manage to move these functions along the horizontal axis** by an amount θ/λ . For example, the sine can be obtained from the cosine by a phase shift of $\pi/2$ since $\cos(t \pm \pi/2) = \mp \sin t$. Naturally:

$$\sin(t \pm \pi/2) = \pm \cos t \quad (1.66)$$

However, since

$$\cos(\lambda t - \theta) = \cos \theta \cos \lambda t + \sin \theta \sin \lambda t, \quad (1.67)$$

another way to achieve the same effect is through a linear combination of $\sin \lambda t$ and $\cos \lambda t$ with coefficients whose Euclidean norm is 1 so that the resulting function remains between -1 and 1 .

In this respect, another simple way to generalise the sine and cosine functions is by **multiplying them by another parameter ρ** , called the amplitude, which allows the range of these functions to be between $-\rho$ and ρ . Finally, a **displacement along the vertical axis** is obtained by adding a constant μ .

Trigonometric functions have fascinated mathematicians since antiquity, but the most useful relationships between them are:

$$\sin^2 \lambda + \cos^2 \lambda = 1 \quad (\text{Pythagorean Identity}) \quad (1.68a)$$

$$e^{i\lambda} = \cos \lambda + i \sin \lambda, \quad (\text{Euler's Identity}) \quad (1.68b)$$

which in turn yields:

$$\underbrace{\cos \lambda = \Re \left(e^{i\lambda} \right) = \frac{e^{i\lambda} + e^{-i\lambda}}{2}}_{=\text{Real Part}}, \text{ and} \quad (1.69a)$$

$$\underbrace{\sin \lambda = \Im \left(e^{i\lambda} \right) = \frac{e^{i\lambda} - e^{-i\lambda}}{2i}}_{=\text{Imaginary Part}}. \quad (1.69b)$$

This also leads to:

$$\sin(\lambda_1 \pm \lambda_2) = \sin \lambda_1 \cos \lambda_2 \pm \cos \lambda_1 \sin \lambda_2 \quad (1.70a)$$

$$\cos(\lambda_1 \pm \lambda_2) = \cos \lambda_1 \cos \lambda_2 \mp \sin \lambda_1 \sin \lambda_2. \quad (1.70b)$$

1.5.2 Spectral Analysis

So far, we have analysed the structure of the stochastic process in the time domain using the autocorrelation function and the lag structure given by the Wold representation. Let us now characterise stochastic processes by **their frequency domain properties**. Frequency domain methods were very popular in the 1960s and 1970s, but by 1990 they had faded away from mainstream time series analysis, except in certain areas, such as seasonality, long memory, or filtering theory.

Harvey's (1989) treatment of spectral inference methods in his book *Structural Time Series Models and the Kalman Filter* is often on par with his treatment of time domain ones. The advent of fast computers and the development of reliable recursive algorithms contributed to the decline of spectral analysis. Still, I would argue that reports of the death of spectral methods in time series have been greatly exaggerated.

We initially consider covariance stationary, linearly deterministic processes. In particular, we will begin by analysing in some additional **detail the harmonic process with frequency λ , $x_t = a \cos \lambda t + b \sin \lambda t$** , which is the most general constant cycle process that we can obtain from sines and cosines.

Definition 1.24 (Harmonic Processes). Consider the following process:

$$x_t = \cos(\lambda t) + b \sin(\lambda t); \text{ where:} \quad (1.71a)$$

$$\lambda \in \mathbb{R}; \quad (1.71b)$$

$$t \in [-T_i, \dots, 0, \dots, T_f] \text{ or } Tt \in [-T_i, T_f]; \text{ and} \quad (1.71c)$$

$$a, b \text{ RVs.} \quad (1.71d)$$

This is referred to as a **harmonic process**.

Specifically:

$$a \cos \lambda t + b \sin \lambda t = \sqrt{a^2 + b^2} \cos \left[\lambda t - \arctan \left(\frac{b}{a} \right) \right] \implies \quad (1.72a)$$

$$a \cos \lambda t + b \sin \lambda t = \sqrt{a^2 + b^2} \left\{ \cos \left[\arctan \left(\frac{b}{a} \right) \right] \cos \lambda t + \sin \left[\arctan \left(\frac{b}{a} \right) \right] \sin \lambda t \right\}. \quad (1.72b)$$

Thus, we can directly interpret **the polar coordinates of the pair (a, b)** . Specifically, $\sqrt{a^2 + b^2}$ is **the amplitude of the harmonic process**, and $\arctan \left(\frac{b}{a} \right)$ **its phase shift**.

Lemma 1.4. *As we have seen, **this process is stationary** if:*

$$\mathbb{E}(a) = \mathbb{E}(b) = 0; \quad (1.73a)$$

$$\mathbb{V}(a) = \mathbb{V}(b) = \sigma^2; \text{ and} \quad (1.73b)$$

$$\mathbb{E}(ab) = 0 \quad (1.73c)$$

Remark. *For the **harmonic processes**:*

$$\mathbb{E}(x_t) = 0; \quad (1.74a)$$

$$\mathbb{V}(x_t) = \sigma^2; \text{ and} \quad (1.74b)$$

$$\gamma_{t,s} = \text{cov}(x_t, x_s) = \sigma^2 \cos(\lambda s). \quad (1.74c)$$

All variability in the series is trivially associated with a cycle of period $2\pi/\lambda$. Note that this process is **not ergodic**!

Remark. *It is important to highlight that if **the parameter t of a real harmonic process is discrete and takes values $t = 0, \pm 1, \pm 2, \pm 3, \dots$, we can only consider frequencies that are between 0 and π** because for any other frequency, there are alternative processes with either $\lambda < 0$ or $\lambda > \pi$ that would produce indistinguishable realisations given our observation period.*

Definition 1.25 (Nyquist Frequency). Nyquist's theorem states that a periodic signal must be sampled at more than twice the highest frequency component of the signal.

The reason is that on the one hand $\cos \lambda t = \cos(-\lambda t)$ and $\sin \lambda t = -\sin(-\lambda t)$, and on the other $\cos(\lambda t + k\pi t) = \cos \lambda t$ and $\sin(\lambda t + k\pi t) = \sin \lambda t$ for k even, and $\cos(\lambda t + k\pi t) = \cos(\pi - \lambda)t$ and $\sin(\lambda t + k\pi t) = -\sin(\pi - \lambda)t$ for k odd. This is the so-called **(temporal) aliasing problem** because if the true process evolves at a frequency higher than the Nyquist one, which is the highest observable, then it will be confused with a lower frequency process, and thus assigned the wrong “alias.” Importantly, **it implies that for monthly data, the shortest cycle that can be observed, which corresponds to the highest frequency, is two months.**

Example 1.9 (Old Movies & Trains). In analogue movies, temporal aliasing results from the low frame rate of the underlying sequence of photographs, which causes the so-called wagon-wheel effect, whereby a spoked wheel appears to rotate too slowly or even with a reversal of direction, as in a seemingly negative frequency.

If the observations occurred at $t = 0, \pm\Delta t, \pm 2\Delta t, \pm 3\Delta t$, the highest frequency would be $\pi/\Delta t$. The lowest frequency is $\lambda = 0$, in which case the stochastic process becomes a constant one. Thus, for real processes observed once per period, we can use 0 to π . For complex processes, we could distinguish frequencies from $-\pi$ to π . In any event, the **harmonic process is empirically unrealistic for economic and financial time series.**

However, it can be easily generalised by **superposition**. Specifically, we can form **linear combinations of harmonic processes at different frequencies**:

$$x_t = \sum_{j=1}^m a_{\lambda_j} \cos \lambda_j t + b_{\lambda_j} \sin \lambda_j t = \sum_{j=1}^m r_j \cos(\lambda_j t - \theta_j). \quad (1.75)$$

A special case of some interest is the so-called purely periodic process, which is such that $x_t = x_{t-\tau}$, for $\tau \in \mathbb{N}$:

$$x_t = \sum_{j=1}^m \left\{ a_{\lambda_j} \cos \left[\frac{2\pi(j-1)t}{\tau} \right] + b_{\lambda_j} \sin \left[\frac{2\pi(j-1)t}{\tau} \right] \right\} \quad (1.76)$$

where $\frac{2\pi}{\tau}$ is the **fundamental frequency** and:

$$m = \begin{cases} \frac{\tau-1}{2} + 1 & \text{if } \tau \text{ is odd; or} \\ \frac{\tau}{2} + 1 & \text{if } \tau \text{ is even.} \end{cases} \quad (1.77)$$

Theorem 1.5 (Covariance Stationarity). *The general process is **covariance stationary** if:*

1. *each component is **covariance stationary and therefore**:*

$$\mathbb{E}(a_{\lambda_j}) = \mathbb{E}(b_{\lambda_j}) = 0; \quad (1.78a)$$

$$\mathbb{V}(a_{\lambda_j}) = \mathbb{V}(b_{\lambda_j}) = \sigma_j^2; \text{ and} \quad (1.78b)$$

$$\mathbb{E}(a_{\lambda_j} b_{\lambda_j}) = 0; \text{ and} \quad (1.78c)$$

2. *the components are **orthogonal to each other**:*

$$\mathbb{E}(a_{\lambda_j} a_{\lambda_k}) = 0; \quad (1.79a)$$

$$\mathbb{E}(b_{\lambda_j} b_{\lambda_k}) = 0; \text{ and} \quad (1.79b)$$

$$\mathbb{E}(a_{\lambda_j} b_{\lambda_k}) = 0 \quad (1.79c)$$

Then, it is easy to show that $E(x_t) = 0$, $V(x_t) = \gamma(0) = \sum_{j=1}^m \sigma_j^2$, and $\gamma(s) = \sum_{j=1}^m \sigma_j^2 \cos \lambda_j s$ because each of the underlying harmonic processes is orthogonal from the rest.

To measure the importance of each cyclic component, that is, of each frequency λ_j , we can use the corresponding variance σ_j^2 . If low frequencies have high variances, we will have a smooth series in which long cycles tend to dominate. If on the other hand, high frequencies have high variances, then we will tend to obtain a series that fluctuates quickly. This relationship can be represented in a variance-frequency diagram, in an analogous way to how the probability function of a discrete random variable that takes the values $\lambda_1, \lambda_2, \dots, \lambda_m$ with probabilities $\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2$. Such a **graph will give us the so-called spectral “probability” function**. If we standardised the process scaling by $V^{1/2}(x_t)$, then the “probabilities” $\sigma_j^2/V^{1/2}(x_t)$ would add up to 1.

We can also consider the cumulative graph, which indicates the importance of frequencies less than or equal to a given one.

Definition 1.26 (Spectral Distribution Function).

$$F(\lambda) \equiv \sum_{j=1}^m \sigma_j^2 \mathbb{I}(\lambda_j \leq \lambda) \quad (1.80)$$

indicates the **contribution to the variance of the series of the frequencies less than or equal to λ** . The **function $F(\lambda)$ is known as the spectral distribution function**.

Once again, if we standardised x_t by scaling with $V^{1/2}(x_t)$, we would obtain a function such that:

$$F^*(\pi) = \frac{F(\pi)}{\gamma(0)} = 1 \quad (1.81)$$

If we allow m to go to infinite, we can obtain increasingly complex linearly deterministic processes. The covariance stationarity conditions are entirely analogous, with $V(x_t) = \gamma(0) = \sum_{j=1}^{\infty} \sigma_j^2 < \infty$. This way, though, we would never cover all possible frequencies between 0 and π , which prevents us from generating linearly regular processes. However, if we replace **sums with integrals and carefully define the a^T s and b^T s associated with each frequency, then we can cover a continuum of frequencies**. Thus, we will have:

$$x_t = \int_0^\pi [\cos \lambda t dA(\lambda) + \sin \lambda t dB(\lambda)] \quad (1.82)$$

where $A(\lambda)$ and $B(\lambda)$ are in turn stochastic processes with continuous parameter λ , with $\lambda \in [0, \pi]$. These are **stochastic integrals**, and they can be understood as **mean-square limits of the underlying Riemman-Stieltjes sums**.

Stochastic calculus is devoted to the analysis of such integrals. To **achieve stationarity, the “increments” for $A(\lambda)$ and $B(\lambda)$ must satisfy**:

$$\mathbb{E}[dA(\lambda)] = \mathbb{E}[dB(\lambda)] = 0 \quad \forall \lambda; \quad (1.83a)$$

$$\mathbb{E}[dA(\lambda_1) dA(\lambda_2)] = 0 \quad \forall \lambda_1 \neq \lambda_2; \quad (1.83b)$$

$$\mathbb{E}[dB(\lambda_1) dB(\lambda_2)] = 0 \quad \forall \lambda_1 \neq \lambda_2; \quad (1.83c)$$

$$\mathbb{E}[dA(\lambda_1) dB(\lambda_2)] = 0 \quad \forall \lambda_1, \lambda_2; \text{ and} \quad (1.83d)$$

$$\mathbb{E}[dA(\lambda)]^2 = \mathbb{E}[dB(\lambda)]^2 = dF(\lambda) \quad \forall \lambda. \quad (1.83e)$$

They **resemble Wiener processes, but unlike these, $dF(\lambda)$ is not necessarily $\sigma^2 d\lambda$.**

Given that $e^{i\lambda r} = \cos \lambda r + i \sin \lambda r$, x_t is sometimes written as:

$$x_t = \int_{-\pi}^{\pi} e^{i\lambda t} dZ(\lambda) \text{ with} \quad (1.84a)$$

$$\mathbb{E}[dZ(\lambda)] = 0 \forall \lambda; \quad (1.84b)$$

$$\mathbb{E}[dZ(\lambda_1) dZ(\lambda_2)] = 0 \quad \forall \lambda_1 \neq \lambda_2; \text{ and} \quad (1.84c)$$

$$\mathbb{E}[dZ(\lambda)]^2 = dF(\lambda). \quad (1.84d)$$

In that case:

$$\mathbb{E}(x_t) = 0; \quad (1.85a)$$

$$\mathbb{V}(x_t) = \gamma(0) = \int_0^{\pi} dF(\lambda); \text{ and} \quad (1.85b)$$

$$\text{cov}(x_t, x_{t-s}) = \gamma(s) = \int_0^{\pi} \cos \lambda s dF(\lambda) \quad (1.85c)$$

so $F(\lambda)$ yields the spectral distribution function, which tells us how each frequency, or in other words, each subprocess, contributes to the variance of the overall process. The usefulness of this **so-called Cramér representation lies in the fact that any stationary process in covariance can be written like this.**

Although any function can be written as

$$F(\lambda) = F_c(\lambda) + F_d(\lambda) + F_s(\lambda) \quad (1.86)$$

with $F_c(\lambda)$ continuous, $F_d(\lambda)$ discrete and a somewhat bizarre reminder term $F_s(\lambda)$, the continuously singular function $F_s(\lambda)$ **can be ignored in most practical applications because it is continuous but not absolutely so, which means that it cannot be written over a compact set as its initial value plus the integral of its first derivative**, and it is both non-constant and non-decreasing even though its derivative is zero except at a set of points of measure 0.

Example 1.10 (Devil’s Staircase). An example is the so-called devil’s staircase often used as a counterexample to show that there exist probability distributions that have neither a probability function because the probability of each point is 0, nor a density function because their distribution function is not absolutely continuous.

Consequently, **most processes can be split into two parts**: one with spectral density $F_c(\lambda)$ and another one with spectral density $F_d(\lambda)$. Specifically:

$$x_t = \int_0^{\pi} \cos \lambda a(\lambda) d\lambda + \sin \lambda b(\lambda) d\lambda + \sum_{j=1}^{\infty} a_{\lambda_j} \cos \lambda_j t + b_{\lambda_j} \sin \lambda_j t. \quad (1.87)$$

The **linearly regular part corresponds to the absolutely continuous part**. The **linearly deterministic part corresponds to the discrete part**. This is nothing other than the Wold representation theorem but in the frequency domain.

Assuming that the process is linearly regular, we can define the spectral density function as:

$$f(\lambda) = \frac{1}{2} \frac{dF(\lambda)}{d\lambda}. \quad (1.88)$$

The **reason for the 1/2 is that many books define λ between $-\pi$ and π** . This function is also called “spectrum,” and allows us to see the importance of frequency λ because

$$\mathbb{V}(x_t) = \gamma(0) = 2 \int_0^\pi f(\lambda) d\lambda; \text{ and} \quad (1.89a)$$

$$\text{cov}(x_t, x_{t-s}) = \gamma(s) = 2 \int_0^\pi \cos \lambda s f(\lambda) d\lambda. \quad (1.89b)$$

Interestingly, the **autocovariance can be regarded as the inverse Fourier transforms of the spectral density**. Consequently, we can obtain the spectral density function from the autocovariance **generation function as the direct Fourier transform**:

$$f(\lambda) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} \gamma(s) e^{-i\lambda r} = \frac{1}{2\pi} \psi_{xx} \left(e^{-i\lambda} \right). \quad (1.90)$$

Therefore, **spectral analysis does not provide new information, but rather reorganises it and presents it differently**. It is therefore an alternative to the autocorrelation function, which may be more illustrative in certain cases. For example, if a process has an irregular seasonal component, its spectral density function would have a local maximum around the frequency corresponding to the period of seasonality. Likewise, if the process is relatively “persistent”, in the sense that long-term cycles dominate, we would expect to see high values of $f(0)$. As we **shall see later in this topic, whether or not a process has long memory is determined by the behaviour of $f(\lambda)$ as $\lambda \rightarrow 0$** .

1.6 ARMA Models

Any **covariance stationary process can be written as**:

$$x_t = \sum_{j=0}^{\infty} \delta_j u_{t-j} + P(x_t | H_{-\infty}), u_t = x_t - P(x_t | H_{t-1}), \quad (1.91)$$

where

$$\overbrace{\sum_{j=0}^{\infty} \delta_j^2}^{\text{Square Summable}} < \infty; \quad (1.92a)$$

$$\mathbb{E}(u_t) = 0; \quad (1.92b)$$

$$\mathbb{E}(u_t^2) = \sigma_u^2; \text{ and} \quad (1.92c)$$

$$\mathbb{E}(u_t u_s) = 0, t \neq s \quad (1.92d)$$

Originally, ARMA models estimate the coefficients of u_s , but we ignore it for the time being. Still, we have an infinite number of coefficients for infinite data.

Example 1.11 (White Noise). Consider **white noise**, WN :

$$x_t = \varepsilon_t = u_t. \quad (1.93)$$

u_t is the one-period-ahead prediction error. Hence:

$$\delta_j = 0 \quad \forall j \geq 1 \quad (1.94a)$$

$$D(z) = 1 \quad (1.94b)$$

$$\psi_{xx}(z) = \sigma_\varepsilon^2 \quad (1.94c)$$

Its spectral density function is very simple since $f(\lambda) = \sigma_\varepsilon^2/2\pi \forall \lambda$. **This reflects the fact that no frequency dominates.** By analogy to the white light spectrum, such a flat spectrum is what gives white noise its name.

1.6.1 MA(1) Process

Consider the **moving average process**:

$$x_t = \varepsilon_t + \beta \varepsilon_{t-1} \text{ with } \varepsilon_t \sim WN. \quad (1.95)$$

We have that:

$$\mathbb{E}(x_t) = 0; \quad (1.96a)$$

$$\gamma_0 = \sigma_\varepsilon^2 (1 + \beta^2); \quad (1.96b)$$

$$\gamma_1 = \beta \sigma_\varepsilon^2; \quad (1.96c)$$

$$\gamma_j = 0 \quad \forall j \geq 2; \quad (1.96d)$$

$$D(z) = 1 + \beta z; \quad (1.96e)$$

$$\psi_{xx}(z) = \sigma_\varepsilon^2 (1 + \beta z) (1 + \beta z^{-1}) = \sigma_\varepsilon^2 [1 + \beta^2 + \beta (z + z^{-1})]; \quad (1.96f)$$

$$\rho_1 = \frac{\beta}{1 + \beta^2}; \text{ and } \quad (1.96g)$$

$$\rho_j = 0 \quad \forall j \geq 2. \quad (1.96h)$$

The **spectral density function** is:

$$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} (1 + \beta^2 + 2\beta \cos \lambda) \quad (1.97)$$

Theorem 1.6 (Autocorrelation of MA(1) Process). *The sign of β determines the sign of the first-order autocorrelation, whose magnitude cannot exceed .5 because the autocorrelation function would no longer give rise to a positive semidefinite matrix when written in (tridiagonal) matrix form.*

If β is positive, lower frequencies dominate. If it's negative, higher frequencies dominate.

When β is negative, the process will fluctuate around its mean more than the underlying WN. The **maximum of its spectral density will happen at $\lambda = \pi$** . In contrast, when β is positive, the MA(1) will fluctuate less than ε_t . This time, the maximum of $f(\lambda)$ will happen at $\lambda = 0$. The magnitude of β determines the variance of the process.

If we plot ρ_1 as a function of β we will realise that there may be a second MA(1) process with the same autocovariance structure. Specifically if we choose the dynamic coefficient to be β^{-1} and the variance of the WN to be $\sigma_\varepsilon^2 \beta^2$ we get exactly the same γ_0 and γ_1 . Those two processes differ in a crucial aspect.

Definition 1.27 (Invertibility). **$\{x_t\}$ is invertible iff $\varepsilon_t \in H_t$.** If $\varepsilon_t \in \text{int}(H_t)$ then $\{x_t\}$ is **strictly invertible**.

Proposition 1.6.1. *When the process is invertible, ε_t will trivially coincide with the one-period ahead prediction errors underlying the Wold decomposition.^{1.6} When the process is not invertible, though, we can only recover ε_t from the present and future of x_t .*

^{1.6} The process is invertible if I can recover the underlying the process from the current and past observations.

Parametrically, the condition is $|\beta| \leq 1$. When $|\beta| < 1$ we can make use of the lag (or backward shift) operator L .

Definition 1.28 (Lag Operator).

$$Lx_t = x_{t-1} \quad (1.98)$$

Although it is an operator that transforms a random variable viewed as a function of the sample space into another random variable, **we can often treat it as an algebraic symbol**. For example, $L^2x_t = L(Lx_t) = Lx_{t-1} = x_{t-2}$. In addition, it is a linear operator. Its inverse operator is the "lead" (or forward shift) operator L^{-1} , which is such that:

$$L^{-1}x_t = x_{t+1}. \quad (1.99)$$

Example 1.12 (Extracting ϵ from MA(1)). Since $x_t = \epsilon_t + \beta\epsilon_{t-1} = (1 + \beta L)\epsilon_t$, **we can try to recover ϵ_t by finding the inverse operator to $(1 + \beta L)$** .

This is an infinite-order operator. To find it, we write it as:

$$\pi_0 + \pi_1 L + \pi_2 L^2 + \dots \quad (1.100)$$

multiply it by $1 + \beta L$ and equate the result to 1 (because it's the reciprocal - so multiplying by $1 + \beta L$ gives 1):

$$(\pi_0 + \pi_1 L + \pi_2 L^2 + \dots)(1 + \beta L) = 1 \quad (1.101)$$

Hence:

$$\underbrace{\pi_0}_{=1} + \underbrace{(\pi_1 + \beta\pi_0)}_{=0} L + \underbrace{(\pi_2 + \beta\pi_1)}_{=0} L^2 + \dots = 1. \quad (1.102)$$

We can re-arrange it to get:

$$\pi_j = (-\beta)^j, \quad (1.103)$$

which converges iff $|\beta| < 1$. Then:

$$\pi(L)x_t = \epsilon_t. \quad (1.104)$$

When $|\beta| > 1$, then the inverse operator is:

$$\beta^{-1}L^{-1} - \beta^{-2}L^{-2} + \beta^{-3}L^{-3} - \beta^{-4}L^{-4} + \dots \quad (1.105)$$

Hence, the **Wold representation** is:

$$x_t = \begin{cases} (1 + \beta L)\epsilon_t & \text{if } |\beta| < 1 \\ (1 + \beta^{-1}L)u_t & \text{if } |\beta| > 1. \end{cases} \quad (1.106)$$

1.6.2 MA(2) Process

Consider:

$$x_t = \epsilon_t + \beta_1\epsilon_{t-1} + \beta_2\epsilon_{t-2} \text{ with } \epsilon_t \sim WN. \quad (1.107)$$

It can be shown that:

$$\mathbb{E}(x_t) = 0; \quad (1.108a)$$

$$\gamma_0 = \sigma_\epsilon^2 (1 + \beta_1^2 + \beta_2^2); \quad (1.108b)$$

$$\gamma_1 = \beta_1 (1 + \beta_2) \sigma_\epsilon^2; \quad (1.108c)$$

$$\gamma_2 = \beta_2 \sigma_\epsilon^2; \quad (1.108d)$$

$$\gamma_j = 0 \quad \forall j \geq 3; \text{ and} \quad (1.108e)$$

$$D(z) = 1 + \beta_1 z + \beta_2 z^2. \quad (1.108f)$$

Moreover, we have:

$$\psi_{xx}(z) = \sigma_\varepsilon^2 (1 + \beta_1 z + \beta_2 z^2) (1 + \beta_1 z^{-1} + \beta_2 z^{-2}) \implies \quad (1.109a)$$

$$\psi_{xx}(z) = [1 + \beta_1^2 + \beta_2^2 + \beta_1(1 + \beta_2)(z + z^{-1}) + \beta_2(z^2 + z^{-2})]. \quad (1.109b)$$

We have:

$$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} [1 + \beta_1^2 + \beta_2^2 + 2\beta_1(1 + \beta_2)\cos\lambda + 2\beta_2\cos 2\lambda]. \quad (1.110)$$

We have **limited seasonality**.^{1.7}

$$x_t = u_t + \beta_2 u_{t-2}. \quad (1.111)$$

Theorem 1.7 (Invertibility of MA(2)). β must solve:

$$\lambda^2 + \beta_1 \lambda + \beta_2 = 0, \quad (1.112)$$

with **roots on or inside the unit circle**.^{1.8}

Example 1.13 (Extracting ϵ from MA(2) Process). If the roots, which could be real or complex conjugates, are strictly inside the unit circle, we will be able to invert the polynomial $1 + \beta_1 L + \beta_2 L^2$ by using the same procedure as in the MA(1) case. Specifically, we write it as $\pi_0 + \pi_1 L + \pi_2 L^2 + \dots$, multiply it by $1 + \beta_1 L + \beta_2 L^2$ and equate the result to 1. This yields the second order linear difference equation $\pi_k + \beta_1 \pi_{k-1} + \beta_2 \pi_{k-2} = 0$, with initial conditions $\pi_0 = 1$ and $\pi_1 = -\beta_1$. If some of the roots lie outside the unit circle, then we will have to rely on L^{-1} . In any case, there are up four different representations of the MA(2) process, only one of which will be invertible. This will coincide with the Wold representation.

Comment: Finding inverses and the associated conditions are just purely DE.

1.6.3 MA(k) Process

Consider:

$$x_t = \varepsilon_t + \sum_{j=1}^k \beta_j \varepsilon_{t-j} \text{ with } \varepsilon_t \sim WN. \quad (1.113)$$

Here, we have:

$$D(z) = 1 + \sum_{j=1}^k \beta_j z^j; \quad (1.114a)$$

$$\psi_{xx}(z) = \sigma_\varepsilon^2 \left(1 + \sum_{j=1}^k \beta_j z^j\right) \left(1 + \sum_{j=1}^k \beta_j z^{-j}\right); \quad (1.114b)$$

$$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \left(1 + \sum_{j=1}^k \beta_j e^{-ij\lambda}\right) \left(1 + \sum_{j=1}^k \beta_j e^{ij\lambda}\right); \quad (1.114c)$$

$$\gamma_s = \sigma_\varepsilon^2 \sum_{j=1}^{k-s} \beta_j \beta_{j+s} \quad s \in \{0, 1, \dots, k\}; \quad (1.114d)$$

$$\gamma_j = 0 \quad \forall j > k \quad (1.114e)$$

Theorem 1.8 (Invertibility of MA(k) Processes).

$$\lambda^k + \beta_1 \lambda + \dots + \beta_k = 0 \quad (1.115)$$

with **roots or on inside the unit circle**.

^{1.7} It's because we will have something akin to pseudo-cycles.

^{1.8} This is just the **characteristic equation** of the associated DE.

There are up to 2^k different representations, but **only the Wold representation will be invertible**. Although we gain generality by adding more and more coefficients, the autocorrelations will be 0 after the k^{th} one. For that reason, we must look for an **alternative route to have infinite lags without having an infinite number of coefficients**.

1.6.4 AR(1) Process

Consider:

$$x_t = \sum_{j=0}^{\infty} \alpha^j u_{t-j}. \quad (1.116)$$

The **Wold decomposition requires** $|\alpha| < 1$. Assume $|\alpha| < 1$. Hence:

$$\mathbb{E}(x_t) = 0; \quad (1.117a)$$

$$\gamma_0 = \sigma_u^2 \sum_{j=0}^{\infty} \alpha^{2j} = \sigma_u^2 (1 - \alpha^2)^{-1}; \quad (1.117b)$$

Because $\mathbb{E}(u_s u_t) = 0$

$$\gamma_j = \alpha^j \gamma_0; \quad (1.117c)$$

$$\rho_j = \alpha^j; \quad (1.117d)$$

$$D(z) = \sum_{j=0}^{\infty} \alpha^j z^j = (1 - \alpha z)^{-1}; \quad (1.117e)$$

$$\psi_{xx}(z) = \frac{\sigma_u^2}{(1 - \alpha z)(1 - \alpha z^{-1})} = \frac{\sigma_u^2}{1 + \alpha^2 - \alpha(z + z^{-1})}; \text{ and} \quad (1.117f)$$

$$f(\lambda) = \frac{\sigma_u^2}{2\pi} \left[\frac{1}{1 + \alpha^2 - 2\alpha \cos \lambda} \right] \quad (1.117g)$$

Again, the sign of α determines how often x_t fluctuates around its mean.

- 📖 If $1 > \alpha > 0$, the maximum of the spectral density will be at $\lambda = 0$, and the bigger α , the bigger the importance of the components associated with low frequencies.
- 📖 This simply reflects the fact that since an **AR(1) is an infinite moving average of white noise with positive coefficients for $\alpha > 0$, and therefore smoother than white noise**.
- 📖 As α approaches 1 the series becomes smoother and smoother because more and more lagged values of u_t will get a non-negligible weight.
- 📖 In contrast, when $0 > \alpha > -1$, **the series will fluctuate a lot, and the high frequencies associated with short cycles will dominate**.
- 📖 Nevertheless, the magnitude of α determines the variance of the process. The **closer $|\alpha|$ is to 1, the higher the variance**.

The AR(1) process can be written as a **first-order stochastic difference equation**:

$$x_t = \alpha x_{t-1} + u_t. \quad (1.118)$$

Therefore, it is always trivially invertible. However, in this case the **initial condition x_0 matters for stationarity: x_0 should be drawn from the steady state of the process**. If $|\alpha| < 1$, the initial condition becomes negligible as T increases.

Remark (Markov Chains). *Remember that Markov chains are given by the below:*^{1.9}

$$\mathbb{P}(x_t = 0 \mid x_{t-1} = 0, x_{t-2}, \dots) = \mathbb{P}(x_t \mid x_{t-1} = 0) = p_t; \text{ and} \quad (1.119a)$$

$$\mathbb{P}(x_t = 1 \mid x_{t-1} = 1, x_{t-2}, \dots) = \mathbb{P}(x_t \mid x_{t-1} = 1) = q_t \quad (1.119b)$$

Example 1.14 (Special Case: Binary Markov Chain).

$$\underbrace{\begin{array}{c|c|c} x_t(\downarrow) & x_{t-1}(\rightarrow) & \\ \hline & 0 & 1 \\ \hline 0 & p & 1-q \\ 1 & 1-p & q \end{array}}_{\equiv \text{Transition Matrix}} \quad (1.120)$$

We have:

$$\mathbb{E}(x_t) = \mathbb{P}(x_t = 1) = \frac{1-p}{2-p-q} \implies \quad (1.121a)$$

$$\underbrace{x_t - \mathbb{E}(x_t) = (p+q-1)[x_{t-1} - \mathbb{E}(x_{t-1})] + u_t}_{=\text{AR}(1) \text{ for Deviations from Mean}}; \text{ and} \quad (1.121b)$$

$$\mathbb{E}(u_t \mid x_{t-1} = 0) = \mathbb{E}(u_t \mid x_{t-1} = 1) = 0 \implies \quad (1.121c)$$

$$\mathbb{E}(u_t \mid x_{t-1}) = 0 \implies \quad (1.121d)$$

$$\underbrace{\mathbb{E}(u_t \mid x_{t-1}, x_{t-2}, \dots)}_{=\text{Martingale Difference}} = 0. \quad (1.121e)$$

If $p+q-1 = 0$ **the process has no memory!** It becomes an **i.i.d. sequence of Bernoulli trials**. The stability condition is:

$$|p+q-1| < 1. \quad (1.122)$$

Interestingly, when $p = q \neq 0$ and $p = q \neq 1$, the we “escape” the initial condition in a **finite number of moves**.

The AR(1) has a massive limitation regarding its autocorrelations. Either they decay to 0 or they exhibit a cyclical behaviour. This means that we might wish to have more parameters while modelling richer processes.

1.6.5 AR(2) Process

Consider:

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + u_t \quad \text{given the initial condition: } (x_0, x_{-1}). \quad (1.123)$$

If we assume **mean and covariance stationarity**:

$$\mathbb{E}(x_t) = 0; \quad (1.124a)$$

$$\gamma_0 = \mathbb{E}[x_t(\alpha_1 x_{t-1} + \alpha_2 x_{t-2} + u_t)] = \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + \sigma_u^2; \quad (1.124b)$$

$$\gamma_1 = \mathbb{E}[x_{t-1}(\alpha_1 x_{t-1} + \alpha_2 x_{t-2} + u_t)] = \alpha_1 \gamma_0 + \alpha_2 \gamma_1; \text{ and} \quad (1.124c)$$

$$\gamma_2 = \mathbb{E}[x_{t-2}(\alpha_1 x_{t-1} + \alpha_2 x_{t-2} + u_t)] = \alpha_1 \gamma_1 + \alpha_2 \gamma_0. \quad (1.124d)$$

Theorem 1.9 (Yule-Walker).

$$\gamma_j = \alpha_1 \gamma_{j-1} + \alpha_2 \gamma_{j-2} \quad (1.125)$$

Asymptotic covariance-stationary condition: $\lambda^2 - \alpha_1 \lambda - \alpha_2$ has roots inside the unit circle. If outside (on), then explosive (integrated).

^{1.9} For more, see the *Stochastic Modelling* course notes.

The **spectral density function** is:

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left[\frac{1}{1 + \alpha_1^2 + \alpha_2^2 - 2\alpha_1(1 - \alpha_2)\cos\lambda - 2\alpha_2\cos 2\lambda} \right]. \quad (1.126)$$

This function may look like that of an AR(1) process, but it may also have different forms. In particular, an **AR(2) process can generate irregular cyclic behaviour, in the sense that $f(\lambda)$ can have an interior maximum**. An example would be the “**seasonal**” process $x_t = \alpha_2 x_{t-2} + u_t$, with $0 < \alpha_2 < 1$. Note that this will work only with the processes characterised by two seasons per period. Also, the largest root dominates other roots.

Remark. *If we want to have stationarity, we need to draw (x_0, x_{-1}) from the **joint stationary distribution**!*

1.6.6 AR(k) Process

Consider:

$$(1 - \alpha_1 L - \alpha_2 L^2 - \dots - \alpha_k L^k) x_t = u_t : \quad (1.127a)$$

$$\text{The Initial Conditions : } (x_0, x_{-1}, \dots, x_{-k+1}) \quad (1.127b)$$

Theorem 1.10 (Yule Walker Cont’d).

$$\gamma_j = \alpha_1 \gamma_{j-1} + \alpha_2 \gamma_{j-2} + \dots + \alpha_k \gamma_{j-k} \quad j \geq k \quad (1.128)$$

✎ *Asymptotic covariance-stationary condition: $\lambda^k - \alpha_1 \lambda^{k-1} - \dots - \alpha_k$ **has roots inside the unit circle. If outside (on), then explosive (integrated).***

✎ ***All AR(k)s are invertible!***

✎ *Finite sample stationarity: $(x_0, x_{-1}, \dots, 0, x_{-k+1})^T$ **drawn from the joint distribution (right mean and covariance matrix).** If the condition is not met, we can look for asymptotic stationarity.*

Remark. *Remember about the following.*

📺 *AR processes are slightly less general than MA processes.*

📺 *MA processes with root on the unit circle cannot be written as AR processes with an infinite p .*

📺 *Nevertheless, the predictions of the MA process can be approximated arbitrarily well by a sequence of predictions for increasingly large AR(p) processes.*

📺 *A general problem with the ARs and MAs is as follows. **If we want flexibility, we need to add parameters (in general, too many).***

1.6.7 ARMA Process

Consider ARMA(1,1):

$$x_t - \alpha x_{t-1} = \varepsilon_t + \beta \varepsilon_{t-1}. \quad (1.129)$$

That is, you can think of that as an **AR(1) process in which the stochastic component is MA(1) instead of WN**.

Theorem 1.11 (ARMA(1,1) Stationarity). *The **stationarity condition (asymptotically)** is that **AR(1) has to be asymptotically stationary; i.e., $|\alpha| < 1$.***

Theorem 1.12 (ARMA(1,1) Invertibility). *The **invertibility condition** is that **MA(1) has to be invertible; i.e., $|\beta| \leq 1$.***

The Wold decomposition comprises of a rational distributed lag:

$$x_t = \begin{cases} \frac{1+\beta L}{1-\alpha L} \varepsilon_t & \text{if invertible,} \\ \frac{1+\beta^{-1}L}{1-\alpha L} u_t & \text{if not} \end{cases}, \quad (1.130)$$

where the relationship between u_t and ε_t :

$$u_t = \frac{1 + \beta L}{1 + \beta^{-1} L} \varepsilon_t. \quad (1.131)$$

Remark.

$$u_t \equiv x_t - \mathbb{P}(x_t \mid H_{t-1}) \quad (1.132)$$

We can find the **variance**:

$$\mathbb{V}(x_t) = \left[1 - (\alpha + \beta)^2 \sum_{j=0}^{\infty} \alpha^{2j} \right] \sigma_{\varepsilon}^2 = \left[1 - \frac{(\alpha + \beta)^2}{1 - \alpha^2} \right] \sigma_{\varepsilon}^2. \quad (1.133)$$

Autocovariances are similar to AR(1) with the first lag being more flexible. The special cases include: $\alpha = -\beta$, then coefficients cancel and x_t is WN. But also when $\alpha = -\beta^{-1}$. Even though it doesn't have many parameters, it can deliver very flexible dynamics. To obtain its autocorrelation structure and spectral density function it is convenient to make use of the autocovariance generating function.

In the covariance stationary case, the AGF will be:

$$\Gamma(z) = \sigma_u^2 \frac{B(z)B(z^{-1})}{A(z)A(z^{-1})} \quad (1.134)$$

so that:

$$f(\lambda) = \frac{\sigma_u^2}{2\pi} \frac{B(e^{-i\lambda})B(e^{i\lambda})}{A(e^{-i\lambda})A(e^{i\lambda})} = \frac{\sigma_u^2}{2\pi} \frac{|B(e^{-i\lambda})|^2}{|A(e^{-i\lambda})|^2} = \frac{\sigma_u^2}{2\pi} \frac{\left| 1 + \sum_{j=1}^k \beta_j e^{-i\lambda j} \right|^2}{\left| 1 - \sum_{j=1}^h \alpha_j e^{-i\lambda j} \right|^2}. \quad (1.135)$$

This is known as a **rational spectral function** since it is the ratio of two trigonometric polynomials. These types of **functions can approximate the spectral density function of many stationary processes quite well**. This is nothing more than a **reflection that ARMA processes can closely approximate the lag structure and associated autocovariance of numerous linearly regular processes**.

1.7 Partial Autocorrelation Function

Consider the **least squares projection of x_t onto a constant and $x_{t-1}, x_{t-2}, \dots, x_{t-k}$** . Let $\phi_{jk}, (j = 1, \dots, k)$ denote the coefficients from this projection.

Example 1.15 (Working Towards Partial Correlation). Assume **covariance stationarity**:

$$\mathbb{P}(x_t \mid \{x_{t-1}\}) = \frac{\gamma_1}{\gamma_0} x_{t-1} = \rho_1 x_{t-1}; \quad (1.136a)$$

$$\mathbb{P}(x_t \mid \{x_{t-1}, x_{t-2}\}) = (\gamma_1 \ \gamma_2) \times \begin{pmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{pmatrix}^{-1} \begin{pmatrix} x_{t-1} \\ x_{t-2} \end{pmatrix} \implies \quad (1.136b)$$

$$\mathbb{P}(x_t \mid \{x_{t-1}, x_{t-2}\}) = \underbrace{\phi_{12}x_{t-1} + \phi_{22}x_{t-2}}_{\text{Separating Contributions}}; \text{ and} \quad (1.136c)$$

$$\mathbb{P}(x_t \mid \{x_{t-1}, x_{t-2}, x_{t-3}\}) = \underbrace{\phi_{13}x_{t-1} + \phi_{23}x_{t-2} + \phi_{33}x_{t-3}}_{\text{By the Same Token}}. \quad (1.136d)$$

Definition 1.29 (Partial Correlation Function). The **partial autocorrelation coefficient of order k is defined as ϕ_{kk} , and the partial autocorrelation function contains the sequence of ϕ_{kk} as a function of k .**

Obviously, $\phi_{11} = \rho_1$. If the true model is an AR(h), then $\phi_{kk} = 0$ for $k > h$. If the true model is an MA(1) with coefficient β , then the partial autocorrelations are of the form $(-\beta)^k$. More generally, it provides the mirror image of the usual autocorrelation function for AR and MA models. A nice property of partial autocorrelations is that they can be between -1 and 1 .

Definition 1.30 (Band Diagonal Matrix). For example, you could have a tri-diagonal matrix. FINISH

1.8 Long Memory Process

ARMA(p, q) models are very flexible, but they cannot well approximate covariance stationary models whose **Wold representation coefficients are square summable but not absolutely summable**. More generally, they cannot capture the spectrum of covariance stationary processes that have:

$$\lim_{\lambda \rightarrow 0} f(\lambda) = \infty \text{ but} \quad (1.137a)$$

$$\underbrace{\int_0^\infty f(\lambda) d\lambda}_{=\text{Variance}} < \infty \quad (1.137b)$$

Those processes are **called “long memory” ones**. The simplest example is the pure fractionally integrated process.

Definition 1.31 (Pure Fractionally Integrated Process).

$$x_t = \left[1 + dL + \frac{1}{2}d(1+d)L^2 + \frac{1}{6}d(1+d)(2+d)L^3 + \dots \right] \varepsilon_t = (1-L)^d \varepsilon_t \quad (1.138)$$

If $d < \frac{1}{2}$ the process is covariance stationary while if $d > (-\frac{1}{2})$ it will be strictly invertible too. The autocorrelations of covariance stationary versions of these “long memory” processes decay at hyperbolic rates, while those of ARMA processes decay at exponential rates.

1.9 Temporal Aggregation

Typically, data are temporally aggregated; e.g., quarterly GDP. What’s the process for annual data?

Example 1.16. Suppose we have x_t , what is the process for x_τ ? It is important to distinguish stocks from flows:

⊕ **stock:** x_τ is x_t with t even (e.g., we observe SPX every two hours);

⊕ **flow:** x_τ is $x_t + x_{t-1}$ with t even

The **key question** is as follows. Suppose we have the process x_t , **how can we derive x_τ from x_t ?**

✖ If x_t is **WN**, then x_τ is **WN** (true for both, stocks and flows).

✖ If x_t is **MA(1)**, for **stocks**, the **serial correlation disappears**; i.e., it becomes WN.

✖ If x_t is **MA(1)** (i.e., $x_t = \varepsilon_t + \beta\varepsilon_{t-1}$) for **flows**, it is also a **MA(1)** with parameter γ that solves:

$$\frac{\gamma}{1 + \gamma^2} = \frac{\beta}{1 + (1 + \beta)^2 + \beta^2}. \quad (1.139)$$

AR(k)s are trickier as they are infinite MA models!

Example 1.17 (AR(1) for Stocks). Assume x_t is AR(1) **for stocks**:

$$x_t = \alpha x_{t-1} + u_t. \quad (1.140)$$

Then:

$$x_\tau = \alpha^2 x_{\tau-1} + u_\tau. \quad (1.141)$$

Hence:

$$\mathbb{V}(x_t) = \mathbb{V}(x_\tau); \text{ and} \quad (1.142a)$$

$$\mathbb{V}(u_\tau) = (1 + \alpha^2) \mathbb{V}(u_t). \quad (1.142b)$$

Notice that the reverse gives rise to two solutions.

Example 1.18 (AR(1) for Flows). Assume x_t is a **flow and is an AR(1) process**:

$$x_t = \alpha x_{t-1} + u_t \quad (1.143)$$

Then:

$$\underbrace{x_\tau = (1 - \alpha^2 L^2) x_t = (1 + \delta^2 L^2) u_t}_{=\text{ARMA}(1,1)}. \quad (1.144)$$

Example 1.19 (Markov Chains). Note that **Markov chains can only be stocks**. I want to find:

$$\mathbb{P}(x_t = 0 \mid x_{t-2} = 0, x_{t-4}, x_{t-6}, \dots); \text{ and} \quad (1.145a)$$

$$\mathbb{P}(x_t = 1 \mid x_{t-2} = 1, x_{t-4}, x_{t-6}, \dots). \quad (1.145b)$$

I need to consider the possible trajectories. Let $p \equiv \mathbb{P}(x_t = 0 \mid x_{t-1} = 0)$ and $q \equiv \mathbb{P}(x_t = 1 \mid x_{t-1} = 1)$. Hence, we have:

$$\mathbb{P}(x_t = 0 \mid x_{t-2} = 0, x_{t-4}, x_{t-6}, \dots) = p^2 + (1-p)(1-q); \text{ and} \quad (1.146a)$$

$$\mathbb{P}(x_t = 1 \mid x_{t-2} = 1, x_{t-4}, x_{t-6}, \dots) = q^2 + (1-q)(1-p). \quad (1.146b)$$

Note that we do not need to worry about other probabilities (as they add up to 1).

$$\begin{pmatrix} p & 1-q \\ 1-p & q \end{pmatrix}^2 = \underbrace{\begin{pmatrix} p^2 + (1-p)(1-q) & 1-q^2 - (1-q)(1-p) \\ 1-p^2 - (1-p)(1-q) & q^2 + (1-q)(1-p) \end{pmatrix}}_{\text{Columns Add Up to 1}} \quad (1.147)$$

Hence, this can be written as an AR(1) process:

$$x_t - \mathbb{E} x_t = \underbrace{[p^2 + (1-p)(1-q) + q^2 + (1-q)(1-p) - 1]}_{=(p+q-1)^2} (x_{t-2} - \mathbb{E} x_{t-2}) + u_t. \quad (1.148)$$

Consider the eigenvalues of the transition matrix:

$$\det \left| \begin{pmatrix} p & 1-q \\ 1-p & q \end{pmatrix} - \lambda \mathbb{I}_2 \right| = 0 \implies \quad (1.149a)$$

$$\mathbf{v}_\lambda = \begin{pmatrix} 1 - \frac{1-p}{2-p-q} \\ \frac{1-p}{2-p-q} \end{pmatrix} \implies \quad (1.149b)$$

$$\lambda = \begin{pmatrix} 1 \\ p-q-1 \end{pmatrix}. \quad (1.149c)$$

Here, I used the triangular decomposition:

$$T = U \Lambda U^{-1} \implies \quad (1.150a)$$

$$T^j = U \Lambda^j U^{-1}. \quad (1.150b)$$

1.9.1 Relationship between Discrete and Continuous Time Processes

Let $r \in [0, 1]$ be the parameter of a sequence of stochastic processes $\{Y_T(r)\}$. Consider a sequence indexed by T . Stochastic processes are indexed by r (argument of the random functions). For fixed r , it is a sequence of random variables. processes:

$$Y_T(r) = \frac{\sqrt{T}}{T} \sum_{s=1}^{[Tr]} u_s = \frac{\sqrt{T}}{T} X_{[Tr]} \quad (1.151)$$

where u_t is a white noise process and $[x]$ is the largest integer smaller or equal than x .

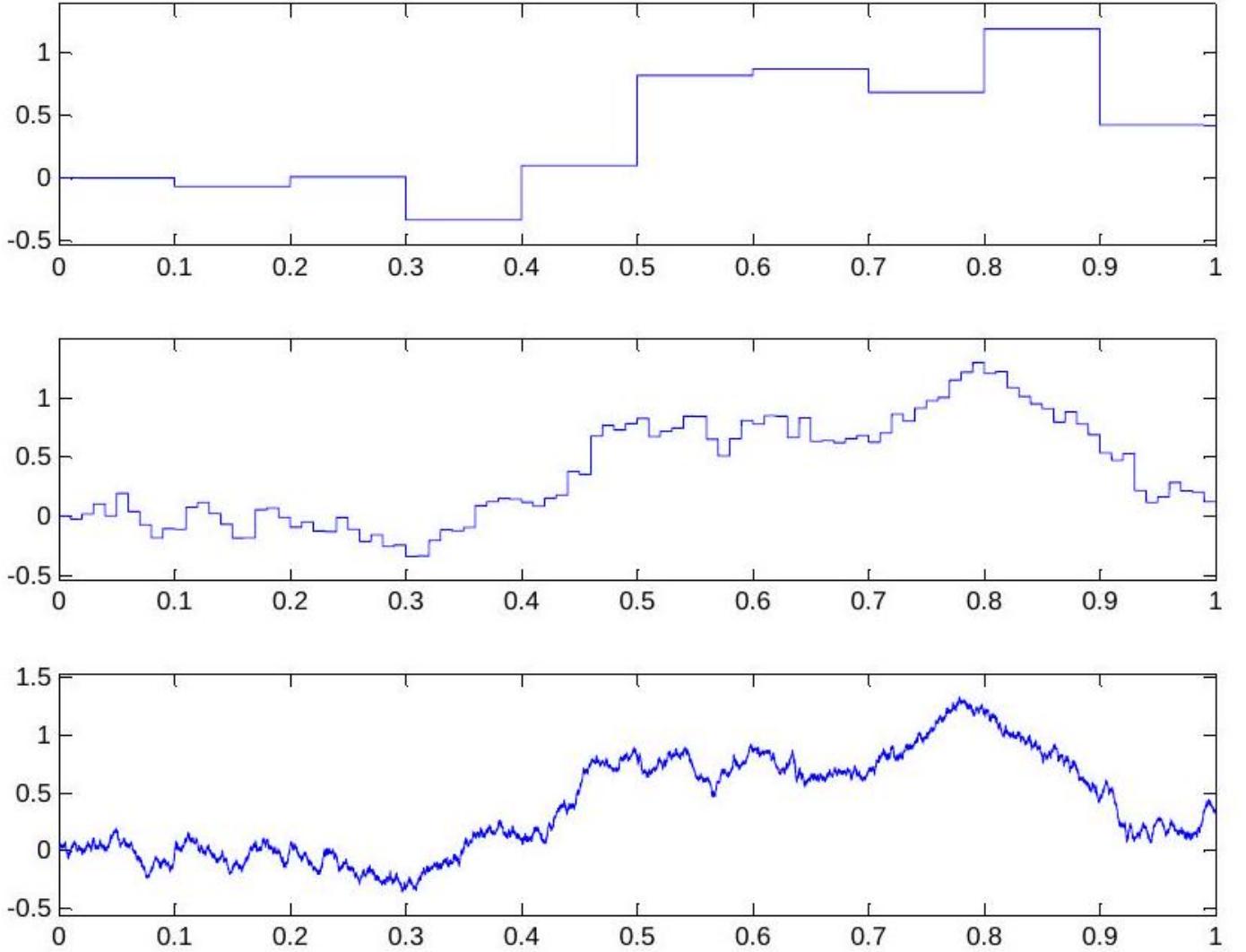


Figure 1.1: Simulating Random Walk/ Brownian Motion.

Let $r_1 < r_2 < r_3$, this means:

$$\begin{pmatrix} Y_T(r_1) \\ Y_T(r_2) \\ Y_T(r_3) \end{pmatrix} = \frac{\sqrt{T}}{T} \times \begin{pmatrix} \sum_{s=1}^{[Tr_1]} u_s \\ \sum_{s=1}^{[Tr_1]} u_s + \sum_{s=[Tr_1]+1}^{[Tr_2]} u_s \\ \sum_{s=1}^{[Tr_1]} u_s + \sum_{s=[Tr_1]+1}^{[Tr_2]} u_s + \sum_{s=[Tr_2]+1}^{[Tr_3]} u_s \end{pmatrix}. \quad (1.152)$$

This converges to:

$$\mathcal{N} \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \sigma_u^2 \begin{pmatrix} r_1 & r_1 & r_1 \\ r_1 & r_2 & r_2 \\ r_1 & r_2 & r_3 \end{pmatrix} \right]. \quad (1.153)$$

This way we obtain the Wiener process:

$$W(0) = 0; \quad (1.154a)$$

$$\mathbb{E}[W(t)] = 0; \quad (1.154b)$$

$$\mathbb{V}[W(t)] = \sigma^2 t; \text{ and} \quad (1.154c)$$

$$\text{cov}[W(t), W(s)] = \sigma^2 \min\{t, s\}. \quad (1.154d)$$

This is not a convergence of random variables, but a convergence of a stochastic process to its limiting stochastic process.

Theorem 1.13 (Functional Central Limit Theorem). *Apply the standard CLT for each element of the sequence. On top of that, apply a tightness condition. That is, apply CLT uniformly (i.e., the maximum difference has to go to zero) in r .*

Theorem 1.14 (Continuous Mapping Theorem). *This is the **stochastic process analogue to Slutsky's theorem**:*

$$Y_T(r) \xrightarrow{P} Y(r) \text{ for fixed } r \quad (1.155a)$$

$$g[Y_T(r)] \xrightarrow{P} g[Y(r)] \text{ for } g \text{ continuous} \quad (1.155b)$$

Example 1.20.

$$\mathbb{E} X_{RW} \xrightarrow{P} \mathbb{E} X_{\text{Wiener}}. \quad (1.156)$$

Is there an inverse relationship from the Wiener process to RW? Yes. Consider H intervals of length $h = \frac{1}{H}$, then:

$$W\left(\frac{t}{H}\right) - W\left(\frac{t-1}{H}\right) \quad (1.157)$$

is like a RW 's innovation. This is a **bijective mapping between RW and Wiener**.

Example 1.21 (Continuous Markov Chain). Is there a **continuous time process that can generate a discrete Markov chain**? Let the transition matrix from X_t to X_{t+1} be:

$$\mathbb{P}_1 = \begin{pmatrix} p & 1-q \\ 1-p & q \end{pmatrix}. \quad (1.158)$$

Then from X_t to X_{t+h} we have that:

$$\mathbb{P}_h = \mathbb{P}^h \quad (1.159)$$

Theorem 1.15 (Chapman-Kolmogorov Equation).

$$\mathbb{P}_{t+s} = \mathbb{P}_t \mathbb{P}_s \quad (1.160)$$

Under certain embeddability conditions, \exists **intensity matrix** Q such that:

$$e^{hQ} = \begin{pmatrix} p_h & 1-q_h \\ 1-p_h & q_h \end{pmatrix}. \quad (1.161)$$

They generate **only positive serial correlation**!

Remark (Exponential of a Matrix). *This is a Taylor expansion:*

$$e^A = \mathbb{I} + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \dots; \text{ and} \quad (1.162a)$$

$$e^A = Ue^\Lambda U^{-1} \quad \text{given} \quad A = U\Lambda U^{-1}. \quad (1.162b)$$

Example 1.22 (Ornstein-Uhlenbeck (OU) Process). Consider the **first-order stochastic differential equation with the differential of the Wiener process as the driving process**:

$$\underbrace{dX(t) = \rho X(t)dt + \gamma dW(t)}_{=\text{Mean-Reverting Process}}. \quad (1.163)$$

If $X(t)$ is a stock process, then:

$$X_t^{(h)} = e^{\rho h} X_{t-h}^{(h)} + v_t^{(h)} \quad \text{with} \quad (1.164a)$$

$$v_t^{(h)} \mid X_{t-h}^{(h)}, \dots \sim \mathcal{N}\left[0, \frac{(\gamma\sigma)^2}{2\rho} (e^{2\rho h} - 1)\right]. \quad (1.164b)$$

Only positive correlation, large (small) when h is small (large). **An AR(1) with negative autocorrelation cannot be generated by an OU (it violates the embeddability conditions)**. WN does not exist in continuous time, you'll need $\rho \rightarrow -\infty$.

2 Multivariate Time Series

2.1 Introduction

2.1.1 Vector Stochastic Processes

Definition 2.1 (Vector Stochastic Process). A **vector stochastic process** is a collection of random vectors defined on the same (Ω, \mathcal{F}, P) :

$$\underbrace{\begin{pmatrix} X_1(\omega, t) \\ \vdots \\ X_N(\omega, t) \end{pmatrix}}_{=N \times 1 \text{ vector}} = X_t, \quad (2.1)$$

Typically $N \ll T$, so we will treat the **cross-sectional dimension as a relatively small number and consider t as the only index**. A realisation of the process is the joint path for every element. Again, we will initially focus on linear relationships.

Definition 2.2 (Trend).

$$\mu_t = \mathbb{E}(X_t) \quad (2.2)$$

This is a $N \times 1$ vector that **collects the trend functions of each element of the process**.

Definition 2.3 (Autocovariance Matrix). The **autocovariance matrix** (or matrix of autocovariance functions) is defined as:

$$\text{cov}(X_t, X_s) \equiv \Gamma(t, s) = \begin{pmatrix} \gamma_{11}(t, s) & \cdots & \gamma_{1N}(t, s) \\ \vdots & \ddots & \vdots \\ \gamma_{N1}(t, s) & \cdots & \gamma_{NN}(t, s) \end{pmatrix} \quad (2.3)$$

➔ The **autocovariance functions** are the **diagonal** elements.

➔ The **cross-covariance functions** are the **off-diagonal elements**.

➔ We can define auto- and cross-correlation functions analogously by dividing $\gamma_{ij}(t, s)$ by the square roots of $\gamma_{ii}(t, t)$ and $\gamma_{jj}(s, s)$.

Is it symmetric? No!

$$\Gamma(t, s) = \Gamma^T(s, t) \quad (2.4)$$

Only $\Gamma(t, t)$ is always symmetric.

What about **positive semidefiniteness**? Stack the X_t s in a $T \times N$ matrix, X whose rows contain X_t and whose columns contain the T observations on a single variable, then $\text{cov}(\text{vec}(X^T))$ is a symmetric positive semidefinite matrix with block element $\Gamma(i, j)$.

Definition 2.4 (Vectorisation). Let A be an $m \times n$ matrix, then:

$$\text{vec}(A) = \text{vec}(\underbrace{a_1 \mid a_2 \mid a_3 \cdots \mid a_n}_{=\text{Columns}}) \equiv \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}. \quad (2.5)$$

Definition 2.5 (Half Vectorisation). Let A be an $m \times n$ **symmetric matrix**, then:

$$\text{vech}(A) = \text{vech}(\underbrace{a_1 \mid a_2 \mid a_3 \cdots \mid a_n}_{=\text{Columns}}) \equiv \begin{pmatrix} a_1 \\ [a_2]_{2:m} \\ \vdots \\ [a_n]_m \end{pmatrix}. \quad (2.6)$$

The cross-correlation functions have the advantage that they remain between -1 and 1 but the **disadvantage is that they lose information about the scale of the variables involved**. When $i = j$ we recover the autocorrelation function for variable i , whose maximum value trivially happens when $t = s$. But when $i \neq j$ then the maximum autocorrelation could happen with $t \neq s$ (e.g. if $X_{1t} = X_{2t-1}$ then $\rho_{12}(t, t-1) = 1$).

Example 2.1 (White Noise).

$$\mathbb{E}(u_t) = 0; \quad (2.7a)$$

$$\Gamma(t, t) = \mathbf{\Omega} \quad \text{where } \text{tr}(\mathbf{\Omega}) < \infty; \text{ and} \quad (2.7b)$$

$$\Gamma(t, s) = 0 \quad \forall s \neq t. \quad (2.7c)$$

Therefore, **its elements are serially uncorrelated but they will generally be contemporaneously correlated**.

Example 2.2 (Harmonic Process).

$$X_t = \mathbf{a} \cos \lambda t + \mathbf{b} \sin \lambda t, \quad \lambda \in \mathbb{R} \quad (2.8)$$

where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N$ are randomly drawn at the beginning of time.

Example 2.3 (Winer Process). The **Wiener process** is as its univariate counterpart (zero mean, uncorrelated increments), but now has:

$$\mathbb{V}[\mathbf{W}(t) - \mathbf{W}(s)] = (t - s)\mathbf{\Omega} \quad (2.9)$$

It is common to set $\mathbf{\Omega} = \mathbf{I}_N$ to yield a standardised multivariate Wiener process/aka multivariate Brownian motion

Example 2.4 (Markov Chains). Replace the vector of **binary Markov chains with a single multinomial chain**. We see a **proliferation of states**. If we have a vector of two elements, we end up with a matrix with 4 states.

Example 2.5 (Factor Process). We have:

$$\mathbf{x}_t = \mathbf{C}\mathbf{f} + \mathbf{u}_t, \quad (2.10)$$

where \mathbf{C} is $N \times k$ and \mathbf{f} is $k \times 1$.

2.1.2 Stationarity

We only observe one realisation. How can we make inferences? We need additional assumptions. To do so, as in the univariate case, we exploit the concept of stochastic equilibrium.

Definition 2.6 (Mean-Stationarity). It's the same as in the univariate case:

$$\mu_t = \mu_s \quad \forall s \neq t. \quad (2.11)$$

Definition 2.7 (Covariance Stationarity). For the vector stochastic process to be **covariance stationary**, it needs to be mean stationary and:

$$\Gamma(t, s) = \Gamma(t - s). \quad (2.12)$$

Notice that it is not enough that each element of the vector is covariance stationary because this only affects the diagonal elements of $\Gamma(t, s)$.

2.2 Prediction

Definition 2.8 (Projection).

$$\mathbb{P}(y \mid H), \quad (2.13)$$

where H is the set of admissible prediction rules and y is a vector.

The object we want to predict is now multivariate, but this is not important because we are going to predict each element of y separately. The fundamental difference is that H may now contain rules involving several variables.

🍏 The **properties of projections hold: 1) Linearity, 2) Updating, and 3) Law of Iterated Projections**.

Learn about X_t using info from X_{t-1}, X_{t-2}, \dots (all series can be used).

Linear Case Define:

$$H_{t-1, \tau} \equiv \left\{ h = \sum_{j=0}^{\tau} \alpha_j^T X_{t-1-j}, \alpha_j \in \mathbb{R}^N \right\}, \quad (2.14)$$

the **space of linear functions that consider** $X_{t-1}, \dots, X_{t-\tau-1}$ and

$$H_{t-1} \equiv \overline{\bigcup_{\tau=0}^{\infty} H_{t-1, \tau}}. \quad (2.15)$$

Once again:

$$H_t = H_{t-1} \oplus U_t, \quad (2.16)$$

where U_t is the set of rules in H_t which are orthogonal to H_{t-1} . That is:

$$U_t = \{u = h - \mathbb{P}(h \mid H_{t-1}), h \in H_t\}. \quad (2.17)$$

It turns out that U_t is simply given by:

$$U_t = \{u = \alpha^T u_t, \alpha \in \mathbb{R}^N, u_t = X_t - \mathbb{P}(X_t \mid H_{t-1})\}. \quad (2.18)$$

That is, it **contains linear combinations of the one-period ahead linear prediction error of all N series**.

Regression Problem Consider:

$$\mathbb{P}(X_t | U_{t-1}) = \underbrace{\mathbb{E}(X_t u_{t-1}^T) \mathbb{E}^{-1}(u_{t-1} u_{t-1}^T)}_{\equiv D_{t,t-1}} u_{t-1} \equiv D_{t,t-1} u_{t-1} \quad (2.19)$$

where $D_{t,t-1}$ is a $N \times N$ matrix. Hence:

$$X_t = \sum_{j=0}^{\infty} D_{t,t-j} u_{t-j} + \mathbb{P}(X_t | H_{-\infty}). \quad (2.20)$$

where

$$H_{-\infty} = \bigcap_{k=0}^{\infty} H_{t-k} \quad (2.21)$$

is the prediction at the beginning of times for the N series, and $D_{t,t} = \mathbb{I}_N$.

Lemma 2.1. *If in addition, we assume that the **vector process is covariance stationary**, then*

$$D_{t,t-j} = D_j \quad (2.22)$$

for all t .

As a result, if X_t is **covariance stationary we have a multivariate version of the Wold representation theorem**:

$$X_t = \sum_{j=0}^{\infty} D_j u_{t-j} + \mathbb{P}(X_t | H_{-\infty}) \quad (2.23)$$

where

$$u_t = X_t - \mathbb{P}(X_t | H_{t-1}); \text{ and} \quad (2.24a)$$

$$\mathbb{V}(u_t) = \Omega \quad (2.24b)$$

Thus, we decompose X_t in two orthogonal components, a linearly deterministic part, $\mathbb{P}(X_t | H_{-\infty})$ and a linearly regular part.

Lemma 2.2 (Stationarity Implications). *The implication of stationarity is that:*

$$\text{tr} \left(\sum_{j=0}^{\infty} D_j \Omega D_j^T \right) < \infty \quad (2.25)$$

In general, **the multivariate Wold decomposition will be different from the collection of N univariate Wold decompositions**.

The multivariate Wold decomposition is useful for:

1. forecasting;
2. forecast revision; and
3. computation of positive semidefinite autocovariance functions.

Definition 2.9 (Matrix of Autocovariance Generating Functions). The **matrix of autocovariance generating functions** is given by:

$$\Psi_{xx}(z) = \sum_{j=-\infty}^{\infty} \Gamma(j) z^j. \quad (2.26)$$

For the linearly regular part,

$$\Psi_{xx}(z) = D(z) \Omega D^T(z^{-1}) \quad (2.27)$$

where

$$D(z) = \sum_{j=0}^{\infty} D_j z^j = \mathbb{I}_N + D_1 z + D_2 z^2 + \dots \quad (2.28)$$

2.3 Multivariate Spectral Analysis

2.3.1 Multivariate Harmonic Process

Definition 2.10 (Multivariate Harmonic Process). A **multivariate harmonic process** is a stochastic process defined in either discrete time, with $\mathcal{T} = \{-T_i, -T_{i-1}, \dots, -1, 0, 1, \dots, T_f\}$, or continuous time, with $\mathcal{T} = \{t : -T_i < t < T_f\}$, and T_i, T_f possibly unbounded, such that:

$$x_{it} = a_i \cos \lambda t + b_i \sin \lambda t, \quad (2.29)$$

where a_i and b_i are the elements of the $N \times 1$ random vectors $\mathbf{a} = (a_1, \dots, a_N)^T$ and $\mathbf{b} = (b_1, \dots, b_N)^T$, and therefore, random variables themselves, and λ a real number.

As in the univariate case, one draws \mathbf{a} and \mathbf{b} at the dawn of time and then obtains \mathbf{x}_t as a function of t and λ .

The different elements of \mathbf{x}_t are **cyclical functions that share a constant period**, but they differ in their amplitude and phase, not only across realisations for a given series, but also across series for a given realisation. It is easy to check that the multivariate harmonic process will be covariance stationary if:

$$\mathbb{E}(\mathbf{a}) = \mathbb{E}(\mathbf{b}) = \mathbf{0}; \quad (2.30a)$$

$$\mathbb{V}(\mathbf{a}) = \mathbb{V}(\mathbf{b}) = \mathbf{C}; \text{ and} \quad (2.30b)$$

$$\mathbb{E}(\mathbf{a}\mathbf{b}^T) = \mathbf{Q}, \quad (2.30c)$$

where \mathbf{C} is a positive (semi)-definite symmetric matrix and \mathbf{Q} a skew-symmetric matrix such that:

$$\mathbf{Q} = -\mathbf{Q}^T, \quad (2.31)$$

or $q_{ij} = -q_{ji} \forall i, j$, which obviously requires that $q_{ii} = 0 \forall i$. In that case:

$$\mathbb{E}(\mathbf{x}_t) = \mathbf{0}; \quad (2.32a)$$

$$\mathbb{V}(\mathbf{x}_t) = \mathbb{E}[(\mathbf{a} \cos \lambda t + \mathbf{b} \sin \lambda t)(\mathbf{a}^T \cos \lambda t + \mathbf{b}^T \sin \lambda t)] \implies \quad (2.32b)$$

$$\mathbb{V}(\mathbf{x}_t) = \mathbf{C}(\cos^2 \lambda t + \sin^2 \lambda t) + (\mathbf{Q} + \mathbf{Q}^T) \sin \lambda t \cos \lambda t = \mathbf{C}; \text{ and} \quad (2.32c)$$

$$\mathbf{\Gamma}(t, s) = \mathbb{E}[(\mathbf{a} \cos \lambda t + \mathbf{b} \sin \lambda t)(\mathbf{a}^T \cos \lambda s + \mathbf{b}^T \sin \lambda s)] \implies \quad (2.32d)$$

$$\mathbf{\Gamma}(t, s) = \mathbf{C}(\cos \lambda t \cos \lambda s + \sin \lambda t \sin \lambda s) + \mathbf{Q} \cos \lambda t \sin \lambda s + \mathbf{Q}^T \sin \lambda t \cos \lambda s \implies \quad (2.32e)$$

$$\mathbf{\Gamma}(t, s) = \mathbf{C} \cos \lambda(t - s) + \mathbf{Q}(\cos \lambda t \sin \lambda s - \sin \lambda t \cos \lambda s) \implies \quad (2.32f)$$

$$\mathbf{\Gamma}(t, s) = \underbrace{\mathbf{C} \cos \lambda(t - s) - \mathbf{Q} \sin \lambda(t - s)}_{\text{Both } \mathbf{C} \text{ and } \mathbf{Q} \text{ Are Important!}}. \quad (2.32g)$$

As expected, this only depends on $t - s$, but notice that:

$$\mathbf{\Gamma}(s, t) = \mathbf{\Gamma}(s - t) = \mathbf{C} \cos \lambda(s - t) - \mathbf{Q} \sin \lambda(s - t) \implies \quad (2.33a)$$

$$\mathbf{\Gamma}(s, t) = \mathbf{C} \cos \lambda(t - s) + \mathbf{Q} \sin \lambda(t - s) \implies \quad (2.33b)$$

$$\mathbf{\Gamma}(s, t) = \mathbf{C} \cos \lambda(t - s) - \mathbf{Q}^T \sin \lambda(t - s) = \mathbf{\Gamma}^T(t - s) = \mathbf{\Gamma}^T(t, s). \quad (2.33c)$$

It is convenient to interpret both terms.

- ⚙ The **marginal spectral probability distribution** of each process is a step function whose values appear in the diagonal elements of \mathbf{C} .
- ⚙ As a result, the **marginal autocovariance** of each process depends on those diagonal elements only.
- ⚙ Similarly, the **contemporaneous covariances** between the elements of \mathbf{x}_t are given by the off-diagonal elements of \mathbf{C} , as the diagonal of \mathbf{Q} is 0.
- ⚙ In contrast, the **dynamic cross-covariances** between any two elements depend on the off-diagonal elements of both \mathbf{C} and \mathbf{Q} .

2.3.2 Spectral Density Matrix

As in the univariate case, we can generate more complex processes by superposition, namely combining several mutually orthogonal multivariate harmonic processes with different frequencies. The autocovariance matrices of such a process will simply be the sum of the autocovariance matrices of the components given their mutual orthogonality. Similarly, we **can reproduce the analogue to the Wold representation by combining a linearly regular process with a linearly deterministic part**:

$$\mathbf{x}_t = \int_0^\pi \cos \lambda \mathbf{a}(\lambda) d\lambda + \sin \lambda \mathbf{b}(\lambda) d\lambda + \sum_{j=1}^{\infty} \mathbf{a}_{\lambda_j} \cos \lambda_j t + \mathbf{b}_{\lambda_j} \sin \lambda_j t \quad (2.34)$$

More generally, the **Cramér representation of a vector process will be**:

$$\mathbf{x}_t = \int_0^\pi [\cos \lambda t d\mathbf{A}(\lambda) + \sin \lambda t d\mathbf{B}(\lambda)] \quad (2.35)$$

where $\mathbf{A}(\lambda)$ and $\mathbf{B}(\lambda)$ are in turn vector stochastic processes with continuous parameter λ , with $\lambda \in [0, \pi]$.

To **achieve stationarity, the process "increments" for $\mathbf{A}(\lambda)$ and $\mathbf{B}(\lambda)$ must satisfy**:

$$\mathbb{E}[d\mathbf{A}(\lambda)] = \mathbb{E}[d\mathbf{B}(\lambda)] = \mathbf{0} \quad \forall \lambda; \quad (2.36a)$$

$$\mathbb{E}[d\mathbf{A}(\lambda_1) d\mathbf{A}^T(\lambda_2)] = \mathbf{0} \quad \lambda_1 \neq \lambda_2; \quad (2.36b)$$

$$\mathbb{E}[d\mathbf{B}(\lambda_1) d\mathbf{B}^T(\lambda_2)] = \mathbf{0} \quad \lambda_1 \neq \lambda_2; \quad (2.36c)$$

$$\mathbb{E}[d\mathbf{A}(\lambda_1) d\mathbf{B}^T(\lambda_2)] = \mathbf{0} \quad \lambda_1 \neq \lambda_2'; \quad (2.36d)$$

$$\mathbb{E}[d\mathbf{A}(\lambda) d\mathbf{A}^T(\lambda)] = \mathbb{E}[d\mathbf{B}(\lambda) d\mathbf{B}^T(\lambda)] = d\mathbf{C}(\lambda) \quad \forall \lambda \text{ where} \quad (2.36e)$$

$$d\mathbf{C}(\lambda) \text{ is symmetric positive (semi)definite; and} \quad (2.36f)$$

$$\mathbb{E}[d\mathbf{A}(\lambda) d\mathbf{B}^T(\lambda)] = d\mathbf{Q}(\lambda) \quad \forall \lambda, \quad (2.36g)$$

with $d\mathbf{Q}(\lambda)$ skew symmetric.

Notice that this is **compatible with the univariate definition because the diagonal elements of $\mathbf{Q}(\lambda)$ are all 0**, so the diagonal elements of $\mathbf{C}(\lambda)$ contain the marginal spectral distribution functions. But how can we relate the different components of the vector \mathbf{x}_t in the frequency domain?

If we have a vector of the linearly regular stochastic process we know that we can define the (marginal) spectral density of its i^{th} component as:

$$f_{ii}(\lambda) = \frac{1}{2\pi} \left[\gamma_{ii}(0) + 2 \sum_{s=1}^{\infty} \gamma_{ii}(s) \cos \lambda s \right] = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} \gamma_{ii}(r) e^{-i\lambda r} = \frac{1}{2\pi} \Gamma_{ii}(e^{-i\lambda}). \quad (2.37)$$

By analogy, we can define the **cross-spectral density between processes i and j as**:

$$f_{ij}(\lambda) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} \gamma_{ij}(r) e^{-i\lambda r} = \frac{1}{2\pi} \left[\sum_{r=-\infty}^{\infty} \gamma_{ij}(r) (\cos \lambda r - i \sin \lambda r) \right]. \quad (2.38)$$

In general, this cross-spectral density has a **real part**:

$$\Re[f_{ij}(\lambda)] = \frac{1}{2\pi} \left\{ \gamma_{ij}(0) + \sum_{s=1}^{\infty} [\gamma_{ij}(s) + \gamma_{ji}(s)] \cos \lambda s \right\}, \quad (2.39)$$

and an imaginary part:

$$\Im[f_{ij}(\lambda)] = -\frac{i}{2\pi} \left[\sum_{r=-\infty}^{\infty} \gamma_{ij}(r) \sin \lambda r \right] = -\frac{i}{2\pi} \left\{ \sum_{s=1}^{\infty} [\gamma_{ij}(s) - \gamma_{ji}(s)] \sin \lambda s \right\} \quad (2.40)$$

where we have exploited that $\gamma_{ij}(s) = \gamma_{ji}(-s)$, $\sin \lambda = -\sin(-\lambda)$, $\cos(\lambda) = \cos(-\lambda)$ and consequently $\sin 0 = 0$ and $\cos \lambda = 1$.

The real part reflects $d\mathbf{C}(\lambda)$ while the imaginary part $d\mathbf{Q}(\lambda)$. In this respect, **note that $f_{ji}(\lambda)$ is the complex conjugate of $f_{ij}(\lambda)$ because they share the real part and have opposite imaginary ones.**

Hence, the spectral density matrix

$$\mathbf{f}(\lambda) = \mathbf{c}(\lambda) - i\mathbf{q}(\lambda) \quad (2.41)$$

is a positive (semi)definite **Hermitian matrix**, which is the complex counterpart to a symmetric real matrix because its complex conjugate transpose will satisfy:

$$\mathbf{f}^*(\lambda) = \mathbf{c}(\lambda) + i\mathbf{q}(\lambda). \quad (2.42)$$

A useful property of Hermitian matrices is that all their eigenvalues are real, and their eigenvectors are unitary, so their spectral representation in terms of eigenvalues and eigenvectors is simple.

Definition 2.11 (Hermitian Matrix). A **Hermitian matrix** is unitary matrix is such that:

$$\underbrace{\mathbf{U}^{-1} = \mathbf{U}^*}_{\text{Conjugate} = \text{Inverse}} \quad (2.43)$$

so that

$$\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbb{I}. \quad (2.44)$$

Definition 2.12 (Coherency). Real and complex parts are usually combined in the **coherency** between x_{it} and x_{jt} at frequency λ , which is defined as:

$$\mathcal{C}(\lambda) = \frac{|f_{ij}(\lambda)|^2}{f_{ii}(\lambda)f_{jj}(\lambda)}. \quad (2.45)$$

We can **interpret this coherency as the R^2 in the theoretical least squares projection of the cyclical component of x_{it} at frequency λ on the cyclical component of x_{jt} at the same frequency.**

Like any R^2 , coherency is bounded between 0 and 1. If $\mathcal{C}(\lambda)$ is close to 1, it means that the cyclical components of x_{it} and x_{jt} at frequency λ are highly correlated. As in the univariate case, we can obtain the autocovariance matrices from the **spectral density using the inverse Fourier transform**. Specifically:

$$\gamma_{ij}(s) = \int_{-\pi}^{\pi} e^{i\lambda s} f_{ij}(\lambda) d\lambda \quad (2.46a)$$

$$\mathbf{\Gamma}(s) = \int_{-\pi}^{\pi} e^{i\lambda s} \mathbf{f}(\lambda) d\lambda \quad (2.46b)$$

2.4 VARMA Models

2.4.1 VMA(1) Process

It is given by:

$$\overbrace{X_t = \varepsilon_t + B\varepsilon_{t-1}, \quad \varepsilon_t \sim WN(0, \Omega)}^{= \text{VMA}(1)} \implies \quad (2.47a)$$

$$\mathbb{E}(X_t) = 0; \quad (2.47b)$$

$$\mathbb{V}(X_t) = \mathbf{\Gamma}(0) = \Omega + B\Omega B^T; \quad (2.47c)$$

$$\text{cov}(X_t, X_{t-1}) = \mathbf{\Gamma}(1) = B\Omega = \mathbf{\Gamma}^T(-1) = \text{cov}^T(X_{t-1}, X_t); \text{ and } \quad (2.47d)$$

$$\text{cov}(X_t, X_{t-j}) = 0, \quad \forall j > 1. \quad (2.47e)$$

We also observe a **tri-band block-diagonal structure** for **covariance matrix** of $\text{vec}(X^T)$:

$$\mathbb{V} \begin{pmatrix} X_1 \\ \vdots \\ X_T \end{pmatrix} = \underbrace{\begin{pmatrix} \Gamma(0) & \Gamma^T(1) & 0 & \cdots & 0 \\ \Gamma(1) & \Gamma(0) & \Gamma^T(1) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}}_{=\text{An } (NT \times NT) \text{ Matrix.}} \quad (2.48)$$

The **autocovariance generating matrix** is:

$$(I + Bz)\Omega(I + B^T z^{-1}) = \Omega + B\Omega B^T + B\Omega z + \Omega B^T z^{-1} \quad (2.49)$$

The **spectral density matrix** is:

$$\left(I + B e^{-i\lambda}\right) \Omega \left(I + B^T e^{i\lambda}\right), \quad (2.50)$$

which involves sines and cosines unless $B = B^T$.

Lemma 2.3 (Invertibility of VMA(1)). X_t is invertible if $\varepsilon_t \in H_t$.

How can we check invertibility in practice? Write $X_t = (I + BL)\varepsilon_t$ and try to find an inverse operator:

$$\Pi(L) = \Pi_0 + \Pi_1 L + \Pi_2 L^2 + \dots \quad (2.51)$$

such that:

$$\Pi(L)(I + BL) = \mathbb{I}. \quad (2.52)$$

This gives rise to **the system of first-order linear difference equations**:

$$\Pi_j + \Pi_{j-1}B = 0 \quad (2.53)$$

with initial condition $\Pi_0 = \mathbb{I}_N$. The **tentative solution** $\Pi_j = (-B)^j$ will be admissible if and only if **the eigenvalues of the matrix B are inside the unit circle**. If B is **diagonalisable**:

$$B^j = P\Lambda^j P^T \quad (2.54)$$

If not, **use Jordan's decomposition**.

If the polynomial is invertible then X_t will be strictly invertible and $\varepsilon_t = X_t - BX_{t-1} + B^2X_{t-2} - B^3X_{t-3} + \dots$. In general, **there will be multiple representations of a VMA process, only one of which will be invertible, and this will coincide with its Wold representation**.^{2.1}

2.4.2 VMA(k) Process

It is given by:

$$\overbrace{X_t = \varepsilon_t + B_1\varepsilon_{t-1} + \dots + B_k\varepsilon_{t-k}}^{=\text{VMA}(k)}, \quad \varepsilon_t \sim WN(0, \Omega) \implies \quad (2.55a)$$

$$\mathbb{V}(X_t) = \Gamma(0) = \Omega + \sum_{j=1}^k B_j \Omega B_j^T; \quad (2.55b)$$

$$\text{cov}(X_t, X_{t-1}) = \sum_{j=1}^k B_j \Omega B_{j-1}^T; \quad (2.55c)$$

$$\text{cov}(X_t, X_{t-k}) = B_k \Omega; \text{ and} \quad (2.55d)$$

$$\text{cov}(X_t, X_{t-j}) = 0 \quad \forall j > k. \quad (2.55e)$$

^{2.1} For each eigenvalue, I can use its inverse. This is why we have 2^n possible representations.

We see a **(2k+1)-band block-diagonal structure for covariance matrix of $\text{vec}(X^T)$** . The autocovariance generating matrix is:

$$B(z)\Omega B^T(z^{-1}), \quad (2.56)$$

where

$$B(L) = I + B_1L + \dots B_kL^k. \quad (2.57)$$

The spectral density matrix is:

$$B(e^{-i\lambda})\Omega B^T(e^{i\lambda}) \quad (2.58)$$

To find the inverse, try to find $\Pi(L)$ such that:

$$\Pi(L)B(L) = \mathbb{I} \quad (2.59)$$

This gives rise to a system of linear difference equations of order k . Such systems can always be written as **first-order systems of linear difference equations with a larger number of equations**. A more direct solution is as follows. The **(classical) adjoint matrix (i.e. transpose of the cofactor matrix)** of:

$$B(L) = I + B_1L + \dots B_kL^k \quad (2.60)$$

is always well defined. The problem is the inverse of the determinant of $B(L)$:

$$\left| I + B_1L + \dots B_kL^k \right| \quad (2.61)$$

But this **determinant is related to the characteristic equation of the following matrix**:

$$\begin{pmatrix} B_1 & B_2 & \dots & B_k \\ I & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & I & 0 \end{pmatrix}. \quad (2.62)$$

Therefore, we **need the eigenvalues of this matrix to be on or inside the unit circle**. Long story short:

$$\Pi(L) = \frac{1}{|B(L)|} \text{adj}[B(L)]. \quad (2.63)$$

You need to show the **roots of this equation are OUTSIDE the unit circle**:

$$|B(L)| = 0. \quad (2.64)$$

There are **three problems with VMA(k)**:

1. I need to calculate k matrices with n roots each;
2. for the predictive purposes, VMA(k) is painful as I need to find the Wold representation; and
3. when you want to estimate the parameter values, not being able to write predictions is difficult.

2.4.3 VAR(1) Process

It is given by:

$$\overbrace{X_t = AX_{t-1} + u_t, X_0, \quad u_t \sim WN(0, \Omega)}^{=\text{VAR}(1)} \quad (2.65)$$

This **process is always invertible**. Hence:

$$X_t = \sum_{j=0}^{t-1} A^j u_{t-j} + A^t X_0. \quad (2.66)$$

Remark (Initial Condition Disappears). *Notice that $A^t \rightarrow 0$ when $t \rightarrow \infty$ iff $|\lambda_j(A)| < 1$ for all j .*

In that case:

$$X_t = (I - AL)^{-1}u_t \quad (2.67)$$

The effect of initial conditions is asymptotically negligible. But we can immediately achieve stationarity by drawing X_0 from its stationary distribution. If some of the **eigenvalues lie on the unit circle, then the VAR(1) process is neither stationary nor explosive.**

In the **stationary case**:

$$\mathbb{E}(X_t) = 0; \quad (2.68a)$$

$$\mathbb{V}(X_t) = \Gamma(0) = \sum_{j=0}^{\infty} A^j \Omega A^{jT} = A\Gamma(0)A^T + \Omega; \quad (2.68b)$$

$$\underbrace{\text{vec}[\Gamma(0)] = (I - A \otimes A)^{-1} \text{vec}(\Omega)}_{=\text{Vectorisation.}}; \text{ and} \quad (2.68c)$$

$$\Gamma(1) = A\Gamma(0) \implies \quad (2.68d)$$

$$\Gamma(j) = A\Gamma(j-1) = A^j\Gamma(0). \quad (2.68e)$$

VAR(1) models are ideal for carrying out multiperiod linear predictions:

$$\mathbb{P}(x_{t+k} \mid H_{t-1}) = \mathbb{P}(Ax_{t+k-1} + u_{t+k} \mid H_{t-1}) = A\mathbb{P}(x_{t+k-1} \mid H_{t-1}) = A^{k+1}x_{t-1} \quad (2.69)$$

If the **process is stationary, those predictions converge in mean square to 0 (its long-term mean) as $k \rightarrow \infty$.**

Theorem 2.4 (Vectorisation Property).

$$\text{vec}(ABC) = (C^T \otimes A) \text{vec}(B) \quad (2.70)$$

Definition 2.13 (Kronecker Product).

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} \quad (2.71)$$

Note that they do not need to be compatible.

Theorem 2.5 (Eigenvalues of Kronecker Products). *We have:*

$$\lambda(A \otimes B) = \lambda(A) \times \lambda(B). \quad (2.72)$$

Example 2.6 (Exploiting Vectorisation and Kronecker Products for VAR(1)).

$$\text{vec}[\Gamma(0)] = \text{vec}[A\Gamma(0)A^T] + \text{vec}[\Omega] \implies \quad (2.73a)$$

$$\text{vec}[\Gamma(0)] = \underbrace{(A \otimes A)}_{\text{We Need Its } |\lambda_j| \neq 1.} \text{vec}[\Gamma(0)] + \text{vec}[\Omega] \implies \quad (2.73b)$$

$$\text{vec}[\Gamma(0)] = (I - A \otimes A)^{-1} \text{vec}(\Omega). \quad (2.73c)$$

Example 2.7 (Autocorrelations). If we feel adventurous, we can get autocorrelations:

$$R(j) = \text{diag}[\Gamma(0)]^{-\frac{1}{2}} \times \Gamma(j) \times \text{diag}[\Gamma(0)]^{-\frac{1}{2}}. \quad (2.74)$$

2.4.4 VAR(h) Process

It is based on the same idea as the univariate case:

$$\overbrace{\left(I_N - A_1 L - \dots - A_h L^h \right)}^{=\text{VAR}(h)} X_t = u_t. \quad (2.75)$$

It can be written in the **companion form as a Super-VAR(1)**:

$$\underbrace{\begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-h+2} \\ X_{t-h+1} \end{pmatrix}}_{Nh \times 1 \text{ Dimension}} = \begin{pmatrix} A_1 & A_2 & \dots & A_{h-1} & A_h \\ I & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ X_{t-2} \\ \vdots \\ X_{t-h+1} \\ X_{t-h} \end{pmatrix} + \begin{pmatrix} u_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \quad (2.76)$$

Example 2.8 (AR(2) Univariate). We can express it as VAR(1):

$$\begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ x_{t-2} \end{pmatrix} + \begin{pmatrix} u_t \\ 0 \end{pmatrix} \implies \quad (2.77a)$$

$$\lambda^2 - \alpha_1 \lambda - \alpha_2 = 0. \quad (2.77b)$$

From here, you could use the VAR(1) formulae.

2.4.5 VARMA(h,k) Process

Consider:

$$\overbrace{\left(I_N - A_1 L - \dots - A_h L^h \right)}^{=\text{VARMA}(h,k)} X_t = \left(I_N + B_1 L + \dots + B_k L^k \right) \varepsilon_t. \quad (2.78)$$

It can be written as a **hyper-VAR(1)**:

$$\begin{pmatrix} X_t \\ \vdots \\ X_{t-h+1} \\ \varepsilon_t \\ \vdots \\ \varepsilon_{t-k+1} \end{pmatrix} = \begin{pmatrix} A_1 & \dots & A_h & B_1 & \dots & B_k \\ I & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & I & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & I & \ddots & \vdots \\ 0 & \dots & 0 & 0 & I & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ \vdots \\ X_{t-h} \\ \varepsilon_{t-1} \\ \vdots \\ \varepsilon_{t-k} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ \vdots \\ 0 \\ \varepsilon_t \\ \vdots \\ 0 \end{pmatrix}. \quad (2.79)$$

Example 2.9 (VARMA(1,1)). We have:

$$\begin{pmatrix} x_t \\ \varepsilon_t \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ \varepsilon_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ \varepsilon_t \end{pmatrix}. \quad (2.80)$$

2.5 Granger Causality

Consider the **following process**:

$$\underbrace{\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \sum_{j=0}^{\infty} \begin{pmatrix} \delta_{11j} & \delta_{12j} \\ \delta_{21j} & \delta_{22j} \end{pmatrix} \begin{pmatrix} u_{1t-j} \\ u_{2t-j} \end{pmatrix}}_{=\text{Bivariate Wold Representation.}} \quad (2.81)$$

where:

$$u_{it} = x_{it} - \mathbb{P}(x_{it} \mid H_{t-1}). \quad (2.82)$$

Alternatively, we may look at:

$$x_{1t} = \sum_{j=0}^{\infty} \delta_{1j}^u v_{1t-j}; \text{ and} \quad (2.83a)$$

$$v_{1t} = x_{1t} - \mathbb{P}(x_{1t} | H_{1t-1}). \quad (2.83b)$$


We have it because:


$$H_{1t-1} \subset H_{t-1} \implies \quad (2.84a)$$

$$\mathbb{E}(u_{1t}^2) \leq \mathbb{E}(v_{1t}^2). \quad (2.84b)$$

Definition 2.14 (Lack of Granger Causality). If $\mathbb{E}(u_{1t}^2) = \mathbb{E}(v_{1t}^2)$, so that $\mathbb{P}(x_{1t} | H_{1t-1}) = \mathbb{P}(x_{1t} | H_{t-1})$, then x_{2t} **does not Granger Cause** X_{1t} .

This means that there is no forecasting gain in using the lag values of x_{2t} . It's a **predictive concept** with nothing to do with causality. How do we do it?

 **VMA(k)**: find Wold representation (i.e. invertible) and check that all δ_{12j} s are 0.

 **VAR(h)**: there are two ways:

1. the Wold representation's δ_{12j} s are all zero; or
2. $a_{12i} = 0, i = 1, \dots, h$ because $D(L) = A^{-1}(L)$ will be diagonal if and only if so is $A(L)$ - for example, for x_2 to have no Granger causality on x_1 we need:

$$A(L)x_t = u_t \implies \quad (2.85a)$$

$$x_t = \underbrace{A(L)^{-1}}_{\text{Lower Triangular}} u_t. \quad (2.85b)$$

 **VARMA(h,k)**: same steps as with VMA(k).

Remark. Remember that this is not symmetric. x_{1t} can Granger cause x_{2t} , but not other way round!


More Than Two Series

Definition 2.15 (Granger Causally Prior). x is Granger Causally Prior (or GCP) to y **iff it is possible to group all the variables in the system into two blocks, Y_1 and Y_2 , such that y is in Y_2 and x is in Y_1 and Y_2 does not Granger-cause Y_1 .**

Example 2.10 (VAR(1), Trivariate). Consider:

$$\begin{pmatrix} y_t \\ x_t \\ z_t \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \\ z_{t-1} \end{pmatrix} + \varepsilon_t. \quad (2.86)$$

 z is **GCP** to y **iff either $a_{31} = a_{32} = 0$ or $a_{31} = a_{21} = 0$.**

 In the first case, the block triangular structure of A would be apparent if we put z first, while in the second case we would need to put y at the bottom of the vector.

2.6 Linear Transformations, Marginalisations, and Contemporaneous Aggregation

2.6.1 Linear transformations

We have X_t but we are interested in $Z_t = PX_t$. If P is a square matrix of full rank then $H_{X_t} = H_{Z_t}$ (where H is the space of all the linear combinations that I can form with Z and X). Start with the Wold representation:

$$X_t = D(L)u_t \implies \quad (2.87a)$$

$$Z_t = PX_t \implies \quad (2.87b)$$

$$Z_t = PD(L)u_t \implies \quad (2.87c)$$

$$Z_t = PD(L)P^{-1}Pu_t \implies \quad (2.87d)$$

$$Z_t = D^*(L)w_t; \quad (2.87e)$$

$$D_j^* = PD_jP^{-1}; \quad (2.87f)$$

$$w_t = Pu_t; \text{ and} \quad (2.87g)$$

$$\mathbb{V}(w_t) = \Omega^* = P\Omega P^T. \quad (2.87h)$$

In the projection terms:

$$\mathbb{P}(Z_t | H_{Z_{t-1}}) = \mathbb{P}(PX_t | H_{Z_{t-1}}) \implies \quad (2.88a)$$

$$\mathbb{P}(Z_t | H_{Z_{t-1}}) = P \mathbb{P}(X_t | H_{X_{t-1}}) \implies \quad (2.88b)$$

$$w_t = Z_t - \mathbb{P}(Z_t | H_{Z_{t-1}}) = Pu_t. \quad (2.88c)$$

Example 2.11 (VMA(1)). Consider:

$$X_t = \varepsilon_t + B\varepsilon_{t-1}. \quad (2.89)$$

Then:

$$Z_t = w_t + B^*w_{t-1} \text{ with} \quad (2.90a)$$

$$B^* = PBP^{-1}. \quad (2.90b)$$

Notice that if $|\lambda_j(B)| < 1$ for all j , then $|\lambda_j(B^*)| < 1$ for all j because the transformation is similar. Thus, the invertibility conditions are the same.

Example 2.12 (VAR(1)). We have:

$$X_t = AX_{t-1} + u_t. \quad (2.91)$$

Then:

$$Z_t = A^*Z_{t-1} + w_t \text{ with} \quad (2.92a)$$

$$A^* = PAP^{-1} \quad (2.92b)$$

Notice that if $|\lambda_j(A)| < 1$ for all j , then $|\lambda_j(A^*)| < 1$ for all j . Thus, the stationarity conditions are the same.

2.6.2 Marginalisation

We have the **Wold representation for X_t and want to find the representation of one element of it, say X_{1t} . Start with VMA(1).** The starting point is:

$$X_t = \varepsilon_t - B\varepsilon_{t-1}. \quad (2.93)$$

Since the **serial dependence of X_t dies away after one lag**, we want to find δ_1 and $\sigma_{v_1}^2$ for:

$$X_{1t} = v_{1t} + \delta_1 v_{1t-1}. \quad (2.94)$$

This means:

$$\rho_{11} = \frac{b_{11}\omega_{11} + b_{12}\omega_{12}}{(1 + b_{11}^2)\omega_{11} + b_{12}^2\omega_{22} + 2b_{11}b_{12}\omega_{12}} = \frac{\delta_1}{1 + \delta_1^2}. \quad (2.95)$$

Choose the invertible one and

$$\sigma_{v_1}^2 = \frac{\mathbb{V}(X_{1t})}{(1 + \delta_1^2)}. \quad (2.96)$$

When you work on such problems:

- * **VMA(k)**: same idea with k unknowns - always keep an invertible solution; and
- * **VAR(1)**: the starting point is $X_t = AX_{t-1} + u_t$.

VAR(1) Case Start with:

$$\Gamma(j) = A\Gamma(j-1); \text{ and} \quad (2.97a)$$

$$\text{vec}(\Gamma(0)) = [\mathbb{I} - (A \otimes A)]^{-1} \text{vec}(\Omega) \quad (2.97b)$$

Lemma 2.6.

$$(\mathbb{I}_n - AL)^{-1} = \frac{1}{|\mathbb{I}_n - AL|} \text{adj}(\mathbb{I}_n - AL) \quad (2.98)$$

where

$$|\mathbb{I}_n - AL| = 1 - \text{tr}(A)L + |A|L^2 \quad (2.99)$$

Hence:

$$\text{adj}(\mathbb{I}_n - AL) = \begin{pmatrix} 1 - a_{22}L & a_{12}L \\ a_{21}L & 1 - a_{11}L \end{pmatrix}. \quad (2.100)$$

The **final form of the process** is:

$$\underbrace{(1 - \text{tr}(A)L + |A|L^2) X_t = \text{adj}(\mathbb{I}_n - AL) u_t}_{=\text{The Final Form.}} \quad (2.101)$$

The **right-hand side is a VMA(1)**, so each of its components will be univariate MA(1)s. The **left-hand side is a common AR(2) polynomial, so both marginal processes will be at most an ARMA(2,1)**.

However, cancellation may reduce the final orders. If instead of 2 series, we had N , **then at most ARMA(N,N-1)**.

Remark. VARMA(h, k) processes are treated in the same way: write the final form, find the MA coefficients of the right-hand side and use the determinant as an AR polynomial.

2.6.3 Contemporaneous Aggregation

Starting from $X_t = (X_{1t}, X_{2t})^T$ suppose we are interested in:

$$Y_{1t} = w_1 X_{1t} + w_2 X_{2t}. \quad (2.102)$$

We can use the **two previous concepts sequentially**.

2.7 Time-Invariant Linear Filters

We focus on **linear homogeneous (i.e. time-invariant) filters in which the filter weights do not depend on t** . The idea is to generate:

$$Y_t = \Delta(L)X_t \quad (2.103)$$

and find the **matrix of autocovariance generating functions $\Psi_{yy}(z)$ from $\Psi_{xx}(z)$** .

📖 X_t is the **input of the filter**.

📖 Y_t is the **output of the filter**.

📖 $\Delta(L) = \sum_{r=-p}^q \Delta_r L^r$ is called the **transfer function of the filter**.

In the univariate case:

$$y_t = \sum_{r=-p}^q \delta_r x_{t-r}. \quad (2.104)$$

We use the filters because they can highlight some features of the series.

Example 2.13 (Standard MA Filter).

$$y_t = \underbrace{n^{-1} \sum_{i=0}^{n-1} x_{t-i}}_{\text{Flat}} \quad (2.105)$$

Example 2.14 (Exponentially Weighted MA Filter). Consider:

$$y_t = \sum_{i=0}^{\infty} \phi^i x_{t-i}. \quad (2.106)$$

It **gives more weight to recent observations!**

Both are used for **smoothing out oscillating series and capturing the trend**.

Example 2.15 (First Difference).

$$y_t = x_t - x_{t-1} = \Delta x_t \quad (2.107)$$

Example 2.16 (Seasonal Difference).

$$y_t = x_t - x_{t-d} = \Delta_d x_t \quad (2.108)$$

E.g., $d = 4$ for **quarterly data**.

Both are used to emphasise **changes in trends in smooth processes**.

The **important result** is that:

$$\Psi_{yy}(z) = \Delta(z) \Psi_{xx}(z) \Delta^T(z^{-1}). \quad (2.109)$$

As expected, the same relationship holds for the spectral density of the input and output series. Specifically:

$$f_{yy}(\lambda) = \Delta(e^{-i\lambda}) f_{xx}(\lambda) \Delta^T(e^{i\lambda}) \implies \quad (2.110a)$$

$$f_{yy}(\lambda) = \Delta(e^{-i\lambda}) f_{xx}(\lambda) \Delta^*(e^{-i\lambda}). \quad (2.110b)$$

Remark. Δ^* is the **complex conjugate transpose!**

Consider a VARMA process:

$$A(L)X_t = B(L)u_t \implies \quad (2.111a)$$

$$X_t = A(L)^{-1}B(L)u_t \implies \quad (2.111b)$$

$$f_{xx}(\lambda) = A^{-1} \left(e^{-i\lambda} \right) B \left(e^{-i\lambda} \right) \Omega B^T \left(e^{-i\lambda} \right) A^{-T} \left(e^{-i\lambda} \right). \quad (2.111c)$$

Given that all **VARMA processes can be regarded as rational filters applied to WN, their spectral density matrices satisfy this relationship**. In the univariate case, it is convenient to express the frequency response function of the filter in polar form so that:

$$\delta \left(e^{-i\lambda} \right) = \underbrace{G(\lambda)}_{\text{Amplitude Ch.}} \times \exp \{ i \underbrace{g(\lambda)}_{\text{Phase Ch.}} \} \quad (2.112)$$

where $G(\lambda)$ is known as the **gain of the filter** and $\phi(\lambda)$ as its **phase** because they capture the effects of the filter on the amplitude and phase shift at each frequency.

Some filters are designed for amplifying some frequencies and dampening others, so they focus on the gain without altering much of the phase, while others focus on phase shifts with a fairly constant gain.

2.8 Structural VARs and Impulse Responses

We observe X and suppose we know its Wold decomposition which, ignoring the linearly deterministic part, reduces to:

$$X_t = \sum_{j=0}^{\infty} D_j u_{t-j} \text{ where} \quad (2.113a)$$

$$u_t = X_t - \mathbb{P}(X_t | H_{t-1}). \quad (2.113b)$$

Note that:

$$D_j u_t = \mathbb{P}(X_{t+j} | H_t) - \mathbb{P}(X_{t+j} | H_{t-1}) \quad (2.114)$$

tells us **how predictions are revised when observation t becomes available**. Often, however, we have a macroeconomic model in mind with **some underlying structural shocks ε_t** such that

$$\mathbb{V}(\varepsilon_t) = \mathbb{I}_N; \text{ and} \quad (2.115a)$$

$$X_t = \sum_{j=0}^{\infty} B_j \varepsilon_{t-j}. \quad (2.115b)$$

The **orthogonality of the shocks matters since we want to modify something** (e.g. a monetary policy shock) leaving the rest unchanged. **Impulse response analysis pays attention to the impact on X of shocks in ε . Specifically, we want to compare X_t with:**

$$X_t^\dagger = \sum_{j=0}^{\infty} B_j \varepsilon_{t-j}^\dagger \quad (2.116)$$

with

$$\varepsilon_{it-j}^\dagger = \varepsilon_{it-j} + 1 \quad (2.117)$$

and **all other elements are the same**.

Dynamic Identification Problem

≡ There is no reason why $X_t = \sum_{j=0}^{\infty} B_j \varepsilon_{t-j}$ should be invertible.

- ⏏ Invertibility fails **whenever** ε_t **contains more elements than** u_t , but it can also happen even if the dimensions are the same.

Static Specification Problem We have:

$$u_t = B_0 \varepsilon_t \implies \quad (2.118a)$$

$$\mathbb{V}(u_t) = B_0 B_0^T = \Omega. \quad (2.118b)$$

But for any orthonormal Q :

$$Q Q^T = Q^T; \quad (2.119a)$$

$$Q Q^T = Q^T Q = \mathbb{I}; \quad (2.119b)$$

$$B_0 B_0^T = B_0 Q Q^T B_0^T = \Omega. \quad (2.119c)$$

- ⏏ **Nothing else can be obtained from the autocovariance of the model, but restrictions from macroeconomic theory might help.**

- ⏏ Such restrictions usually have **an interpretation as assumptions about delays in the reaction of particular variables to certain shocks**, or long-run neutral effects of other shocks.

2.9 Dynamic Models with Latent Variables

2.9.1 Linear, Time-Invariant, State Space Models

Consider:

$$\begin{aligned} & \overbrace{\mathbf{y}_t = \boldsymbol{\pi} + \mathbf{C}(\boldsymbol{\theta}) \mathbf{x}_t}^{\text{=Observed Equation}} \\ & \overbrace{\mathbf{x}_t = \mathbf{A}(\boldsymbol{\theta}) \mathbf{x}_{t-1} + \mathbf{B}(\boldsymbol{\theta}) \mathbf{u}_t}^{\text{=Transition Equation}} \\ & \mathbf{u}_t \mid \mathcal{I}_{t-1}; \boldsymbol{\pi}, \boldsymbol{\theta} \sim N[\mathbf{0}, \boldsymbol{\Omega}(\boldsymbol{\theta})]. \end{aligned} \quad (2.120)$$

Note that:

- ⏏ \mathbf{y}_t contains N observed variables;
- ⏏ $\boldsymbol{\pi}$ contains its means;
- ⏏ \mathbf{x}_t are M state variables, $M \geq N$;
- ⏏ \mathbf{u}_t are k white noise shocks to the state variables, $M \geq k \geq N$;
- ⏏ \mathcal{I}_{t-1} is an **information set that contains the values of \mathbf{y}_t and \mathbf{x}_t up to, and including time $t-1$** ;
- ⏏ \mathbf{A} is the companion matrix of the transition equation;
- ⏏ \mathbf{B} selects the shocks that affect the state variables;
- ⏏ \mathbf{C} contains the coefficients of the measurement equations; and
- ⏏ $\boldsymbol{\theta}$ are the unknown model parameters (other than $\boldsymbol{\pi}$).

Example 2.17 (Basic UCARIMA model). We have:

$$y_t = \pi + x_t + u_t; \quad (2.121a)$$

$$\alpha_x(L)x_t = \beta_x(L)f_t; \quad (2.121b)$$

$$\alpha_u(L)u_t = \beta_u(L)v_t; \text{ and} \quad (2.121c)$$

$$\begin{pmatrix} f_t \\ v_t \end{pmatrix} \Big| \mathcal{I}_{t-1}; \pi, \boldsymbol{\theta} \sim \mathcal{N} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_f^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix} \right]. \quad (2.121d)$$

x_t is the “signal” component. u_t the orthogonal “non-signal” component. $\alpha_x(L)$ and $\alpha_u(L)$ are one-sided polynomials of orders p_x and p_u . $\beta_x(L)$ and $\beta_u(L)$ are one-sided polynomials of orders q_x and q_u , coprime with $\alpha_x(L)$ and $\alpha_u(L)$, respectively. The **UCARINMA** stands for **unobserved components of ARIMA**,

Example 2.18 (Exact Dynamic Factor Model with a Single Common Factor). Consider:

$$\begin{pmatrix} y_{1,t} \\ \vdots \\ y_{N,t} \end{pmatrix} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_N \end{pmatrix} + \begin{pmatrix} c_{1,0} \\ \vdots \\ c_{N,0} \end{pmatrix} x_t + \begin{pmatrix} c_{1,1} \\ \vdots \\ c_{N,1} \end{pmatrix} x_{t-1} + \begin{pmatrix} u_{1,t} \\ \vdots \\ u_{N,t} \end{pmatrix}; \quad (2.122a)$$

$$\alpha_x(L)x_t = \beta_x(L)f_t, \quad \alpha_{u_i}(L)u_{i,t} = \beta_{u_i}(L)v_{i,t}; \text{ and} \quad (2.122b)$$

$$(f_t, v_{1,t}, \dots, v_{N,t}) \mid \mathcal{I}_{t-1}; \boldsymbol{\theta} \sim \mathcal{N} [0, \text{diag} (1, \gamma_1, \dots, \gamma_N)]. \quad (2.122c)$$

$\alpha_x(L)$ and $\alpha_{u_i}(L)$ are polynomials of orders p_x and p_{u_i} , respectively, while $\beta_x(L)$ and $\beta_{u_i}(L)$ are polynomials of orders q_x and q_{u_i} . There are **three different dynamic features**:

1. common factors follow ARMA processes;
2. specific factors follow ARMA processes; and
3. factor loadings are dynamic.

If we eliminated all three, we would go back to static factor analysis.

2.9.2 Structural and Reduced Forms: UCARIMA Model

Under stationarity, the **spectral density of y_t will be**:

$$g_{yy}(\lambda) = g_{xx}(\lambda) + g_{uu}(\lambda); \quad (2.123a)$$

$$g_{xx}(\lambda) = \sigma_f^2 \frac{\beta_x(e^{-i\lambda t}) \beta_x(e^{i\lambda t})}{\alpha_x(e^{-i\lambda t}) \alpha_x(e^{i\lambda t})}; \text{ and} \quad (2.123b)$$

$$g_{uu}(\lambda) = \sigma_v^2 \frac{\beta_u(e^{-i\lambda}) \beta_u(e^{i\lambda})}{\alpha_u(e^{-i\lambda t}) \alpha_u(e^{i\lambda t})}. \quad (2.123c)$$

Hence, the reduced form model will be an ARMA process with maximum orders:

$$p_y = p_x + p_u; \text{ and} \quad (2.124a)$$

For the AR polynomial $\alpha_y(\cdot) = \alpha_x(\cdot)\alpha_u(\cdot)$.

$$q_y = \max(p_x + q_u, q_x + p_u) \quad (2.124b)$$

For the MA polynomial $\beta_y(\cdot)$.

Cancellation will trivially occur when $\alpha_x(\cdot)$ and $\alpha_u(\cdot)$ share common roots, but there could also be other cases. $\beta_y(L)$ and σ_a^2 (the variance of the univariate Wold innovations, a_t) are obtained by matching autocovariances. Assuming strict invertibility of the MA part, we could then obtain the reduced form innovations a_t from the observed process by means of the one-sided filter:

$$\frac{\alpha_y(e^{-i\lambda})}{\beta_y(e^{-i\lambda})}. \quad (2.125)$$

2.9.3 Structural and Reduced Forms: Dynamic Factor Model

Once again, we assume that the observed series are covariance stationary, at least after a suitable transformation. Under **stationarity, the spectral density matrix is proportional to**:

$$\mathbf{G}_{yy}(\lambda) = \mathbf{c} \begin{pmatrix} e^{-i\lambda} \end{pmatrix} g_{xx}(\lambda) \mathbf{c}^T \begin{pmatrix} e^{i\lambda} \end{pmatrix} + \mathbf{G}_{uu}(\lambda); \quad (2.126a)$$

$$g_{xx}(\lambda) = \frac{\beta_x(e^{-i\lambda}) \beta_x(e^{i\lambda})}{\alpha_x(e^{-i\lambda}) \alpha_x(e^{i\lambda})}; \quad (2.126b)$$

$$\mathbf{G}_{uu}(\lambda) = \text{diag}[g_{u_1 u_1}(\lambda), \dots, g_{u_N u_N}(\lambda)]; \text{ and} \quad (2.126c)$$

$$g_{u_i u_i}(\lambda) = \gamma_i \frac{\beta_{u_i}(e^{-i\lambda}) \beta_{u_i}(e^{i\lambda})}{\alpha_{u_i}(e^{-i\lambda}) \alpha_{u_i}(e^{i\lambda})}, \quad (2.126d)$$

which inherits the exact single factor structure of the unconditional covariance matrix of a static factor model. The **reduced form** is:

$$\mathbf{y}_t = \boldsymbol{\pi} + \mathbf{c}x_t + \mathbf{u}_t; \quad (2.127a)$$

$$\mathbf{u}_t = \mathbf{v}_t; \text{ and} \quad (2.127b)$$

$$x_t = \alpha_{x1}x_{t-1} + f_t. \quad (2.127c)$$

It is easy to see that **its autocovariance structure corresponds to a special case of a VARMA (1, 1) model since**:

$$(1 - \alpha_{x1}L)(\mathbf{y}_t - \boldsymbol{\pi}) = \mathbf{c}f_t + (1 - \alpha_{x1}L)\mathbf{v}_t, \quad (2.128)$$

whose **right hand side has the autocovariance structure of a VMA(1)**. Although the VAR part is clearly scalar, figuring out the coefficients of the invertible representation of its VMA part is not a trivial task.

2.9.4 Identification of UCARIMA Models

Without **restrictions on the orders of the different AR and MA polynomials, one cannot non-parametrically identify the signal and non-signal components**, as the only constraints are:

$$g_{yy}(\lambda) = g_{xx}(\lambda) + g_{uu}(\lambda); \text{ and} \quad (2.129a)$$

$$g_{xx}(\lambda), g_{uu}(\lambda) \geq 0. \quad (2.129b)$$

Let c denote the number of common roots of $\alpha_x(L)$ and $\alpha_u(L)$. The basic UCARIMA model will be parametrically identified (except at a set of parameter values of measure 0) when there are no additional restrictions on the AR and MA polynomials if and only if either $p_x \geq q_x + c + 1$ or $p_u \geq q_u + c + 1$ (see Hotta (1989)). Thus, at least one of the components must be a “top-heavy” ARMA process (i.e., a process in which the AR order exceeds the MA one).

2.9.5 Identification of dynamic single factor models

There are two identification issues:

1. the **nonparametric identification of $\mathbf{c} \begin{pmatrix} e^{-i\lambda} \end{pmatrix} g_{xx}(\lambda) \mathbf{c}^T \begin{pmatrix} e^{i\lambda} \end{pmatrix}$ and $\mathbf{G}_{uu}(\lambda)$** ; and
2. the **parametric identification of $\mathbf{c} \begin{pmatrix} e^{-i\lambda} \end{pmatrix}$ and $g_{xx}(\lambda)$** .

The answer to the first question is easy when $\mathbf{G}_{uu}(\lambda)$ is a diagonal, full rank matrix provided N is sufficiently large. **$N \geq 3$ is the so-called Ledermann bound for single factor models**. To see this, consider a univariate model. We need to identify 4 parameters with only 3 pieces of information:

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} c_1^2 + \gamma_1 & c_1 c_2 \\ c_1 c_2 & c_2^2 + \gamma_2 \end{pmatrix}. \quad (2.130)$$

The **second question is trickier, as we could always write any dynamic factor model in terms of white noise common factors**. Nevertheless, if the N polynomials $c_i(\cdot)$ do not share a common root, then it is easy to prove parametric identification. In this regard, the **assumption of ARMA (p_x, q_x) dynamics for the common factor can be regarded as a parsimonious way of modelling a common infinite distributed lag in the $c_i(\cdot)$ s**. Additional restrictions are necessary in the case of multiple factors, but the advantage of using spectral methods is that we can rely on the extensive literature on static factor analysis.

2.9.6 The Wiener-Kolmogorov Filter

Example 2.19 (Intuition: Univariate Case, Linear Projection).

$$y_t = x_t + u_t; \quad (2.131a)$$

$$\mathbb{E} \begin{pmatrix} x_t \\ u_t \end{pmatrix} = \mathbf{0}; \text{ and} \quad (2.131b)$$

$$\mathbb{P}(x_t | \langle y_t \rangle) = \frac{\text{cov}(x_t, y_t)}{\mathbb{V}(y_t)} y_t = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} y_t. \quad (2.131c)$$

The filter conducts a similar decomposition, frequency-by-frequency.

Let:

$$\mathbf{y}_t - \boldsymbol{\pi} = \int_{-\pi}^{\pi} e^{i\lambda t} d\mathbf{Z}_{\mathbf{y}}(\lambda); \text{ and} \quad (2.132a)$$

$$\mathbb{V}[d\mathbf{Z}_{\mathbf{y}}(\lambda)] = \mathbf{G}_{\mathbf{y}\mathbf{y}}(\lambda) d\lambda, \quad (2.132b)$$

denote **the spectral decomposition of the observed process**.

Definition 2.16 (The Wiener-Kolmogorov Two-Sided Filter). The **Wiener-Kolmogorov two-sided filter for the state variables \mathbf{x}_t at each frequency is given by:**

$$\mathbf{G}_{\mathbf{xx}}(\lambda) \mathbf{C}^T \mathbf{G}_{\mathbf{yy}}^{-1}(\lambda) d\mathbf{Z}_{\mathbf{y}}(\lambda), \quad (2.133)$$

so that the **spectral density of the resulting smoother $\mathbf{x}_{t|\infty}^K$ is:**

$$\mathbf{G}_{\mathbf{xx}}(\lambda) \mathbf{C}^T \mathbf{G}_{\mathbf{yy}}^{-1}(\lambda) \mathbf{C} \mathbf{G}_{\mathbf{xx}}(\lambda). \quad (2.134)$$

Hence, the **spectral density of the final estimation error $\mathbf{x}_t - \mathbf{x}_{t|\infty}^K$ will be given by:**

$$\mathbf{G}_{\mathbf{xx}}(\lambda) - \mathbf{G}_{\mathbf{xx}}(\lambda) \mathbf{C}^T \mathbf{G}_{\mathbf{yy}}^{-1}(\lambda) \mathbf{C} \mathbf{G}_{\mathbf{xx}}(\lambda). \quad (2.135)$$

Having obtained these, we can easily obtain the smoother for $\mathbf{u}_t, \mathbf{u}_{t|\infty}^K$, by applying to $\mathbf{x}_{t|\infty}^K$ the one-sided filter

$$[\mathbf{B}^T(\boldsymbol{\theta}) \mathbf{B}(\boldsymbol{\theta})]^{-1} \mathbf{B}^T(\boldsymbol{\theta}) [\mathbf{I} - \mathbf{A}(\boldsymbol{\theta}) e^{-i\lambda}]. \quad (2.136)$$

Likewise, we can derive its spectral density, as well as the spectral density of its final estimation error $\mathbf{u}_t - \mathbf{u}_{t|\infty}^K$. Finally, we can obtain the autocovariance matrices of $\mathbf{x}_{t|\infty}^K, \mathbf{u}_{t|\infty}^K$ and their final estimation errors by applying the usual inverse Fourier transformation:

$$\text{cov}(\mathbf{z}_t, \mathbf{z}_{t-k}) = \int_{-\pi}^{\pi} e^{i\lambda k} \mathbf{G}_{\mathbf{zz}}(\lambda) d\lambda. \quad (2.137)$$

Example 2.20 (The Wiener-Kolmogorov Filter: UCARIMA Models). In this case, the Wiener-Kolmogorov two-sided filter for the signal x_t at each frequency is given by:

$$g_{xx}(\lambda) g_{yy}^{-1}(\lambda) dZ_y(\lambda), \quad (2.138)$$

so that the spectral density of the smoother $x_{t|\infty}^K$ is:

$$g_{x^K x^K}(\lambda) = \frac{g_{xx}^2(\lambda)}{g_{yy}(\lambda)} \implies \quad (2.139a)$$

$$g_{x^K x^K}(\lambda) = \frac{g_{xx}(\lambda)}{g_{xx}(\lambda) + g_{uu}(\lambda)} g_{xx}(\lambda) \implies \quad (2.139b)$$

$$g_{x^K x^K}(\lambda) = R_{xx}^2(\lambda) g_{xx}(\lambda). \quad (2.139c)$$

$R_{xx}^2(\lambda)$ provides a frequency by frequency measure of signal strength.

Hence, the spectral density of the final estimation error $x_t - x_{t|\infty}^K$ will be given by:

$$g_{xx}(\lambda) - g_{x^K x^K}(\lambda) = [1 - R_{xx}^2(\lambda)] g_{xx}(\lambda) = \omega_{xx}(\lambda). \quad (2.140)$$

Having obtained these, **we can easily obtain the smoother for $f_t, f_{t|\infty}^K$, by applying to $x_{t|\infty}^K$ the one-sided filter:**

$$\frac{\alpha_x(e^{-i\lambda})}{\beta_x(e^{-i\lambda})}. \quad (2.141)$$

Likewise, we **can derive its spectral density, as well as the spectral density of its final estimation error $f_t - f_{t|\infty}^K$.**

Example 2.21 (The Wiener-Kolmogorov Filter: Dynamic Factor Models). In this case, the Wiener-Kolmogorov two-sided filter for the common factor x_t at each frequency is given by:

$$\mathbf{c}^T(e^{i\lambda}) g_{xx}(\lambda) \mathbf{G}_{yy}^{-1}(\lambda) d\mathbf{Z}_y(\lambda), \quad (2.142)$$

so that the spectral density of the smoother $x_{t|\infty}^K$ is:

$$g_{xx}^2(\lambda) \mathbf{c}^T(e^{i\lambda}) \mathbf{G}_{yy}^{-1}(\lambda) \mathbf{c}(e^{-i\lambda}). \quad (2.143)$$

Computations can be considerably sped up by exploiting that:

$$\mathbf{G}_{yy}^{-1}(\lambda) = \mathbf{G}_{uu}^{-1}(\lambda) - \omega(\lambda) \mathbf{G}_{uu}^{-1}(\lambda) \mathbf{c}(e^{-i\lambda t}) \mathbf{c}^T(e^{i\lambda t}) \mathbf{G}_{uu}^{-1}(\lambda); \text{ and} \quad (2.144a)$$

$$\omega(\lambda) = \left[g_{xx}^{-1}(\lambda) + \mathbf{c}^T(e^{i\lambda t}) \mathbf{G}_{uu}^{-1}(\lambda) \mathbf{c}(e^{-i\lambda t}) \right]^{-1}. \quad (2.144b)$$

For example:

$$g_{xx}^2(\lambda) \mathbf{c}^T(e^{i\lambda}) \mathbf{G}_{yy}^{-1}(\lambda) \mathbf{c}(e^{-i\lambda}) = \omega(\lambda) g_{xx}(\lambda) \mathbf{c}^T(e^{i\lambda}) \mathbf{G}_{uu}^{-1}(\lambda) \mathbf{c}(e^{-i\lambda}). \quad (2.145)$$

Hence, the **spectral density of the final estimation error $x_t - x_{t|\infty}^K$ will be given by:**

$$g_{xx}(\lambda) - g_{x^K x^K}(\lambda) \mathbf{c}^T(e^{i\lambda}) \mathbf{G}_{yy}^{-1}(\lambda) \mathbf{c}(e^{-i\lambda}) = \omega(\lambda). \quad (2.146)$$

Having obtained these, **we can easily obtain the smoother for $f_t, f_{t|\infty}^K$, by applying to $x_{t|\infty}^K$ the one-sided filter:**

$$\frac{\alpha_x(e^{-i\lambda})}{\beta_x(e^{-i\lambda})}. \quad (2.147)$$

Likewise, we **can derive its spectral density, as well as the spectral density of its final estimation error $f_t - f_{t|\infty}^K$.**

2.9.7 Dynamic Models with Latent Variables

Consider:

$$\overbrace{X_t = HY_t + Rv_t;}^{\equiv \text{Measurement Eq.}} \quad (2.148a)$$

$$\overbrace{Y_t = FY_{t-1} + G\varepsilon_t;}^{\equiv \text{Transition Eq.}} \quad (2.148b)$$

$$\mathbb{E}^* \left[\begin{pmatrix} v_t \\ \varepsilon_t \end{pmatrix} \middle| \begin{matrix} X_{t-1}, X_{t-2}, \dots \\ Y_{t-1}, Y_{t-2}, \dots \end{matrix} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \text{ and} \quad (2.148c)$$

$$\mathbb{V} \left[\begin{pmatrix} v_t \\ \varepsilon_t \end{pmatrix} \right] = \begin{pmatrix} \Sigma & 0 \\ 0 & \Omega \end{pmatrix} \quad (2.148d)$$

🍏 The **measurement equation** relates observed variables to latent variables.

🍏 The **transition equation** specifies the evolution of state variables.

🍏 Our purpose is to obtain $\mathbb{E}^*(Y_\tau | X_{t-1}, X_{t-2}, \dots)$.

🍏 This could be a forecasting ($\tau > t$), filtering ($\tau = t$) or smoothing ($\tau < t$) exercise.

For example, we have three measures of output: expenditure, income and production (value added) but there is only one true underlying GDP.

2.9.8 Kalman Filter

Note that:

$$\mathbb{P}(X_t | \mathcal{H}_{x,t-1}) = H \mathbb{P}(Y_t | \mathcal{H}_{x,t-1}); \text{ and} \quad (2.149a)$$

$$\mathbb{P}(Y_t | \mathcal{H}_{x,t-1}) = F \mathbb{P}(Y_{x,t-1} | \mathcal{H}_{t-1}) \implies \quad (2.149b)$$

$$\mathbb{P}(Y_t | \mathcal{H}_{x,t-1}) = FY_{t-1|t-1} \implies \quad (2.149c)$$

$$H = \mathbb{E} \left\{ [X_t - \mathbb{P}(X_t | \mathcal{H}_{x,t-1})] [X_t - \mathbb{P}(X_t | \mathcal{H}_{x,t-1})]^T \right\}. \quad (2.149d)$$

The **forecasting equations** are:

$$X_{t|t-1} = \mathbb{E}^*(X_t | X_{t-1}, X_{t-2}, \dots) = HY_{t|t-1}; \text{ and} \quad (2.150a)$$

$$Y_{t|t-1} = \mathbb{E}^*(Y_t | X_{t-1}, X_{t-2}, \dots) = FY_{t-1|t-1}. \quad (2.150b)$$

The **forecasting errors (unpredictable)** are:

$$\mathbb{V}(Y_t - Y_{t|t-1}) = F \mathbb{P}_{t-1|t-1} F^T + G\Omega G^T = P_{t|t-1}; \quad (2.151a)$$

$$\mathbb{V}(X_t - X_{t|t-1}) = H P_{t|t-1} H^T + R\Sigma R^T; \text{ and} \quad (2.151b)$$

$$\text{cov}(X_t - X_{t|t-1}, Y_t - Y_{t|t-1}) = H P_{t|t-1}. \quad (2.151c)$$

We **can update the equations**:

$$Y_{t|t} = Y_{t|t-1} + \mathbb{P}(Y_t - Y_{t|t-1} | \{X_t - X_{t|t-1}\}) \implies \quad (2.152a)$$

$$Y_{t|t} = Y_{t|t-1} + \mathbb{P}_{t|t-1} H^T (H \mathbb{P}_{t|t-1} H^T + R\Sigma R^T)^{-1} (X_t - X_{t|t-1}); \text{ and} \quad (2.152b)$$

$$\mathbb{P}_{t|t} = \mathbb{P}_{t|t-1} - P_{t|t-1} H^T (H P_{t|t-1} H^T + R\Sigma R^T)^{-1} H \mathbb{P}_{t|t-1}. \quad (2.152c)$$

Definition 2.17 (Kalman Gain). $\mathbb{P}_{t|t-1} H^T (H \mathbb{P}_{t|t-1} H^T + R\Sigma R^T)^{-1}$ is called the **Kalman gain**.

- ⊕ The algorithm is recursive and thus ideal for a digital computer.
- ⊕ The **highest computational cost is the calculation of $\mathbb{P}_{t|t-1}$ and $\mathbb{P}_{t|t}$.**
- ⊕ Often, these matrices will stabilise if the filter reaches a steady state, in which case there is no need to update them any longer.
- ⊕ But how do we start? Choosing $Y_{0|0}$ and $P_{0|0}$ (or $Y_{1|0}$ and $P_{1|0}$).
- ⊕ If Y is a covariance stationary process, it makes sense to set $Y_{0|0} = E(Y_t)$ and $P_{0|0} = V(Y_t)$.
- ⊕ If not, $Y_{0|0} = 0$ and $P_{0|0} = kI$ with $k \rightarrow \infty$.
- ⊕ **Multiperiod prediction is easy since the state variables follow a VAR(1):**

$$X_{t+k|t-1} = H y_{t+k|t-1}, Y_{t+k|t-1} = F^{k+1} Y_{t-1|t-1}. \quad (2.153)$$

2.9.9 Kalman Smoother

The **Kalman Filter: $Y_{t|t}$ for each t** . Sometimes (e.g., guiding a spacecraft) this is what we need.

In economics often we have a fixed data set for $t = 1, \dots, T$ and would like to know: $Y_{t|T}$ for each t . This makes the full sample smoother. The recursions will be given by:

$$Y_{t|T} = Y_{t|t} + \mathbb{P}_{t|t} F^T \mathbb{P}_{t+1|t}^{-1} (Y_{t+1|T} - Y_{t+1|t}); \text{ and} \quad (2.154a)$$

$$P_{t|T} = \mathbb{P}_{t|t} - \mathbb{P}_{t|t} F^T \mathbb{P}_{t+1|t}^{-1} (\mathbb{P}_{t+1|t} - \mathbb{P}_{t+1|T}) \mathbb{P}_{t+1|t}^{-1} F \mathbb{P}_{t|t}, \quad (2.154b)$$

with **initial conditions $Y_{T|T}$ and $P_{T|T}$** . Alternatively, we may be interested in the fixed interval smoother $Y_{t|t+\tau}$ for fixed τ as t changes. Or the fixed point smoother $Y_{t|t+\tau}$ as τ increases for a fixed t . We can **mechanically obtain the latter with the augmented state space model**:

$$X_\tau = \begin{pmatrix} H & 0 \end{pmatrix} \begin{pmatrix} Y_{\tau-1} \\ Y_{\tau-1}^* \end{pmatrix} + R v_\tau \quad (2.155a)$$

$$\begin{pmatrix} Y_\tau \\ Y_\tau^* \end{pmatrix} = \begin{pmatrix} F & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} Y_{\tau-1} \\ Y_{\tau-1}^* \end{pmatrix} + \begin{pmatrix} G \\ 0 \end{pmatrix} \varepsilon_\tau \quad (2.155b)$$

with initial conditions $Y_{\tau|\tau} = Y_{\tau|\tau}^* = Y_{t|t}$ and $P_{t|t}$.

3 Integration and Non-Linear Models

3.1 Models for Non-Stationary Series

Many time series are non-stationary. So we need to learn how to transform them so that they become stationary. We will consider just a few ways of doing so. We begin by **focusing on models in which x_t has a trend so that $\mathbb{E}(x_t) = \mu(t)$, i.e. lack of mean stationarity**. How to deal with them? One possibility is to model the trend and then get rid of it.

Definition 3.1 (Trand Stationary Process). For instance, take:

$$y_t = m(t) + x_t \quad (3.1)$$

where **$m(t)$ is the trend function** and **x_t is covariance stationary**.

The usual problem is that **we observe only one realisation of the process, so we need to make additional assumptions**.

Definition 3.2 (Polynomial Trend Function).

$$m(t) \equiv \sum_{j=0}^k a_j t^j \quad (3.2)$$

Definition 3.3 (Seasonal Trend Function).

$$m_S(t) \equiv \sum_{j=0}^S \gamma_j \mathbb{I}\{t \in j\} \quad (3.3)$$

Alternatively, we **could use sine and cosine functions with cycles of one day**, week, month, etc. for modelling e.g. residential demand for electricity or trading volumes in financial markets, petrol consumption, etc.

Definition 3.4 (Exponential Trend Function). It is widely used when dealing with cumulative processes. However, **variances explode, so an additional instantaneous transformation is often required**:

$$y_t = \ln x_t. \quad (3.4)$$

Example 3.1. Consider GDP, take logs, and then fit a linear trend. However, in some cases, this doesn't work. What happens if there is a big shock like the first oil crisis or the recent financial crisis? Does GDP eventually resume its fluctuations around its pre-crisis deterministic trend, or do those shocks have permanent effects?

Definition 3.5 (Random Walk (RW) With Drift). It is given by:

$$y_t = \mu + y_{t-1} + u_t. \quad (3.5)$$

If you assume:

$$y_0 = 0 \implies \quad (3.6a)$$

$$\mathbb{E}(y_t) = \mu_t \text{ since} \quad (3.6b)$$

$$\mathbb{E}(u_s) = 0. \quad (3.6c)$$

You could then think of modelling:

$$\tilde{y}_t \equiv y_t - \mathbb{E}(y_t) = \sum_{s=1}^t u_s \quad (3.7)$$

which still is a **driftless RW whose variance increases linearly with t** .

To achieve stationarity in mean and variance, we need:

$$\Delta \equiv (1 - L) \text{ so that} \quad (3.8a)$$

$$\Delta y_t = \mu + u_t. \quad (3.8b)$$

Similarly, if there is nonstationarity in the seasonality, use:

$$\Delta_s \equiv 1 - L^s. \quad (3.9)$$

Occasionally we may need a combination of those two filters, or even the duplication of one of them.

Definition 3.6 (Integrated Processes). $y_t \sim I(d)$ if $\Delta^d y_t$ is covariance stationary and strictly invertible.

The **strictly invertible requirement prevents over-differencing**. For example, take:

$$x_t = u_t - u_{t-1} \sim I(-1) \quad (3.10)$$

since the **underlying White Noise u_t , which can be regarded as the first differences of a driftless random walk, is already $I(0)$** . Here, we take too many differences! u_t is already $I(0)$.

An important **special case is the ARIMA(h,d,k) process**, whose d^{th} differences are ARMA(h,k), which can also be regarded as ARMA(p+d,k) processes in which the AR polynomials contain d roots equal to 1.

Definition 3.7 (Seasonally Integrated Processes). **$\Delta_s^p y_t$ is covariance stationary and strictly invertible.**

Definition 3.8 (Fractionally Integrated Processes). Consider:

$$(1 - L)^d = 1 - dL + \frac{1}{2}d(1 - d)L^2 - \frac{1}{6}d(1 - d)(2 - d)L^3 + \dots \quad (3.11)$$

Although $(1 - L)^d$, with $d \in \mathbb{R}$, can be regarded as a generalisation of the ordinary difference filter, the main objective of the process above is to capture autocorrelation structures that decay slowly. As we saw briefly in section 1 if $d < \frac{1}{2}$, such fractionally integrated processes are covariance stationary (i.e. square summable) but not absolute summable.

Definition 3.9 (Covariance Stationary Long Memory Processes). In general, **covariance stationary “long memory processes” are such that their autocovariance generating function evaluated at 1, i.e. $\psi_{xx}(1)$, is unbounded.**

3.2 Cointegration

Definition 3.10 (Cointegration). Let $x_t \sim I(1), y_t \sim I(1)$. **x_t and y_t are cointegrated of orders (1,1) (or $CI(1,1)$ for short) if there exist $\gamma_1, \gamma_2 \neq 0$:**

$$\gamma_1 x_t + \gamma_2 y_t \sim I(0). \quad (3.12)$$

Example 3.2 (Cointegrated Processes). **Examples of cointegrated processes include:**

- ▮▮▮ per capita consumption and per capita income: $\ln c_t - \ln y_t \sim I(0)$; and
- ▮▮▮ the dividend-price ratio for a given company or an aggregate index.

We see **two requirements**:

1. x_t and y_t share the same order of integration; and
2. one has to find the cointegrating vector $(\gamma_0, \gamma_1)^T$.

Example 3.3 (Bivariate VAR(2)). Consider:

$$X_t = A_1 X_{t-1} + A_2 X_{t-2} + u_t, \quad (3.13)$$

or, in short:

$$A(L)X_t = u_t. \quad (3.14)$$

$|A(L)|$ should have at least one unit root! We know from section 2 that marginals are ARMA(4,2) because in:

$$|A(L)|X_t = \text{adj}[A(L)]u_t \quad (3.15)$$

$|A(L)|$ is of order 4 and $\text{adj}[A(L)]$ of order 2.

In principle, though, some of the MA roots may cancel with some of the AR roots, so the orders could be lower. In addition, **even if there is no cancellation for the two elements of X_t , if we consider all possible linear combinations of their elements we may find one for which cancellation occurs.** But how can we find the cointegrating vector? **Let's write the so-called vector "error correction" representation:**

$$\underbrace{\Delta X_t = -A(1)X_{t-1} - A_2\Delta X_{t-1} + u_t}_{\equiv \text{Error Correction Representation}} \quad (3.16)$$

This is derived from:

$$X_t = A_1X_{t-1} + A_2X_{t-2} + u_t \implies \quad (3.17a)$$

$$X_t - X_{t-1} = (A_1 - \mathbb{I})X_{t-1} + A_2X_{t-2} + u_t \implies \quad (3.17b)$$

$$X_t - X_{t-1} = (A_1 + A_2 - \mathbb{I})X_{t-1} + A_2(X_{t-2} - X_{t-1}) + u_t \implies \quad (3.17c)$$

$$\Delta X_t = A(1)X_{t-1} + A_2\Delta X_{t-1} + u_t, \text{ which is not the same as:} \quad (3.17d)$$

$$\Delta X_t = A_1\Delta X_{t-1} + A_2\Delta X_{t-2} + \Delta u_t. \quad (3.17e)$$

Given that **the elements of ΔX_t are $I(0)$ by assumption, and so are the elements of $u_t, A(1)X_{t-1}$ has to be $I(0)$.** If $\text{rank}(A(1)) = 2, X_t \sim I(0)$ **which is a contradiction because any vector will be a cointegrating vector.** In contrast, **if $\text{rank}(A(1)) = 0, X_t \sim I(1)$ and \nexists Cl vector.** Finally, if $\text{rank}(A(1)) = 1, X_t \sim I(1)$ and \exists Cl vector.

In this last interesting case, **we can write $A(1) = h\gamma^T$ (recall that the outer product of two vectors is of rank one) and get a nice interpretation for them.** γ is the unique (up to scale and sign changes) Cl vector, $\gamma^T X_{t-1}$ the error correction term (ECM). h is what drives the adjustment to the steady state.

Remark. *A lack of cointegration does not mean a lack of correlation!*

The most obvious example is (log) exchange rates. Bilateral exchange rates are clear examples of $I(1)$ processes **which do not tend to revert to a long-term mean.** Short-term changes in exchange rates against a common numeraire currency are highly (cross-sectionally) correlated. Nevertheless, the triangular arbitrage relationship would imply that if two exchange rates against the same currency were cointegrated with vector $(1, -1)$, their cross rate would tend to revert to a long-term mean. In the **multivariate case**, the same principle applies, but there could be several linearly independent cointegrating vectors.

Multivariate Case In fact, $\text{rank}[A(1)]$ determines the dimension of the space of cointegrating vectors and

$$A(1) = HPP^{-1}G^T. \quad (3.18)$$

The matrix **$A(1)$ is called the matrix of long-run multipliers.** However, how do we transform cointegrated systems so that they become covariance stationary and strictly invertible? If we simply take the first differences of all the series involved, we end up with a not strictly invertible process. Specifically, the linear combination:

$$\gamma_1 x_{1t} + \gamma_2 x_{2t} \quad (3.19)$$

will not be strictly invertible.

Remark. *Intuitively, by taking the first differences of all the series we lose information about the equilibrium relationship in levels!*

This can be checked by noticing that $\Psi_{\Delta x \Delta x}(1)$ will have less than full rank. In the bivariate case, the best solution is to work with the first difference of one of the series (or a linear combination of the two first differences), together with the cointegrating relationship itself. This will allow us to provide **forecasts that satisfy the long-run relationship.** In the multivariate case, we **can achieve the same goal by working with a basis of the k -dimensional linear space of cointegrating relationships, and the first differences of $N - k$ other variables (or linear combinations of variables).**

3.3 Non-Linear Models

3.3.1 Readons for Using Non-Linear Models

Improving Predictions Note that:

$$\underbrace{\mathbb{E} \{X_t - \mathbb{E}[X_t | X_{t-1}, \dots]\}^2}_{=\text{Non-Linear Prediction MSE}} \leq \underbrace{\mathbb{E} \{X_t - \mathbb{E}^*[X_t | X_{t-1}, \dots]\}^2}_{=\text{Linear Prediction MSE}}. \quad (3.20)$$

Example 3.4. (Chaotic Process) Consider:

$$X_t = aX_{t-1} + c - \left\lfloor \frac{aX_{t-1} + c}{m} \right\rfloor m \quad (3.21)$$

where $\lfloor \cdot \rfloor$ **denotes the integer part**. By construction, X_t is the remainder of dividing $aX_{t-1} + c$ by m , so it will be an integer between 0 and $m - 1$.

If we define:

$$Y_t = \frac{X_t}{m} \in [0, 1) \quad (3.22)$$

we **obtain a pseudo-uniform random number generator**. A common choice of parameters is:

$$a = 84589; \quad (3.23a)$$

$$c = 45989; \text{ and} \quad (3.23b)$$

$$m = 211728. \quad (3.23c)$$

Interest on Features of the Conditional Distribution Other than the Mean Examples include:

$$\underbrace{\mathbb{V}(X_t | X_{t-1}, \dots)}_{=\text{Conditional Variance}}; \quad (3.24a)$$

$$\underbrace{\mathbb{Q}_\tau(X_t | X_{t-1}, \dots)}_{=\text{Value at Risk}}; \text{ and} \quad (3.24b)$$

$$\underbrace{\mathbb{E} \{X_t | X_{t-1}, \dots; X_t \leq \mathbb{Q}_\tau[X_t | X_{t-1}, \dots]\}}_{=\text{Expected Shortfall}} \quad (3.24c)$$

Stochastic Processes with a Restricted Domain Examples include non-negative random variables, such as nominal interest rates or VIX. Alternatively, we can take a look at the variables between 0 and 1 such as the default rates and changing correlations.

Non-Linear Characteristics of X_t For example, suppose X_t is the GDP growth rate, and we are interested in whether the economy is in an expansion $s_t = 1$ or recession $s_t = 0$. However, we do not believe in mechanical rules such as “a recession is characterised by two successive quarters of negative GDP growth”.

3.3.2 TAR & STAR Models

They are mainly used to improve predictions. “T” stands for **thresholds** and “S” for **smoothing**.

Definition 3.11 (TAR(1)). It is given by:

$$X_t = \begin{cases} \mu_1 + \alpha_1 X_{t-1} + u_t, & \text{if } X_{t-1} < \theta; \text{ and} \\ \mu_2 + \alpha_2 X_{t-1} + u_t, & \text{otherwise,} \end{cases} \quad (3.25)$$

with $\theta \in \mathbb{R}$ and:

$$u_t \sim MD(0, \sigma_u^2). \quad (3.26)$$

Notice that WN is not enough, we need u_t to be at least a *Martingale Difference (MD) sequence*. One could have one of the α 's above 1 and still retain covariance stationarity. The **Wold representation, though, will yield worse linear forecasts**. With three regimes, it is a good discrete time model for interest rate dynamics.

Issues The issues include:

1. determining the **number and location of the thresholds**;
2. making sure that the regression (i.e., the conditional mean) **function is continuous** (differentiability achieved with smooth transition version, which “sands down” the rough edges); and
3. the **number of parameters increases fast**.

It can form the basis for a non-parametric procedure.

Remark. *A TAR(q) process could be covariance stationary even if a part of the graph has the slope above $\frac{\pi}{4}$!*

Remark. *TAR(q) models have kinks. This causes issues with inference. The STAR(q) models replace the kink(s) with a smooth curve. The simplest example is the absolute value function:*

$$f(x) = |x|. \quad (3.27)$$

This could be smoothened with:

$$f^s(x) = \frac{x^2}{\sqrt{x^2 + \delta}} \simeq |x| \iff \delta \simeq 0. \quad (3.28)$$

3.3.3 ARCH & SV Models

They are used to **model time-varying conditional variance**.

Definition 3.12 (ARCH(1)). An **ARCH(1) model** is given by:

$$\mathbb{E}(X_t | X_{t-1}, \dots) = 0; \text{ and} \quad (3.29a)$$

$$\mathbb{V}(X_t | X_{t-1}, \dots) = \mathbb{E}(X_t^2 | X_{t-1}, \dots) = \omega + \alpha X_{t-1}^2 = \sigma_t^2. \quad (3.29b)$$

While in a **binary Markov chain**, we have two conditional distributions depending on the lagged value of the chain, here we have a different one for each value of X_{t-1} .

The positivity constraints are:

$$\omega > 0; \text{ and} \quad (3.30a)$$

$$\alpha \geq 0. \quad (3.30b)$$

The **Law of Iterated Expectations** (LIE) immediately implies that:

$$\mathbb{E}(X_t) = 0; \text{ and} \quad (3.31a)$$

$$\mathbb{V}(X_t) = \mathbb{E}(X_t^2) = \frac{\omega}{1 - \alpha} \quad \text{if } \alpha < 1. \quad (3.31b)$$

This is derived from:

$$\sigma_t^2 = \mathbb{E}(X_t^2 | X_{t-1}) = \omega + \alpha X_{t-1}^2 \implies \quad (3.32a)$$

$$X_t^2 = \omega + \alpha X_{t-1}^2 + \eta_t, \text{ where} \quad (3.32b)$$

$$\mathbb{E} \eta_t = 0. \implies \quad (3.32c)$$

$$X_t^2 = \omega + \alpha \sigma_{t-1}^2 + \underbrace{\alpha (X_{t-1}^2 - \sigma_{t-1}^2)}_{=\eta_{t-1}} + \eta_t \implies \quad (3.32d)$$

$$\mathbb{E}(X_t^2 | X_{t-1}) = \mathbb{E}(\omega + \alpha \sigma_{t-1}^2 + \alpha \eta_{t-1} + \eta_t | X_{t-1}) \implies \quad (3.32e)$$

$$\sigma_t^2 = \omega + \alpha \sigma_{t-1}^2 + \alpha \eta_{t-1}. \quad (3.32f)$$

We also **have that the LIE implies that**:

$$\text{cov}(X_t, X_s) = 0. \quad (3.33)$$

However, we have that:

$$\text{cov}(X_t^2, X_s^2) \neq 0. \quad (3.34)$$

Indeed, X^2 **follows an AR(1)**. Further, σ_t^2 **also follows an AR(1)**.

This (higher-order) serial correlation can **capture volatility clusters**; i.e., the **persistent periods of high volatility alternating with persistent periods of low volatility**. For higher-order moments, it is convenient to focus on the **conditionally standardised variable**:

$$X_t^* \equiv \frac{X_t - \mathbb{E}(X_t | X_{t-1}, \dots)}{\sqrt{\mathbb{V}(X_t | X_{t-1}, \dots)}} \quad (3.35)$$

This is different from the unconditionally standardised variable:

$$\frac{X_t - \mathbb{E}(X_t)}{\sqrt{\mathbb{V}(X_t)}} \quad (3.36)$$

If we assume that:

$$\mathbb{E}(X_t^{*4}) = \kappa \quad (3.37)$$

is **bounded**, we can prove that:

$$\mathbb{E}(X_t^4) = \frac{(1 + \alpha)\omega^2\kappa}{(1 - \alpha)(1 - \kappa\alpha^2)}. \quad (3.38)$$

provided that $\alpha^2\kappa < 1$. Since we can re-write this expression as:

$$\mathbb{E}(X_t^4) = \kappa \cdot \mathbb{V}^2(X_t) \cdot \frac{1 - \alpha^2}{1 - \kappa\alpha^2} \quad (3.39)$$

and $\kappa \geq 1$, **the unconditional distribution of X_t will have fatter tails than the conditional distributions**. Intuitively, the reason is that changes in the volatility of X_t lead to a higher variance for X_t^2 . GARCH is like an ARMA for X^2 .

Example 3.5 (Log SV Model). Consider:

$$\overbrace{X_t = X_t^* \cdot \sigma_t}^{\text{=Measurement Eq.}} ; \quad (3.40a)$$

$$\overbrace{\ln \sigma_t^2 = \mu + \alpha \ln \sigma_{t-1}^2 + u_t}^{\text{=Transition Eq.}} ; \text{ and} \quad (3.40b)$$

$$(X_t^*, u_t) \sim i.i.d \quad (3.40c)$$

The **latent variable is “volatility”, but the state-space model is non-linear**. The optimal filter is **numerical** (Markov Chain Monte Carlo or particle filters). A suboptimal “linear” filter can be achieved by taking logs of the square of the observed variables, which yields:

$$\ln X_t^2 = \ln \sigma_t^2 + \ln X_t^{*2} \quad (3.41)$$

In particular, we lose information about the sign of X_t .

3.3.4 Compound Autoregressive Processes (CAR)

A conditionally homoskedastic Gaussian $AR(1)$ model:

$$X_t = \mu + \alpha X_{t-1} + u_t; \text{ and} \quad (3.42a)$$

$$u_t \mid I_{t-1} \sim \mathcal{N}(0, \sigma^2) \quad (3.42b)$$

can be fully **characterised by its conditional cumulant generating function**.

Definition 3.13 (Cumulant Generating Function). The cumulant generating function $\varphi(\lambda)$ is simply the log of the moment generating function $m(\lambda)$:

$$\varphi(\lambda) = \ln m(\lambda). \quad (3.43)$$

Since in the Gaussian case:

$$m_t(\lambda) = \mathbb{E}[\exp(\lambda x_t) \mid I_{t-1}] = \exp\left(\lambda\mu + \lambda\alpha X_{t-1} + \frac{1}{2}\lambda^2\sigma^2\right) \quad (3.44)$$

the **conditional cumulant generating function will be the following affine function**:

$$\varphi(\lambda) = \omega(\lambda) + \rho(\lambda)X_{t-1}; \text{ where} \quad (3.45a)$$

$$\omega(\lambda) = \mu\lambda + .5\sigma^2\lambda^2; \text{ and} \quad (3.45b)$$

$$\rho(\lambda) = \alpha\lambda. \quad (3.45c)$$

- ☑ A nice feature of **this model is that it temporally aggregates, in the sense that the k -period ahead conditional cumulant generating function is of the same form**.
- ☑ This aggregation works both ways, since a Gaussian $AR(1)$ can be **regarded as a discrete-time version of an Ornstein-Uhlenbeck process in continuous time**.
- ☑ **Compound autoregressive processes (CAR) are specified with alternative conditional cumulant generating functions**.

The most popular one is the **Autoregressive Gamma Process (ARG)**.

Definition 3.14 (Autoregressive Gamme Process). The **Autoregressive Gamma Process (ARG)** is given by:

$$\omega(\lambda) = -\delta \ln(1 - \lambda c); \text{ and} \quad (3.46a)$$

$$\rho(\lambda) = \frac{\lambda c}{1 - \lambda c} \beta \text{ such that} \quad (3.46b)$$

$$\left. \frac{2X_t}{c} \right| I_{t-1} \sim \chi_{2\delta}^2(2\beta X_{t-1}). \quad (3.46c)$$

Let:

$$\mathbf{y} \sim N(\boldsymbol{\psi}, \mathbb{I}_N) \implies \quad (3.47a)$$

$$\mathbf{y}^T \mathbf{y} \sim \chi_N^2(\boldsymbol{\psi}^T \boldsymbol{\psi}). \quad (3.47b)$$

Therefore, it is **suitable for positive random variables**. Like in the Gaussian case, the **k -period ahead conditional cumulant generating function is of the same form**. It can also be regarded as a **discrete version of the so-called Feller (or “Square root”) continuous time process**:

$$dX(t) = \kappa X(t)dt + \sqrt{X(t)}dW(t). \quad (3.48)$$

3.3.5 Binary Markov Chains

Definition 3.15 (Binary Markov Chain). A **binary Markov chain in discrete time** is a binary stochastic process that only takes two values, 0 and 1, in which its time-series dependence is limited to a single period.

This means that the conditional distribution of x_t given $x_{t-1}, x_{t-2}, x_{t-3}, \dots$ only depends on the value of x_{t-1} . In fact, **there are only two conditional distributions, one given $x_{t-1} = 0$ and another one given $x_{t-1} = 1$** . Both these distributions are Bernoulli, but in general, the parameter is different. These probabilities are usually **described using the following transition matrix between states**:

$$\begin{pmatrix} p_t & 1 - q_t \\ 1 - p_t & q_t \end{pmatrix} \quad (3.49)$$

where

$$p_t \equiv \mathbb{P}(x_t = 0 \mid x_{t-1} = 0); \text{ and} \quad (3.50a)$$

$$q_t \equiv \mathbb{P}(x_t = 1 \mid x_{t-1} = 1). \quad (3.50b)$$

Theorem 3.1 (Perron-Frobenius Theorem). *The eigenvalues of the **transition matrix** are 1 and $p_t + q_t - 1$.*

Definition 3.16 (Homogeneous Binary MC/ Time-Invariant MC). When $p_t = p$ and $q_t = q \forall t$, the chain is said to be **homogenous or time-invariant**.

Definition 3.17 (Reducible/ Irreducible MC). If p and q are strictly less than 1, the chain is **irreducible**, while if any of them is 1, then it is **reducible**.

Definition 3.18 (Transient & Absorbing States). If $p = 1$ then the second state is called **transient** while the first state is called **absorbing** because once the chain reaches it, it is not possible to leave.

Theorem 3.2 (Stationarity of Binary MC). *Given a suitable initial condition, a **homogeneous binary Markov chain will be strictly stationary** if it is irreducible and the second eigenvalue of its transition matrix is less than 1 in absolute value.*

In that case, the unconditional probability that it is 1 will be:

$$\frac{1 - q}{2 - p - q} \quad (3.51)$$

In contrast, **periodic chains, with $p = q = 0$, are not stationary**. Finally, it is interesting to note that as far as their autocorrelation structure is concerned, homogeneous binary Markov chains are a special case of an AR(1) in which the autoregressive parameter is precisely $p + q - 1$, the second eigenvalue of its transition matrix. This explains **why the (asymptotic) stationarity condition is $|p + q - 1| < 1$** .

Definition 3.19 (Geometric Distribution). The **geometric distribution** is the probability distribution of the number of Bernoulli trials needed to get one success, supported on the set $1, 2, \dots$

There is a very interesting and useful property linking binary Markov chains and geometric distributions. The **time the process remains in state 0 is geometrically distributed with parameter $(1 - p)$** . In turn, the time the process remains in state 1 is also geometrically distributed, but this time with parameter $(1 - q)$. This **property is very convenient for defining continuous-time Markov chains**.

Specifically, given that **the exponential distribution is the continuous analogue of the geometric distribution**, we can define binary Markov chains in continuous time as a binary process in which **the times between states are exponentially rather than geometrically distributed**.

If X is an exponentially distributed random variable with parameter λ , then $Y = \lfloor X \rfloor$, where $\lfloor X \rfloor$ denotes the floor (or greater integer) function, is geometrically distributed with parameter:

$$p = 1 - e^{-\lambda} \quad (3.52)$$

so that

$$\lambda = -\ln(1 - p) \quad (3.53)$$

although in this case the geometric distribution is defined over $0, 1, 2, \dots$. The relationship works both ways. If X is geometrically distributed with parameter $p = \lambda/n$, then as $n \rightarrow \infty$ the distribution of X/n approaches an exponential distribution with rate λ , so that

$$\mathbb{P}\left(\frac{X}{n} > x\right) = \mathbb{P}(X > nx) \implies \quad (3.54a)$$

$$\mathbb{P}\left(\frac{X}{n} > x\right) = \lim_{n \rightarrow \infty} \left(1 - \frac{p}{n}\right)^x \implies \quad (3.54b)$$

$$\mathbb{P}\left(\frac{X}{n} > x\right) = e^{-\lambda x}. \quad (3.54c)$$

3.3.6 k-ary Markov chains

K-ary stands for binary, ternary, quaternary, etc. The crucial ingredient is the transition matrix P . This a matrix of non-negative elements in which the elements of every column add up to 1. (many textbooks define P^T as the transition matrix). As a result, **Perron-Frobenius's theorem implies that it has at least one eigenvalue equal to 1**.

Definition 3.20 (Stationary Distribution of K-Ary MC). The **stationary distribution** is the vector π such that:

$$P\pi = \pi. \quad (3.55)$$

The vector π is such that:

$$\pi^T \ell_k = 1. \quad (3.56)$$

However, **this may not be unique**.

Theorem 3.3 (Uniqueness of Stationary Distribution). *If the chain is irreducible and aperiodic, then π is unique.*

Periodic **Markov chains are those irreducible ones in which there is more than one eigenvalue of P on the unit circle**.

3.3.7 Absorbing Markov Chains

Definition 3.21 (Absorbing/ Reducible Markov Chain). A Markov chain is an **absorbing** (or reducible) chain if:

1. there is at least one absorbing state; and
2. it is possible to go from any state to at least one absorbing state in a finite number of steps.

In an **absorbing Markov chain**, a state that is not absorbing is called **transient**. Let an absorbing Markov chain with transition matrix P have t transient states and r absorbing states, with $t + r = k$. Then we can always write its transition matrix P in the following canonical form:

$$P = \begin{pmatrix} Q & 0 \\ R & \mathbb{I}_r \end{pmatrix}, \quad (3.57)$$

where Q is a $t \times t$ matrix, R is a nonzero $r \times t$ matrix, 0 is an $t \times r$ zero matrix, and \mathbb{I}_r is the identity matrix of order r .

Thus, Q describes the probability of transitioning from some transient state to another while R describes the probability of transitioning from some transient state to an absorbing one. The expected number of visits to a transient state starting from a transient state is given by:

$$\mathbb{I}_t + Q + Q^2 + Q^3 + \dots = (\mathbb{I}_t - Q)^{-1}. \quad (3.58)$$

The expected number of steps before being absorbed is:

$$(\mathbb{I}_t - Q)^{-1} \ell_t, \quad (3.59)$$

where ℓ_t is a vector of t ones. The probability of eventually being absorbed in the absorbing state i when starting from transient state j is given by the (j, i) -entry of the matrix $R(\mathbb{I}_t - Q)^{-1}$.

Example 3.6 (Absorbing Markov Chain). You have been hired to manage the portfolio of a rich client. The client says that he will move his portfolio to a different firm after two consecutive years of underperformance relative to a benchmark. Naturally, you will lose your job when he does. For simplicity, **assume that your outperformance and underperformance in any given year is completely random, with probability $\frac{1}{2}$.**

We can represent the problem by employing a ternary Markov chain. The states are 0 consecutive years of underperformance, 1 consecutive year of underperformance, and 2 consecutive years of underperformance. The **transition probability matrix** is:

$$\begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{pmatrix} \quad (3.60)$$

Therefore, the expected number of years you will remain employed will be 6 since you are hired, but only 4 after experiencing an underperforming year. Unfortunately, the probability that you will eventually be fired is 1 regardless of whether you have overperformed or underperformed in the current year. This is a simple version of the claim that a monkey typing randomly will eventually write this sentence with probability one.

Example 3.7 (VAR(1) and Markov Chains). Let ξ_t denote a $k \times 1$ **multinomial random variable whose j^{th} element is equal to 1 if the chain is at state j in period t and 0 otherwise.** The conditional expectation of ξ_t given that the chain was at state i at period $t - 1$ is given by the i^{th} column of the matrix P . Consequently:

$$\mathbb{E}(\xi_t \mid \xi_{t-1}) = P\xi_{t-1} \quad (3.61)$$

Hence, we can then write:

$$\xi_t = P\xi_{t-1} + v_t, \quad (3.62)$$

where v_t is a vector martingale difference sequence. Therefore:

$$\mathbb{E}(\xi_{t+k} \mid \xi_{t-1}) = P^k \xi_{t-1}. \quad (3.63)$$

The VAR(1) representation of the chain **seems odd because P contains a unit root while an ergodic chain is strictly stationary.** However, this is only a reflection of the fact that we have included all k states in ξ_t , which implies that the covariance matrix of v_t is singular, but in a way that annihilates the unit root in P . In the binary case, we avoided this problem by getting rid of one of the two elements of ξ_t , which introduced a drift in the AR(1) representation of the chain.

Definition 3.22 (Embeddable Markov Chain). **Embeddable Markov chains are those for which there is a real intensity matrix Q such that:**

$$P = \exp(Q) \quad (3.64)$$

Given that the eigenvalues of Q are the natural logs of the eigenvalues of P , we effectively need the latter to be strictly positive to avoid complex matrices, and thereby achieve embeddability. **Q , however, might not be unique. The problem arises when there are multiple eigenvalues and their arithmetic and geometric multiplicities do not coincide. In other words, when P cannot be diagonalised.**

3.3.8 Regime switching (RS) Models

Suppose we observe y_t but there is a latent variable -which follows a first-order binary Markov chain- that drives the dynamic properties of the series:

$$y_t \mid s_t = 0, y_{t-1}, y_{t-2}, \dots \sim \mathcal{N}(\mu_0 + \alpha_0 y_{t-1}, \sigma_0^2) \quad (3.65a)$$

$$y_t \mid s_t = 1, y_{t-1}, y_{t-2}, \dots \sim \mathcal{N}(\mu_1 + \alpha_1 y_{t-1}, \sigma_1^2) \quad (3.65b)$$

These can be regarded as **measurement equations**. The **transition equation is characterised by the transition matrix of the binary Markov chain**.

Hamilton Filter We can obtain the **forecasting equations**:

$$\begin{aligned} \mathbb{P}(s_t = 1 \mid y_{t-1}, \dots) &= \mathbb{P}(s_t = 1 \mid s_{t-1} = 0, y_{t-1}, \dots) \mathbb{P}(s_{t-1} = 0 \mid y_{t-1}, \dots) \\ &\quad + \mathbb{P}(s_t = 1 \mid s_{t-1} = 1, y_{t-1}, \dots) \mathbb{P}(s_{t-1} = 1 \mid y_{t-1}, \dots) \end{aligned} \quad (3.66)$$

where

$$\mathbb{P}(s_t = 1 \mid s_{t-1} = 0, y_{t-1}, \dots) \equiv 1 - p; \text{ and} \quad (3.67a)$$

$$\mathbb{P}(s_t = 1 \mid s_{t-1} = 1, y_{t-1}, \dots) \equiv q. \quad (3.67b)$$

By the same token:

$$\begin{aligned} g(y_t \mid y_{t-1}, \dots) &= g(y_t \mid s_t = 0, y_{t-1}, \dots) \mathbb{P}(s_t = 0 \mid y_{t-1}, \dots) \\ &\quad + g(y_t \mid s_t = 1, y_{t-1}, \dots) \mathbb{P}(s_t = 1 \mid y_{t-1}, \dots), \end{aligned} \quad (3.68)$$

where $g(y_t \mid s_t = 1, y_{t-1}, \dots)$ is the **conditional pdf of y_t given $s_t = 1, y_{t-1}, \dots$** . Using Bayes' Theorem, we get the **updating equations**:

$$\mathbb{P}(s_t = 1 \mid y_t, y_{t-1}, \dots) = \frac{g(y_t \mid s_t = 1, y_{t-1}, \dots) \mathbb{P}(s_t = 1 \mid y_{t-1}, \dots)}{\sum_{i \in \{0,1\}} g(y_t \mid s_t = i, y_{t-1}, \dots) \mathbb{P}(s_t = i \mid y_{t-1}, \dots)} \implies \quad (3.69a)$$

$$\mathbb{P}(s_t = 1 \mid y_t, y_{t-1}, \dots) = \frac{g(y_t \mid s_t = 1, y_{t-1}, \dots) \mathbb{P}(s_t = 1 \mid y_{t-1}, \dots)}{g(y_t \mid y_{t-1}, \dots)}. \quad (3.69b)$$

As usual, an assumption about pre-sample observations has to be made to initialise the filter. There are also smoothers (Hamilton's book describes an algorithm due to Kim (1993)).

4 Inference with Dependent Observations

4.1 DGPs, Models and Parameters

4.1.1 Introduction

It is very important to distinguish between:

1. **DGP (data generating process)**: true model, ρ denotes the parameters; and
2. **estimated model**: the model whose parameters θ you will estimate.

By the analogy principle, use $\frac{1}{T} \sum X_t$ to estimate $\mathbb{E} X$ and $\frac{1}{T} \sum X_t^2$ to estimate $\mathbb{E}(X^2)$.

Definition 4.1 (Extremum Estimators). The **extremum estimators** is as follows.^{4.1} Suppose you have a criterion function:

$$Q_T(\theta) = Q(\theta \mid y_1, \dots, y_T) \quad (4.1)$$

and would like to find

$$\hat{\theta}_T = \arg_{\theta \in \Theta} \{\min Q_T(\theta)\}. \quad (4.2)$$

^{4.1} It's called the "extremum" estimator because we optimise over some set and look for the minimum value.

This includes MLE, GMM, minimum distance, etc. **since we can always transform a maximisation by changing the sign of the criterion function.** We assume that θ is in Θ (compact). The purpose of this section is to **figure out what exactly $\hat{\theta}_T$ estimates and to learn about its sampling properties.**

In that regard, it is important to **distinguish between the numerical properties of the estimators $\hat{\theta}_T$ and their statistical properties.** The numerical properties have to do with things like the numerically efficient and reliable evaluation of the criterion function, its score and Hessian, as well as with the first and second-order conditions of the optimisation programme for a given sample. The **statistical properties have to do with the sampling distribution of the estimators across different samples.** Since $Q_T(\theta)$ **varies from sample to sample for any fixed value of θ , it is convenient to look first at what would happen in very large samples.** Although $\hat{\theta}_T$ is an extremum estimator, looking at the limiting case of infinite samples will allow us to make use of the analogy principle to re-interpret it.

Limiting Objective Function It may sound surprising at first, but $Q_T(\theta)$ can be regarded as a multidimensional stochastic process indexed by the parameters $\theta \in \Theta$. The fact that it is a stochastic process is pretty clear once we realise that:

▣ $Q_T(\theta)$ is a random variable for a fixed value of θ ; and

▣ $Q_T(\theta')$ and $Q_T(\theta'')$ will be correlated, and the same is true for any sequence of parameter values.

This type of stochastic process is **called an empirical process** because its variability arises from the variability of the underlying sample. As such we can try to define the limit of the sequence of random functions $Q_T(\cdot)$. For a fixed value of θ , ordinary limits in probability would suffice. Specifically, we could define the limiting value of the objective function at $\theta = \theta'$ as

$$Q(\theta') = p \lim_{T \rightarrow \infty} Q_T(\theta'), \quad (4.3)$$

which will often be non-stochastic. However, **pointwise convergence may not be enough. We will need a stronger concept such as uniform convergence.** i.e.

Definition 4.2 (Uniform Convergence). $Q_T(\theta)$ **uniformly converges** to $Q(\theta)$ if:

$$p \lim_{T \rightarrow \infty} \left\{ \sup_{\theta \in \Theta} |Q_T(\theta) - Q(\theta)| \right\} = 0. \quad (4.4)$$

Assuming the limiting function is well defined, we can think of:

$$\theta_\infty = \arg_{\theta \in \Theta} \{ \min Q(\theta) \} \quad (4.5)$$

as the **“pseudo-true” value of the parameters.** But first, we need to be able to compute $Q_T(\theta)$. Let’s consider some likelihood examples.

Example 4.1 (Random Sample of Bernoulli RV). Let:

$$\mathbb{P}(u = 1) = \lambda. \quad (4.6)$$

We are interested in finding λ , the parameter of interest. We have a **random sample of size T :**

$$u_t \sim \text{i.i.d. Bernoulli}(\lambda) \quad \forall t. \quad (4.7)$$

The contribution of observation t to log-likelihood is:

$$l_t(\lambda) = \ln p(u_t | \lambda) = (1 - u_t) \ln(1 - \lambda) + u_t \ln(\lambda) \quad (4.8)$$

The **(average) log-likelihood function**:

$$\bar{l}_T(\lambda) = \left(1 - \frac{1}{T} \sum_{t=1}^T u_t\right) \ln(1 - \lambda) + \left(\frac{1}{T} \sum_{t=1}^T u_t\right) \ln \lambda. \quad (4.9)$$

The **sufficient statistic** is:

$$\frac{1}{T} \sum_{t=1}^T u_t. \quad (4.10)$$

The **contribution of observation t to log-likelihood score** is:

$$s_t(\lambda) = \frac{u_t - \lambda}{\lambda(1 - \lambda)}. \quad (4.11)$$

The **(average) log-likelihood score vector** is:

$$\bar{s}_T(\lambda) = \frac{\left[\left(\frac{1}{T} \sum_{t=1}^T u_t\right) - \lambda\right]}{\lambda(1 - \lambda)} \quad (4.12)$$

The **maximum likelihood estimator** is:

$$\hat{\lambda}_T = \frac{1}{T} \sum_{t=1}^T u_t \quad (4.13)$$

The **contribution of observation t to log-likelihood Hessian** is:

$$h_t(\lambda) = \frac{-1}{\lambda(1 - \lambda)} + \frac{2(u_t - \lambda)\left(\frac{1}{2} - \lambda\right)}{\lambda^2(1 - \lambda)^2}. \quad (4.14)$$

The (average) Hessian matrix is:

$$\bar{h}_T(\lambda) = \frac{-1}{\lambda(1 - \lambda)} + \frac{2\left(\hat{\lambda}_T - \lambda\right)\left(\frac{1}{2} - \lambda\right)}{\lambda^2(1 - \lambda)^2} \Rightarrow \quad (4.15a)$$

$$\bar{h}_T(\lambda) = \frac{3\lambda - 1 - \hat{\lambda}_T + \sqrt{1 - \hat{\lambda}_T(1 - \hat{\lambda}_T)}}{3\lambda(1 - \lambda)} \times \frac{3\lambda - 1 - \hat{\lambda}_T - \sqrt{1 - \hat{\lambda}_T(1 - \hat{\lambda}_T)}}{3\lambda(1 - \lambda)} \leq 0 \Rightarrow \quad (4.15b)$$

$$\bar{h}_T(\hat{\lambda}_T) = \frac{-1}{\hat{\lambda}_T(1 - \hat{\lambda}_T)} < 0 \Rightarrow \quad (4.15c)$$

$$\hat{\lambda}_T = \arg \max_{\lambda} \bar{l}_T(\lambda). \quad (4.15d)$$

Example 4.2 (Normal Distribution). We have

$$x \sim \mathcal{N}(\nu, \omega^2) \quad (4.16)$$

This means:

$$\mathbb{P}(x \leq c \mid \nu, \omega^2) = \int_{-\infty}^c \frac{1}{\sqrt{2\pi}\omega} \exp\left[-\frac{1}{2} \frac{(x - \nu)^2}{\omega^2}\right] dx, \quad (4.17)$$

with:

$$\mathbb{P}(x \mid \nu, \omega^2) = (2\pi)^{-1/2} \omega^{-1} \exp\left[-\frac{1}{2} \frac{(x - \nu)^2}{\omega^2}\right]. \quad (4.18)$$

The **parameters of interest** are $\theta = (\nu, \omega^2)$. We consider a random sample of size T :

$$x_t \sim i.i.d. \mathcal{N}(\nu, \omega^2) \quad \forall t. \quad (4.19)$$

	u_1	u_2	u_3	u_4	likelihood	ML estimator
0	0	0	0	0	$(1 - \lambda)^4$	$\hat{\lambda}_4 = 0$
1	0	0	0	1	$(1 - \lambda)^3 \lambda$	$\hat{\lambda}_4 = 1/4$
2	0	0	1	0	$(1 - \lambda)^3 \lambda$	$\hat{\lambda}_4 = 1/4$
3	0	0	1	1	$(1 - \lambda)^2 \lambda^2$	$\hat{\lambda}_4 = 1/2$
4	0	1	0	0	$(1 - \lambda)^3 \lambda$	$\hat{\lambda}_4 = 1/4$
5	0	1	0	1	$(1 - \lambda)^2 \lambda^2$	$\hat{\lambda}_4 = 1/2$
6	0	1	1	0	$(1 - \lambda)^2 \lambda^2$	$\hat{\lambda}_4 = 1/2$
7	0	1	1	1	$(1 - \lambda) \lambda^3$	$\hat{\lambda}_4 = 3/4$
8	1	0	0	0	$(1 - \lambda)^3 \lambda$	$\hat{\lambda}_4 = 1/4$
9	1	0	0	1	$(1 - \lambda)^2 \lambda^2$	$\hat{\lambda}_4 = 1/2$
10	1	0	1	0	$(1 - \lambda)^2 \lambda^2$	$\hat{\lambda}_4 = 1/2$
11	1	0	1	1	$(1 - \lambda) \lambda^3$	$\hat{\lambda}_4 = 3/4$
12	1	1	0	0	$(1 - \lambda)^2 \lambda^2$	$\hat{\lambda}_4 = 1/2$
13	1	1	0	1	$(1 - \lambda) \lambda^3$	$\hat{\lambda}_4 = 3/4$
14	1	1	1	0	$(1 - \lambda) \lambda^3$	$\hat{\lambda}_4 = 3/4$
15	1	1	1	1	λ^4	$\hat{\lambda}_4 = 1$

Table 4.1: Bernoulli Random Sample $T = 4$.

The **contribution of observation t to log-likelihood** is:

$$l_t(\theta) = \ln p(x_t | \theta) = -\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2(\theta)}{2} - \frac{\varepsilon_t^{*2}(\theta)}{2}; \text{ where} \quad (4.20a)$$

$$\varepsilon_t^*(\theta) = \frac{x_t - \mu(\theta)}{\sigma(\theta)} = \frac{x_t - \nu}{\omega}. \quad (4.20b)$$

The (average) log-likelihood function:

$$\bar{l}_T(\theta) = -\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2(\theta)}{2} - \frac{1}{2T} \sum_{t=1}^T \varepsilon_t^{*2}(\theta). \quad (4.21)$$

The sufficient statistics are:

$$\frac{1}{T} \sum_{t=1}^T x_t; \text{ and} \quad (4.22a)$$

$$\frac{1}{T} \sum_{t=1}^T x_t^2 \quad (4.22b)$$

The **contribution of observation t to score** is:

$$s_t(\theta) = \frac{\varepsilon_t^*(\theta)}{\sigma(\theta)} \frac{\partial \mu(\theta)}{\partial \theta} + \frac{[\varepsilon_t^{*2}(\theta) - 1]}{2\sigma^2(\theta)} \frac{\partial \sigma^2(\theta)}{\partial \theta}; \text{ where} \quad (4.23a)$$

$$\frac{\partial \mu(\theta)}{\partial \theta} = (1, 0)^T; \quad (4.23b)$$

$$\frac{\partial \sigma^2(\theta)}{\partial \theta} = (0, 1)^T; \quad (4.23c)$$

$$s_{\nu t}(\theta) = \frac{\varepsilon_t^*(\theta)}{\omega}; \text{ and} \quad (4.23d)$$

$$s_{\omega^2 t}(\theta) = \frac{\varepsilon_t^{*2}(\theta) - 1}{2\omega^2}. \quad (4.23e)$$

The **(average) score** is:

$$\bar{s}_T(\theta) = \frac{1}{T} \sum_{t=1}^T \frac{\varepsilon_t^*(\theta)}{\sigma(\theta)} \frac{\partial \mu(\theta)}{\partial \theta} + \frac{1}{2T} \sum_{t=1}^T \frac{[\varepsilon_t^{*2}(\theta) - 1]}{\sigma^2(\theta)} \frac{\partial \sigma^2(\theta)}{\partial \theta} \quad (4.24)$$

The maximum likelihood estimators are:

$$\hat{\nu}_T = \frac{1}{T} \sum_{t=1}^T x_t; \text{ and} \quad (4.25a)$$

$$\hat{\omega}_T^2 = \frac{1}{T} \sum_{t=1}^T (x_t - \hat{\nu}_T)^2 = \frac{1}{T} \sum_{t=1}^T x_t^2 - \left(\frac{1}{T} \sum_{t=1}^T x_t \right)^2. \quad (4.25b)$$

The **contribution of observation t to Hessian** is:

$$\begin{aligned} h_t(\theta) = & \frac{\varepsilon_t^*(\theta)}{\sigma(\theta)} \frac{\partial^2 \mu(\theta)}{\partial \theta \partial \theta^T} + \frac{[\varepsilon_t^{*2}(\theta) - 1]}{2\sigma^2(\theta)} \frac{\partial^2 \sigma^2(\theta)}{\partial \theta \partial \theta^T} \\ & - \frac{\varepsilon_t^*(\theta)}{\sigma^3(\theta)} \left[\frac{\partial \mu(\theta)}{\partial \theta} \frac{\partial \sigma^2(\theta)}{\partial \theta^T} + \frac{\partial \sigma^2(\theta)}{\partial \theta} \frac{\partial \mu(\theta)}{\partial \theta^T} \right] \\ & - \frac{1}{\sigma^2(\theta)} \frac{\partial \mu(\theta)}{\partial \theta} \frac{\partial \mu(\theta)}{\partial \theta^T} - \frac{[2\varepsilon_t^{*2}(\theta) - 1]}{2\sigma^4(\theta)} \frac{\partial \sigma^2(\theta)}{\partial \theta} \frac{\partial \sigma^2(\theta)}{\partial \theta^T}. \end{aligned} \quad (4.26)$$

The **(average) Hessian matrix** is $\bar{h}_T(\theta)$, which is indefinite. It is:

$$\bar{h}_T(\hat{\theta}_T) = \underbrace{\frac{-1}{\sigma^2(\hat{\theta}_T)} \frac{\partial \mu(\hat{\theta}_T)}{\partial \theta} \frac{\partial \mu(\hat{\theta}_T)}{\partial \theta^T} - \frac{1}{2\sigma^4(\hat{\theta}_T)} \frac{\partial \sigma^2(\hat{\theta}_T)}{\partial \theta} \frac{\partial \sigma^2(\hat{\theta}_T)}{\partial \theta^T}}_{\text{Negative Definite}} \implies \quad (4.27a)$$

$$\hat{\theta}_T = \arg \max_{\theta} \bar{l}_T(\theta). \quad (4.27b)$$

Example 4.3 (Homogeneous Binary Markov Chain). If $t = 1$, we have:

$$P(u_t = 1) = v; \text{ and} \quad (4.28a)$$

$$P(u_t = 0) = 1 - v. \quad (4.28b)$$

Further, if $t > 1$:

$$\mathbb{P}(u_t = 1 \mid u_{t-1} = 1, u_{t-2}, \dots, u_1) = q; \quad (4.29a)$$

$$\mathbb{P}(u_t = 0 \mid u_{t-1} = 1, u_{t-2}, \dots, u_1) = 1 - q; \quad (4.29b)$$

$$\mathbb{P}(u_t = 1 \mid u_{t-1} = 0, u_{t-2}, \dots, u_1) = 1 - p; \text{ and} \quad (4.29c)$$

$$\mathbb{P}(u_t = 0 \mid u_{t-1} = 0, u_{t-2}, \dots, u_1) = p. \quad (4.29d)$$

The **parameters of interest** are $\theta = (p, q)^T$. To compute the joint log-likelihood function we make use of the “prediction error decomposition”, which sequentially factorises the joint log-likelihood of the sample u_1, \dots, u_T starting from the end. The **contribution of observation $t \geq 2$ to log-likelihood** is:

$$l_t(p, q) = \ln p(u_t \mid u_{t-1}, \dots, u_1; p, q) \implies \quad (4.30a)$$

$$l_t(p, q) = (1 - u_{t-1}) [(1 - u_t) \ln p + u_t \ln(1 - p)] + u_{t-1} [(1 - u_t) \ln(1 - q) + u_t \ln q] \quad (4.30b)$$

The **contribution of observation $t = 1$ to log-likelihood**:

$$l_1(v) = (1 - u_1) \ln(1 - v) + u_1 \ln v. \quad (4.31)$$

The sufficient statistics are:

$$\frac{1}{T} \sum_{t=2}^T u_t; \quad (4.32a)$$

$$\frac{1}{T} \sum_{t=2}^T u_{t-1}; \text{ and} \quad (4.32b)$$

$$\frac{1}{T} \sum_{t=2}^T u_t u_{t-1}. \quad (4.32c)$$

The **contribution of observation t to log-likelihood score** is:

$$s_t(p, q) = \frac{(1 - u_t - p)(1 - u_{t-1})}{p(1 - p)} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \frac{(u_t - q)u_{t-1}}{q(1 - q)} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (4.33)$$

The **maximum likelihood estimators** are:

$$\hat{v}_T = u_1; \quad (4.34a)$$

$$\hat{p}_T = \frac{\sum_{t=2}^T (1 - u_t)(1 - u_{t-1})}{\sum_{t=2}^T (1 - u_{t-1})}; \text{ and} \quad (4.34b)$$

$$\hat{q}_T = \frac{\sum_{t=2}^T u_t u_{t-1}}{\sum_{t=2}^T u_{t-1}}. \quad (4.34c)$$

Estimators with better finite sample properties but no closed-form can be obtained if we assume stationarity (i.e. $v = \lambda$)

Example 4.4 (Gaussian AR(1)). Set:

$$x_1 \sim \mathcal{N}(\iota, \varsigma^2); \text{ and} \quad (4.35a)$$

$$x_t \mid x_{t-1}, \dots, x_1 \sim \mathcal{N}(\beta x_{t-1}, \varphi^2). \quad (4.35b)$$

The **parameters of interest are $\theta = (\beta, \varphi^2)$** . Again, we will make use of the prediction error decomposition to compute the log-likelihood function. The contribution of observation t to log-likelihood:

$$l_t(\theta) = \ln p(x_t \mid x_{t-1}, \dots, x_1; \theta) \implies \quad (4.36a)$$

$$l_t(\theta) = -\frac{\ln 2\pi}{2} - \frac{\ln \sigma_t^2(\theta)}{2} - \frac{\varepsilon_t^{*2}(\theta)}{2}, \text{ where} \quad (4.36b)$$

$$\varepsilon_t^*(\theta) = \frac{x_t - \mu_t(\theta)}{\sigma_t(\theta)} = \frac{x_t - \beta x_{t-1}}{\varphi}. \quad (4.36c)$$

The **(average) log-likelihood function (ignoring initial conditions)** is:

$$\bar{l}_T(\theta) = -\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2(\theta)}{2} - \frac{1}{2T} \sum_{t=2}^T \varepsilon_t^{*2}(\theta). \quad (4.37)$$

The sufficient statistics are:

$$\frac{1}{T} \sum_{t=2}^T x_t^2; \quad (4.38a)$$

$$\frac{1}{T} \sum_{t=2}^T x_{t-1}^2; \text{ and} \quad (4.38b)$$

$$\frac{1}{T} \sum_{t=1}^T x_t x_{t-1}. \quad (4.38c)$$

The **contributions of observation t to score** are:

$$s_t(\theta) = \frac{\varepsilon_t^*(\theta)}{\sigma_t(\theta)} \frac{\partial \mu_t(\theta)}{\partial \theta} + \frac{[\varepsilon_t^{*2}(\theta) - 1]}{2\sigma_t^2(\theta)} \frac{\partial \sigma_t^2(\theta)}{\partial \theta}; \quad (4.39a)$$

$$\frac{\partial \mu_t(\theta)}{\partial \theta} = (x_{t-1}, 0)^T; \quad (4.39b)$$

$$\frac{\partial \sigma_t^2(\theta)}{\partial \theta} = (0, 1)^T; \quad (4.39c)$$

$$s_{\beta t}(\theta) = \frac{\varepsilon_t^*(\theta)x_{t-1}}{\varphi}; \text{ and} \quad (4.39d)$$

$$s_{\varphi^2 t}(\theta) = \frac{\varepsilon_t^{*2}(\theta) - 1}{2\varphi^2}. \quad (4.39e)$$

The **average score**:

$$\bar{s}_T(\theta) = \frac{1}{T-1} \sum_{t=2}^T \frac{\varepsilon_t^*(\theta)}{\sigma_t(\theta)} \frac{\partial \mu_t(\theta)}{\partial \theta} + \frac{1}{2(T-1)} \sum_{t=2}^T \frac{[\varepsilon_t^{*2}(\theta) - 1]}{\sigma_t^2(\theta)} \frac{\partial \sigma_t^2(\theta)}{\partial \theta} \quad (4.40)$$

The maximum likelihood estimators are:

$$\hat{\beta}_T = \frac{\sum_{t=2}^T x_t x_{t-1}}{\sum_{t=2}^T x_{t-1}^2}; \text{ and} \quad (4.41a)$$

$$\hat{\varphi}_T^2 = \frac{1}{T-1} \sum_{t=2}^T (x_t - \hat{\beta}_T x_{t-1})^2. \quad (4.41b)$$

The **contribution of observation t to Hessian** is:

$$\begin{aligned} h_t(\theta) = & \frac{\varepsilon_t^*(\theta)}{\sigma_t(\theta)} \frac{\partial^2 \mu_t(\theta)}{\partial \theta \partial \theta^T} + \frac{[\varepsilon_t^{*2}(\theta) - 1]}{2\sigma_t^2(\theta)} \frac{\partial^2 \sigma_t^2(\theta)}{\partial \theta \partial \theta^T} - \frac{\varepsilon_t^*(\theta)}{\sigma_t^3(\theta)} \left[\frac{\partial \mu_t(\theta)}{\partial \theta} \frac{\partial \sigma_t^2(\theta)}{\partial \theta^T} + \frac{\partial \sigma_t^2(\theta)}{\partial \theta} \frac{\partial \mu_t(\theta)}{\partial \theta^T} \right] \\ & - \frac{1}{\sigma_t^2(\theta)} \frac{\partial \mu_t(\theta)}{\partial \theta} \frac{\partial \mu_t(\theta)}{\partial \theta^T} - \frac{[2\varepsilon_t^{*2}(\theta) - 1]}{2\sigma_t^4(\theta)} \frac{\partial \sigma_t^2(\theta)}{\partial \theta} \frac{\partial \sigma_t^2(\theta)}{\partial \theta^T} \end{aligned} \quad (4.42)$$

The **average Hessian matrix, $\bar{h}_T(\theta)$, is indefinite**:

$$\bar{h}_T(\hat{\theta}_T) = \underbrace{-\frac{1}{T} \sum_{t=2}^T \frac{1}{\sigma_t^2(\hat{\theta}_T)} \frac{\partial \mu_t(\hat{\theta}_T)}{\partial \theta} \frac{\partial \mu_t(\hat{\theta}_T)}{\partial \theta^T} - \frac{1}{T} \sum_{t=2}^T \frac{1}{2\sigma_t^4(\hat{\theta}_T)} \frac{\partial \sigma_t^2(\hat{\theta}_T)}{\partial \theta} \frac{\partial \sigma_t^2(\hat{\theta}_T)}{\partial \theta^T}}_{\text{Negative Definite}} \Rightarrow \quad (4.43a)$$

$$\hat{\theta}_T = \arg \max_{\theta} \bar{l}_T(\theta). \quad (4.43b)$$

Estimators with better finite sample properties but no closed-form can be obtained if we assume stationarity (i.e., $\iota = 0, \varsigma^2 = \varphi^2 / (1 - \beta^2)$).

4.2 Statistical Properties

Let us now turn to the limiting objective function, its score and Hessian. For simplicity, let us consider the **i.i.d Gaussian log-likelihood as an example**. We saw before that the contribution of observation t to log-likelihood is:

$$l_t(\theta) = \ln p(x_t | \theta) = -\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2(\theta)}{2} - \frac{\varepsilon_t^{*2}(\theta)}{2}, \text{ where} \quad (4.44a)$$

$$\varepsilon_t^*(\theta) = \frac{x_t - \mu(\theta)}{\sigma(\theta)} = \frac{x - \nu}{\omega} \quad (4.44b)$$

The **(average) log-likelihood function** is:

$$\bar{l}_T(\theta) = -\frac{\ln 2\pi}{2} - \frac{\ln \sigma^2(\theta)}{2} - \frac{1}{2T} \sum_{t=1}^T \varepsilon_t^{*2}(\theta) \quad (4.45)$$

To find $Q(\theta)$, **we need to find the p lim of this expression for fixed values of the parameters in θ .**

However, **we must do so without necessarily assuming that the estimated model is correct.** This effectively requires us to obtain the p lim of the sufficient statistics:

$$\frac{1}{T} \sum_{t=1}^T x_t; \text{ and} \quad (4.46a)$$

$$\frac{1}{T} \sum_{t=1}^T x_t^2. \quad (4.46b)$$

Therefore, the limiting criterion function $Q(\theta)$ will be well defined if the **true DGP for x_t is such that a Law of Large numbers can be applied to yield.**

$$p \lim_T \frac{1}{T} \sum_{t=1}^T x_t = \mathbb{E}(x_t) \text{ and} \quad (4.47a)$$

$$p \lim_T \frac{1}{T} \sum_{t=1}^T x_t^2 = \mathbb{E}(x_t^2). \quad (4.47b)$$

Let us **define**

$$\rho_1 = \mathbb{E}(x_t); \text{ and} \quad (4.48a)$$

$$\rho_2 = \mathbb{E}(x_t^2) \quad (4.48b)$$

as parameters of the DGP.

There may be many other parameters, **but ρ_1 and ρ_2 are the only ones that matter for this criterion function.** Assuming suitable LLNs can be applied, the **limiting objective function will be:**

$$Q(\theta; \rho) = \frac{\ln 2\pi}{2} + \frac{\ln \omega^2}{2} + \frac{\rho_2 - 2\nu\rho_1 + \nu^2}{\omega^2} \quad (4.49)$$

The **first derivatives of this limiting criterion function** are:

$$q_\nu(\theta; \rho) = \frac{\nu - \rho_1}{\omega^2}; \text{ and} \quad (4.50a)$$

$$q_{\omega^2}(\theta; \rho) = -\frac{1}{2\omega^2} \left[\frac{\rho_2 - 2\nu\rho_1 + \nu^2}{\omega^2} - 1 \right] \quad (4.50b)$$

which **coincides with the plim of the average scores under the same assumptions that guarantee the proper definition of the limiting criterion function.**

It is straightforward to see that the values of ν and ω^2 that solve these limiting focs are:

$$\nu(\rho_1, \rho_2) = \rho_1; \text{ and} \quad (4.51a)$$

$$\omega^2(\rho_1, \rho_2) = \rho_2 - \rho_1^2. \quad (4.51b)$$

Importantly, **these expressions not only set to zero the gradients of the limiting criterion function, but they also lead to a minimum of the limiting criterion function since they satisfy the**

second-order conditions too. They are the **binding functions**.

The **pseudo-true values are simply the values of these binding functions evaluated at the true values of the DGP parameters ρ_{10} and ρ_{20}** . Therefore when we estimate an i.i.d. Gaussian model we can be sure that we are estimating the unconditional mean and variance of the true distribution, provided x_t is such that the LLN applies to both its level and its square. Importantly, the result in no way depends on x_t being normal or even i.i.d..

4.2.1 Static Gaussian PMLE of a Homogenous Binary Markov Chain

■ The **parameters** characterising DGP are p and q .

■ The **estimated parameters are ν and ω^2** .

Pseudo-maximum likelihood estimators:

$$\tilde{\nu}_T = \frac{1}{T} \sum_1^T x_t = \hat{\lambda}_T; \text{ and} \quad (4.52a)$$

$$\tilde{\omega}^2 = \frac{1}{T} \sum_1^T x_t^2 - \left[\frac{1}{T} \sum_1^T x_t \right]^2 = \hat{\lambda}_T (1 - \hat{\lambda}_T). \quad (4.52b)$$

In this case, we **simply need to make sure that a Law of Large Numbers can be applied to:**

$$p \lim_T \left\{ \frac{1}{T} \left(\sum_{t=1}^T x_t \right) \right\}. \quad (4.53)$$

because:

$$x_t^2 = x_t. \quad (4.54)$$

But we know that x_t has the autocorrelation structure of an $AR(1)$, so provided that we rule out extreme cases (reducible states or cyclical chains), the sample average of x_t will indeed converge to $\mathbb{E}(x_t)$ by **the ergodic theorem for finite Markov chains**. Hence:

$$Q(\nu, \omega^2; p, q) = \mathbb{E}[-l_t(\theta) | p, q] \implies \quad (4.55a)$$

$$Q(\nu, \omega^2; p, q) = \frac{\ln 2\pi}{2} + \frac{\ln \omega^2}{2} + \frac{\mathbb{E}[\varepsilon_t^{*2}(\theta) | p, q]}{2} \implies \quad (4.55b)$$

$$Q(\nu, \omega^2; p, q) = \frac{\ln 2\pi}{2} + \frac{\ln \omega^2}{2} + \frac{\lambda(p, q) - 2\lambda(p, q)\nu + \nu^2}{2\omega^2}, \text{ where} \quad (4.55c)$$

$$\lambda(p, q) = \frac{1-p}{2-p-q} \implies \quad (4.55d)$$

$$\lambda(p, q) = \mathbb{E}(x_t | p, q) \implies \quad (4.55e)$$

$$\lambda(p, q) = \mathbb{E}(x_t^2 | p, q). \quad (4.55f)$$

The **limiting score** is:

$$q_\nu(\theta; p, q) = \frac{\nu - \lambda(p, q)}{\omega^2}; \text{ and} \quad (4.56a)$$

$$q_{\omega^2}(\theta; p, q) = -\frac{1}{2\omega^2} \left[\frac{\lambda(p, q) - 2\nu\lambda(p, q) + \nu^2}{\omega^2} - 1 \right]. \quad (4.56b)$$

The binding functions are:

$$\nu(p, q) = \lambda(p, q); \text{ and} \quad (4.57a)$$

$$\omega^2(p, q) = \lambda(p, q)[1 - \lambda(p, q)]. \quad (4.57b)$$

These functions do not only set to 0 the limiting first-order conditions but they also satisfy the second-order conditions for minimisation.

The pseudo-true values:

$$\nu_{\infty} = \nu(p_0, q_0) = \frac{1 - p_0}{2 - p_0 - q_0} = \lambda_0; \text{ and} \quad (4.58a)$$

$$\omega_{\infty}^2 = \omega^2(p_0, q_0) = \lambda_0(1 - \lambda_0). \quad (4.58b)$$

4.2.2 MA(1) estimated as AR(1)

The true model is:

$$x_t = u_t - \delta u'_{t-1} \quad (4.59a)$$

$$u_t \mid x_{t-1}, x_{t-2}, \dots \sim \text{iid } N(0, \psi^2); \text{ and} \quad (4.59b)$$

$$|\delta| \leq 1, 0 < \psi^2 < \infty. \quad (4.59c)$$

The parameters are:

$$\rho = (\delta, \psi^2)^T. \quad (4.60)$$

The covariance structure is:

$$\gamma_0(\rho) = (1 + \delta^2) \psi^2; \quad (4.61a)$$

$$\gamma_1(\rho) = -\delta \psi^2; \text{ and} \quad (4.61b)$$

$$\gamma_j(\rho) = 0, j > 1. \quad (4.61c)$$

The estimated model is:

$$x_t = \beta x_{t-1} + v_t; \quad (4.62a)$$

$$v_t \sim \text{iid } N(0, \varphi^2); \text{ and} \quad (4.62b)$$

$$-1 < \beta < 1, \varphi^2 \geq 0. \quad (4.62c)$$

The estimated parameters are:

$$\theta = (\beta, \varphi^2)^T \quad (4.63)$$

The covariance structure:

$$\mathbb{V}(x_t) = \frac{\varphi^2}{1 - \beta^2}; \quad (4.64a)$$

$$\text{cov}(x_t, x_{t-1}) = \beta \mathbb{V}(x_t); \text{ and} \quad (4.64b)$$

$$\text{cov}(x_t, x_{t-j}) = \beta \text{cov}(x_{t-1}, x_{t-j}), j > 1. \quad (4.64c)$$

We have nesting only if $\delta = 0$. The pseudo-log-likelihood function:

$$\bar{l}_T(\theta) = -\frac{\ln 2\pi}{2} - \frac{\ln \varphi^2}{2} - \frac{\sum_t (x_t - \beta x_{t-1})^2}{2\varphi^2}. \quad (4.65)$$

The (average) score:

$$\bar{s}_{\beta T}(\theta) = \frac{1}{T} \sum_t \frac{x_t - \beta x_{t-1}}{\varphi} \frac{x_{t-1}}{\varphi}; \text{ and} \quad (4.66a)$$

$$\bar{s}_{\varphi^2 T}(\theta) = \frac{1}{2\varphi^2} \frac{1}{T} \sum_t \left[\frac{(x_t - \beta x_{t-1})^2}{\varphi^2} - 1 \right]. \quad (4.66b)$$

The **first-order conditions**:

$$\frac{1}{T} \sum_t \frac{x_t - \tilde{\beta}_T x_{t-1}}{\tilde{\varphi}_T} \frac{x_{t-1}}{\tilde{\varphi}_T} = 0; \text{ and} \quad (4.67a)$$

$$\frac{1}{2\tilde{\varphi}_T^2} \frac{1}{T} \sum_t \left[\frac{(x_t - \tilde{\beta}_T x_{t-1})^2}{\tilde{\varphi}_T^2} - 1 \right] = 0. \quad (4.67b)$$

The **pseudo-ML estimators**:

$$\tilde{\beta}_T = \frac{\hat{\sigma}_{01}}{\hat{\sigma}_{11}} \geq 0; \text{ and} \quad (4.68a)$$

$$\tilde{\varphi}_T^2 = \hat{\sigma}_{00} - \frac{\hat{\sigma}_{01}^2}{\hat{\sigma}_{11}}. \quad (4.68b)$$

The sufficient statistics comprise of the second moment matrix:

$$\text{vech}(\hat{\Sigma}) = (\hat{\sigma}_{00}, \hat{\sigma}_{01}, \hat{\sigma}_{11}) = \frac{1}{T} \sum_t (x_t^2, x_t x_{t-1}, x_{t-1}^2) \quad (4.69)$$

We need to **obtain the plims of the elements of $\hat{\Sigma}$ to find the limit of the objective function and its score**. Let us begin with $\hat{\sigma}_{00} = \frac{1}{T} \sum_t x_t^2$. Given the MA(1) structure:

$$x_t^2 = u_t^2 + \delta^2 u_{t-1}^2 - 2\delta u_t u_{t-1}, \quad (4.70)$$

the sample averages of u_t^2 and u_{t-1}^2 will converge in probability to ψ^2 by a standard application of Khinchin's Law of Large Numbers. The **sample average of $u_t u_{t-1}$ will converge to 0 by Chebyshev's LLN**. Hence:

$$p \lim_{T \rightarrow \infty} \hat{\sigma}_{00} = (1 + \delta^2) \psi^2 \implies \quad (4.71a)$$

$$p \lim_{T \rightarrow \infty} \hat{\sigma}_{00} = \gamma_0(\rho). \quad (4.71b)$$

Trivially, the same result applies to $\hat{\sigma}_{11}$. As for $\hat{\sigma}_{01} = \frac{1}{T} \sum_t x_t x_{t-1}$, the MA(1) structure implies that:

$$x_t x_{t-1} = u_t u_{t-1} - \delta u_{t-1}^2 - \delta u_t u_{t-2} + \delta^2 u_{t-1} u_{t-2}. \quad (4.72)$$

The only new term we have to worry about is the sample average of $u_t u_{t-2}$, which will also converge to 0 under Chebyshev's LLN. As a result:

$$p \lim_{T \rightarrow \infty} \hat{\sigma}_{01} = -\delta \psi^2 \implies \quad (4.73a)$$

$$p \lim_{T \rightarrow \infty} \hat{\sigma}_{01} = \gamma_1(\rho). \quad (4.73b)$$

The **population score** is:

$$q_\beta(\theta; \rho) = \text{plim}_{T \rightarrow \infty} \left\{ \frac{1}{T} \sum_t \frac{x_t - \beta x_{t-1}}{\varphi} \frac{x_{t-1}}{\varphi} \right\} \Rightarrow \quad (4.74a)$$

$$q_\beta(\theta; \rho) = \mathbb{E} \left[\frac{1}{T} \sum_t \frac{x_t - \beta x_{t-1}}{\varphi} \frac{x_{t-1}}{\varphi} \middle| \rho \right] \Rightarrow \quad (4.74b)$$

$$q_\beta(\theta; \rho) = \frac{1}{\varphi^2} [\gamma_1(\rho) - \beta \gamma_0(\rho)]; \text{ and} \quad (4.74c)$$

$$q_{\varphi^2}(\theta; \rho) = \text{plim}_{T \rightarrow \infty} \left\{ \frac{1}{2\varphi^2} \frac{1}{T} \sum_t \left[\frac{(x_t - \beta x_{t-1})^2}{\varphi^2} - 1 \right] \right\} \Rightarrow \quad (4.74d)$$

$$q_{\varphi^2}(\theta; \rho) = \mathbb{E} \left\{ \frac{1}{2\varphi^2} \frac{1}{T} \sum_t \left[\frac{(x_t - \beta x_{t-1})^2}{\varphi^2} - 1 \right] \middle| \rho \right\} \Rightarrow \quad (4.74e)$$

$$q_{\varphi^2}(\theta; \rho) = \frac{1}{2\varphi^2} \left[\frac{-2\beta\gamma_1(\rho) + (1 + \beta^2) \gamma_0(\rho)}{\varphi^2} - 1 \right], \quad (4.74f)$$

where the **dependence of γ_j on ρ comes through the autocovariances of the MA(1) model**. The **binding functions**:

$$\theta(\rho) : q[\theta(\rho), \rho] = 0; \quad (4.75a)$$

$$\beta(\rho) = \frac{\gamma_1(\rho)}{\gamma_0(\rho)}; \text{ and} \quad (4.75b)$$

$$\varphi^2(\rho) = \gamma_0(\rho) - \frac{\gamma_1^2(\rho)}{\gamma_0(\rho)}. \quad (4.75c)$$

These expressions also **satisfy the second-order conditions and therefore, they maximise the population criterion function**. The **pseudo-true values of the parameters will be**:

$$\beta_\infty = \frac{\gamma_1(\rho_0)}{\gamma_0(\rho_0)}; \text{ and} \quad (4.76a)$$

$$\varphi_\infty^2 = \gamma_0(\rho_0) - \frac{\gamma_1^2(\rho_0)}{\gamma_0(\rho_0)}. \quad (4.76b)$$

As expected, when we estimate a **Gaussian AR(1)** we can be sure that (subject to regularity) we are estimating the first order “autocorrelation” of x_t , which coincides with the coefficient in the population least squares projection of x_t onto x_{t-1} , as well as the associated mean square error.

Interestingly, we can use the binding functions to estimate the parameters of the true MA(1) process δ and ψ^2 from the OLS coefficients.

Alternatively, we could estimate those parameters from the population scores evaluated at the OLS coefficients.

Both these procedures are examples of *Indirect Inference*.

In this simple case, though, they would be equivalent to a GMM procedure in which the moments are $\mathbb{E}(x_t^2)$ and $\mathbb{E}(x_t x_{t-1})$.

Once we know what we are estimating, we can wonder about the sampling properties of $\hat{\theta}$, typically relying on asymptotic approximations.

We will do so by studying three issues: (1) identification; (2) consistency; and (3) limiting distribution.

4.2.3 Identification

Definition 4.3 (Observational Equivalence). $\theta^{**} \neq \theta^*$ is **observationally equivalent** to θ^* iff:

$$Q_T(\theta^*) = Q_T(\theta^{**}) \quad (4.77)$$

for all possible samples of size T .

If $Q_T(\theta)$ does not depend on θ for some values of y_1, \dots, y_T , then **we can say that there are many observationally equivalent values of θ in this particular sample**. However classical statisticians would usually say that what counts is whether $Q_T(\theta)$ depends on θ over the whole range of possible samples y_1, \dots, y_T .

Example 4.5 (Binary Markov Chain). In the binary Markov chain, **there are samples with no information on either p or q , but if we can draw repeated samples of y_t , then we will eventually get one for which the criterion function is not flat.**

Definition 4.4 (Asymptotic Observational Equivalence). $\theta^{**} \neq \theta^*$ is **asymptotically observationally equivalent** to θ^* iff:

$$Q(\theta^*) = Q(\theta^{**}) \quad (4.78)$$

Definition 4.5 (Asymptotically Globally Identifiable). θ_∞ is **asymptotically globally identifiable** iff there is **no other observationally equivalent value of θ over the admissible parameter space Θ** .

Note that the identifiability of a parameter depends on both the criterion function used to estimate it and the admissible parameter space.

Definition 4.6 (Asymptotically Identifiable Model). An asymptotically **identifiable model in one in which θ_∞ is asymptotically globally identifiable**.

Definition 4.7 (Asymptotically Locally Identifiable). θ_∞ is **asymptotically locally identifiable** iff **there is no observationally equivalent value in a neighbourhood of θ_∞** . More formally, if:

$$Q(\theta^i) > Q(\theta_\infty) \quad (4.79)$$

for any sequence θ^i such that:

$$\lim_{i \rightarrow \infty} \{\theta^i\} = \theta_\infty. \quad (4.80)$$

Definition 4.8 (Uncountable Underidentification/ Set Identified). There is a **continuum of values of θ that are observationally equivalent to θ_∞** .

Definition 4.9 (Countable Underidentification). There is a **finite or countably infinite number of values of θ that are observationally equivalent to θ_∞** .

Theorem 4.1 (Sufficient Conditions for Local Identification). *Let:*

$$\mathcal{A}(\theta; \rho) = \frac{\partial^2 Q(\theta; \rho)}{\partial \theta \partial \theta^T} \implies \quad (4.81a)$$

$$\mathcal{A}(\theta; \rho) = \frac{\partial q(\theta; \rho)}{\partial \theta^T}. \quad (4.81b)$$

If $\mathcal{A}(\theta; \rho_0)$ is **continuous at θ_∞** and

$$\text{rank}[\mathcal{A}(\theta_\infty, \rho_0)] = \dim(\theta) \quad (4.82)$$

then **θ_∞ is locally identified**.

Example 4.6. There **are non-linear models in which the rank condition fails**, and yet θ_∞ is locally identified. For example, take:

$$Q(\theta) = \theta^4; \tag{4.83a}$$

$$A(\theta) = 12\theta^2; \text{ and} \tag{4.83b}$$

$$\theta_\infty = 0. \tag{4.83c}$$

In that case, θ_∞ **is said to be first-order underidentifiable**. Notable examples are as follows.

➡ **MA(1):** The **two generally isolated pairs of parameter values (β, σ^2) and $(\beta^{-1}, \beta^2 \sigma^2)$ give rise to the same first and second moments.**

➡ **ARMA(1, 1):** The line $\alpha = \beta; \sigma^2 = \sigma_0^2$ gives rise to the same first and second moments.

Therefore, if $Q_T(\theta)$ is such that it effectively depends on the data through the first and second moments, as in a Gaussian pseudo-likelihood or most GMM procedures, in the previous examples we will have global or local underidentification, respectively. In example 1, though, the lack of global identifiability can be eliminated by requesting the estimated model to be invertible.

4.2.4 Consistency

Theorem 4.2 (Consistency). *Assume that:*

1. Θ **is compact in \mathbb{R}^p ($\theta \in \Theta \subseteq \mathbb{R}^p$)**;
2. $Q_T(\theta)$ **is continuous in y_1, \dots, y_T** ;
3. $\sup_{\theta \in \Theta} |Q_T(\theta) - Q(\theta)|$; **and**
4. θ_∞ **is unique, at least locally.**

Then:

$$\boxed{\hat{\theta}_T \xrightarrow{p} \theta_\infty.} \tag{4.84}$$

This simply says that under **those sufficient conditions, the limit of the minimisers is the minimiser of the limiting function**. Continuity of the sample criterion function is not necessary and can be replaced by weaker conditions. The most difficult regularity condition to check for specific models is uniform convergence.

- 🔔 Despite its importance, though, **consistency is not of much practical use because we only have access to finite sample sets.**
- 🔔 In practice, we would like to be able to make inferences about the **parameter values by studying their finite sample distributions.**
- 🔔 Except in some textbook examples, **exact finite distributions cannot be obtained, and for that reason, we have to resort to approximations.**
- 🔔 The most **common approximation is given by the asymptotic (limiting) distribution, which is obtained after centring and scaling $\hat{\theta}_T$ so that the resulting distribution does not become degenerate as T increases without bound.**
- 🔔 The **classical central limit theorem is the best-known example of limiting distribution, although the approximation of binomial by normal proposed by de Moivre is even older.**

4.2.5 Limiting distribution & Sandwich Formula

Theorem 4.3 (Limiting Distribution). *Consider the following conditions:*

1. $\theta_\infty \in \text{int}(\Theta) \subseteq \mathbb{R}^p$;
2. we have that:

$$q_T(\theta) = \frac{\partial Q_T(\theta)}{\partial \theta}; \text{ and} \quad (4.85a)$$

$$\mathcal{A}_T(\theta) = \frac{\partial^2 Q_T(\theta)}{\partial \theta \partial \theta^T}. \quad (4.85b)$$

are *continuous in y_1, \dots, y_T , with $\mathcal{A}_T(\theta)$ of full rank.*

3. we have that:

$$\sqrt{T} q_T(\theta_\infty) \xrightarrow{d} \mathcal{N}[0, \mathcal{B}(\theta_\infty, \rho_0)]; \text{ and} \quad (4.86)$$

4. we have that: .

$$\mathcal{A}_T(\hat{\theta}_T^*) \xrightarrow{p} \mathcal{A}(\theta_\infty, \rho_0) \text{ if} \quad (4.87a)$$

$$\hat{\theta}_T^* \xrightarrow{p} \theta_\infty \quad (4.87b)$$

with $\mathcal{A}(\theta; \rho_0)$ of full rank,

then, we have that:

$$\boxed{\sqrt{T}(\hat{\theta}_T - \theta_\infty) \xrightarrow{d} \mathcal{N}[0, \mathcal{A}^{-1}(\theta_\infty; \rho_0) \mathcal{B}(\theta_\infty; \rho_0) \mathcal{A}^{-1}(\theta_\infty; \rho_0)]} \quad (4.88)$$

Proof. Multiply the first order conditions by \sqrt{T} , apply the mean-value theorem equation by equation, solve for $\sqrt{T}(\hat{\theta}_T - \theta_\infty)$, and use the assumptions. \square

If x_t is a strictly stationary and ergodic time series process:

$$\mathcal{B}(\theta_\infty, \rho_0) = \sum_{\tau=-\infty}^{\infty} \Gamma_\tau(\theta_\infty, \rho_0) = \Psi_{ss}(1), \text{ where} \quad (4.89a)$$

$$\Gamma_\tau(\theta; \rho) \equiv \mathbb{E}[s_t(\theta) s_{t-\tau}^T(\theta) | \rho]. \quad (4.89b)$$

Example 4.7 (Binary Markov Chain Cont'd). Let us revisit the example of the binary Markov chain estimated as a Gaussian i.i.d. model. However, we fix the variance to some arbitrary value to focus on the score for the mean. We can do this without loss of generality because the Gaussian PML of the mean is numerically the same whether we estimate the variance simultaneously or we fix it to some arbitrary value. As we have already seen, in this case, the *contribution of observation t to the score is*:

$$s_{\nu t}(\theta) = \omega^{-1} \varepsilon_t^*(\theta). \quad (4.90)$$

As a result:

$$\mathbb{E}[s_t(\theta_\infty) | \rho_0] = 0; \quad (4.91a)$$

$$\mathbb{E}[s_t(\theta_\infty) | u_{t-1}; \rho_0] = [(1 - p_0) + (q_0 + p_0 - 1) u_{t-1} - \theta] \frac{\frac{1}{2} - \theta(1 - \theta)}{\theta^2(1 - \theta)^2} \neq 0 \implies \quad (4.91b)$$

$$\mathbb{V}[\sqrt{T} \bar{s}_T(\theta_\infty) | \rho_0] \neq \mathbb{V}[s_t(\theta_\infty) | \rho_0] \implies \quad (4.91c)$$

$$\mathcal{B}(\theta_\infty, \rho_0) = \lim_{T \rightarrow \infty} \left\{ \mathbb{V}[\sqrt{T} \bar{s}_T(\theta_\infty)] \right\} \implies \quad (4.91d)$$

$$\mathcal{B}(\theta_\infty, \rho_0) = \frac{p_0 + q_0}{2 - p_0 - q_0} \frac{[\frac{1}{2} - \lambda_0(1 - \lambda_0)]^2}{\lambda_0(1 - \lambda_0)} \implies \quad (4.91e)$$

$$\mathcal{B}(\theta_\infty, \rho_0) = \frac{p_0 + q_0}{2 - p_0 - q_0} \mathbb{V}[s_t(\theta_\infty) | \rho_0]. \quad (4.91f)$$

Note that we have:

$$0 \leq \frac{p_0 + q_0}{2 - p_0 - q_0} < \infty \implies \quad (4.92a)$$

$$\underbrace{\frac{p_0 + q_0}{2 - p_0 - q_0} = 1 \Leftrightarrow p_0 + q_0 = 1;}_{\text{Serial Independence}} \quad (4.92b)$$

$$\underbrace{\frac{p_0 + q_0}{2 - p_0 - q_0} \rightarrow 0 \text{ as } p_0 + q_0 \rightarrow 0; \text{ or}}_{\text{Cyclical Chain}} \quad (4.92c)$$

$$\underbrace{\frac{p_0 + q_0}{2 - p_0 - q_0} \rightarrow \infty \text{ as } p_0 + q_0 \rightarrow 2}_{\text{Reducible Chain}} \quad (4.92d)$$

Hence:

$$\mathcal{A}(\theta_\infty, \rho_0) = \lim_{T \rightarrow \infty} \{ \mathbb{E} [-\bar{h}_T(\theta_\infty) \mid \rho_0] \} \implies \quad (4.93a)$$

$$\mathcal{A}(\theta_\infty, \rho_0) = \mathbb{E} [-h_t(\theta_\infty) \mid \rho_0] \implies \quad (4.93b)$$

$$\mathcal{A}(\theta_\infty, \rho_0) = \frac{\frac{1}{2} - \theta_0(1 - \theta_\infty)}{\theta_\infty^2(1 - \theta_\infty)^2} \quad (4.93c)$$

Further:

$$\mathcal{A}(\theta_\infty, \rho_0) \neq \mathcal{B}(\theta_\infty, \rho_0); \quad (4.94a)$$

$$\sqrt{T}(\tilde{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}[0, \mathcal{C}(\lambda_0)]; \text{ and} \quad (4.94b)$$

$$\mathcal{C}(\rho_0) = \mathcal{A}^{-1}(\lambda_0) \mathcal{B}(\rho_0) \mathcal{A}^{-1}(\lambda_0) \implies \quad (4.94c)$$

$$\mathcal{C}(\rho_0) = \frac{\lambda_0(1 - \lambda_0) \cdot (p_0 + q_0)}{(2 - p_0 - q_0)}. \quad (4.94d)$$

More generally, we can **consistently estimate the matrix** $\mathcal{B}(\theta_\infty, \rho_0)$ as

$$\tilde{\mathcal{B}}_T(\theta_\infty, \rho_0) = \sum_{\tau=-T^\iota}^{T^\iota} w(\tau) \tilde{\Gamma}_{\tau T}; \text{ and} \quad (4.95a)$$

$$\tilde{\Gamma}_{\tau T} = \frac{1}{T} \sum_{t=\tau+1}^T s_s(\tilde{\theta}_T) s_{t-\tau}^T(\tilde{\theta}_T), \quad (4.95b)$$

where $w(\tau)$: weights suggested by a standard HAC covariance estimation procedure, and ι is the asymptotic rate.

Again, some of these sufficient conditions can be weakened. But in studying asymptotic approximations, often the concern is the things that could go wrong. An **interesting situation arises when θ_∞ belongs to the closure of Θ but not to its interior**. In that case, we can get “half-normal” asymptotic distributions instead if we consider “one-sided” derivatives.

For example, the reciprocal of the degrees of freedom of a Student t is 0 under normality, which is right at the boundary of the admissible parameter space. Another unusual situation arises when **θ_∞ is locally identified, but the Hessian matrix suffers a rank deficiency so that it is not first-order identified**. In those cases, we can get limiting distributions but they will not be Gaussian. Problems may also arise when $\mathcal{B}(\theta_\infty; \rho_0)$ is singular, as in the next topic.

4.2.6 Gaussian Pseudo Maximum Likelihood Estimation

Consider the **following estimated multivariate model**:

$$x_t \mid x_{t-1}, \dots \sim \mathcal{N}[\mu_t(\theta), \Sigma_t(\theta)] \quad (4.96)$$

The **sample average of the Gaussian log-likelihood function** is:

$$Q_T(\theta) = -\frac{N}{2} \ln 2\pi - \frac{1}{2T} \sum_t \ln |\Sigma_t(\theta)| - \frac{1}{2T} \sum_t [x_t - \mu_t(\theta)]^T \Sigma_t^{-1}(\theta) [x_t - \mu_t(\theta)] \quad (4.97)$$

where we have made use of the prediction error decomposition.

A relevant issue we have ignored so far is **truncation**. In Markovian models, such as AR or ARCH, we **can condition on a finite number of initial observations**. For example, a Gaussian AR(1) model with ARCH(1) innovations requires two initial conditions. But **how do we deal with non-Markovian models?**

Consider a Univariate MA(1) Although we will often write:

$$\mu_t = \beta \varepsilon_{t-1}(\beta) \quad (4.98)$$

in fact what we really mean is:

$$\mu_t(\beta) = \sum_{j=1}^{\infty} (-\beta)^j x_{t-j}. \quad (4.99)$$

For t large enough **we can truncate this summation at little cost, as long as β is not too close to 1 or -1, but for the first few observations, the effect will be important**. One solution is to either backcast the pre-sample observations or to fix the values of the pre-sample innovations $\varepsilon_t(\beta)$ to 0 and their squares to their unconditional variance.

A better solution for linear models is to compute the log-likelihood function based on the Kalman filter predictive equations, which automatically give us $\mu_t(\theta)$ and $\Sigma_t(\theta)$ under the additional assumption that the distribution of x_1 is Gaussian. The prediction equations of the Hamilton filter also provide the log-likelihood computation for dynamic Markov switching models, although in this case, the log-likelihood function is a mixture of Gaussian distributions.

Remark. *Another very relevant numerical issue: **never invert a matrix unless you've got to. Use some numerically stable matrix decomposition instead.***

Although the spectral decomposition is always real for positive semidefinite matrices, the Cholesky decomposition is much more convenient:

$$\Sigma_t(\theta) = \Sigma_t^L(\theta) \Sigma_t^{L^T}(\theta) = \Sigma_t^{L1}(\theta) \Sigma_t^D(\theta) \Sigma_t^{L1^T}(\theta) \quad (4.100)$$

The **determinant of $\Sigma_t(\theta)$ is the square of the determinant of $\Sigma_t^L(\theta)$, which coincides with the product of its diagonal values, or the product of the diagonal values of $\Sigma_t^D(\theta)$.**

The **quadratic form $[X_t - \mu_t(\theta)]^T \Sigma_t^{-1}(\theta) [X_t - \mu_t(\theta)]$ can be very quickly obtained** as:

$$\varsigma_t(\theta) = \varepsilon_t^{*T}(\theta) \varepsilon_t^*(\theta) \quad (4.101)$$

where $\varepsilon_t^*(\theta)$ **is the solution to the lower triangular system of equations:**

$$\Sigma_t^L(\theta) \varepsilon_t^*(\theta) = \varepsilon_t(\theta), \varepsilon_t(\theta) = x_t - \mu_t(\theta) \quad (4.102)$$

which can be recursively solved in no time.

The **expression for the Gaussian log-likelihood score is:**

$$\mathbf{s}_{\theta t}(\boldsymbol{\theta}) = [\mathbf{Z}_{lt}(\boldsymbol{\theta}), \mathbf{Z}_{st}(\boldsymbol{\theta})] \begin{bmatrix} \mathbf{e}_{lt}(\boldsymbol{\theta}) \\ \mathbf{e}_{st}(\boldsymbol{\theta}) \end{bmatrix} \implies \quad (4.103a)$$

$$\mathbf{s}_{\theta t}(\boldsymbol{\theta}) = \mathbf{Z}_{dt}(\boldsymbol{\theta}) \mathbf{e}_{dt}(\boldsymbol{\theta}), \quad (4.103b)$$

where

$$\mathbf{Z}_{lt}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\mu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot \boldsymbol{\Sigma}_t^{-1/2'}(\boldsymbol{\theta}); \quad (4.104a)$$

$$\mathbf{Z}_{st}(\boldsymbol{\theta}) = \frac{1}{2} \partial \text{vec}^T [\boldsymbol{\Sigma}_t(\boldsymbol{\theta})] / \partial \boldsymbol{\theta} \cdot \left[\boldsymbol{\Sigma}_t^{-1/2'}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_t^{-1/2'}(\boldsymbol{\theta}) \right]; \text{ and} \quad (4.104b)$$

$$\mathbf{e}_{dt}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{e}_{lt}(\boldsymbol{\theta}) \\ \mathbf{e}_{st}(\boldsymbol{\theta}) \end{bmatrix} = \left\{ \begin{array}{c} \boldsymbol{\varepsilon}_t^*(\boldsymbol{\theta}) \\ \text{vec} [\boldsymbol{\varepsilon}_t^*(\boldsymbol{\theta}) \boldsymbol{\varepsilon}_t^{*'}(\boldsymbol{\theta}) - \mathbf{I}_N] \end{array} \right\}. \quad (4.104c)$$

This might sound very complicated, but it **is the multivariate version of the univariate expression:**

$$s_{\theta t}(\theta) = \frac{\varepsilon_t^*(\theta)}{\sigma_t(\theta)} \frac{\partial \mu_t(\theta)}{\partial \theta} + \frac{[\varepsilon_t^{*2}(\theta) - 1]}{2\sigma_t^2(\theta)} \frac{\partial \sigma_t^2(\theta)}{\partial \theta}. \quad (4.105)$$

The **expression for the Gaussian log-likelihood Hessian is:**

$$\mathbf{h}_{\theta\theta t}(\boldsymbol{\theta}) = \frac{\partial^2 d_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - \frac{1}{2} \frac{\partial^2 \varsigma_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \quad (4.106)$$

with

$$\begin{aligned} \frac{\partial^2 d_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} &= \frac{1}{2} \frac{\partial \text{vec}^T [\boldsymbol{\Sigma}_t(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \left[\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \right] \frac{\partial \text{vec} [\boldsymbol{\Sigma}_t(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^T} \\ &\quad - \frac{1}{2} \left\{ \text{vec}^T [\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})] \otimes \mathbf{I}_p \right\} \frac{\partial \text{vec}}{\partial \boldsymbol{\theta}^T} \left\{ \frac{\partial \text{vec}^T [\boldsymbol{\Sigma}_t(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \right\}. \end{aligned} \quad (4.107)$$

and

$$\begin{aligned} \frac{\partial^2 \varsigma_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} &= 2 \frac{\partial \boldsymbol{\mu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\mu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} + 2 \frac{\partial \text{vec}^T [\boldsymbol{\Sigma}_t(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \\ &\quad \times \left[\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{\varepsilon}_t(\boldsymbol{\theta}) \boldsymbol{\varepsilon}_t^T(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \right] \frac{\partial \text{vec} [\boldsymbol{\Sigma}_t(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^T} \\ &\quad + 2 \frac{\partial \boldsymbol{\mu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left[\boldsymbol{\varepsilon}_t^T(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \right] \frac{\partial \text{vec} [\boldsymbol{\Sigma}_t(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^T} \\ &\quad + 2 \frac{\partial \text{vec}^T [\boldsymbol{\Sigma}_t(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \left[\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{\varepsilon}_t(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \right] \frac{\partial \boldsymbol{\mu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \\ &\quad - 2 \left[\boldsymbol{\varepsilon}_t^T(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \otimes \mathbf{I}_p \right] \frac{\partial \text{vec}}{\partial \boldsymbol{\theta}^T} \left\{ \frac{\partial \boldsymbol{\mu}_t^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\} \\ &\quad - \left\{ \text{vec}^T [\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{\varepsilon}_t(\boldsymbol{\theta}) \boldsymbol{\varepsilon}_t^T(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})] \otimes \mathbf{I}_p \right\} \frac{\partial \text{vec}}{\partial \boldsymbol{\theta}^T} \left\{ \frac{\partial \text{vec}^T [\boldsymbol{\Sigma}_t(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \right\}, \end{aligned} \quad (4.108)$$

where **the exact expressions depend on the specific assumption made for estimation purposes.** Once again, this expression simplifies in the univariate case.

4.2.7 Correctly Specified 1st and 2nd Conditional Moments

Let us make the **following two assumptions:**

1. $\mathbb{E}(x_t \mid X_{t-1}; \rho_0) = \mu_t(\theta^*)$ for some $\theta^* \in \Theta$; and
2. $\mathbb{V}(x_t \mid X_{t-1}; \rho_0) = \Sigma_t(\theta^*)$ for the same $\theta^* \in \Theta$.

This means that although the **estimated model is potentially misspecified, we get the first two conditional moments right**. In that case, we can show that the Gaussian score evaluated at θ^* is a martingale difference sequence. The proof is straightforward once we notice that $\mathbf{e}_{lt}(\theta^*)$ and $\mathbf{e}_{st}(\theta^*)$ will be martingale difference sequences on their own. In turn, this **implies that $\theta_\infty = \theta^*$ since the unconditional expected value of the average score will also be 0 by the law of iterated expectations**.

It also implies that the **variance of the average score coincides with the average of the variances of the scores for each observation t , so no HAC weights are necessary**. In addition, the expression for the conditional expected value of the Hessian simplifies considerably.

Specifically, we are left with:

$$\mathbf{Z}_{lt}(\boldsymbol{\theta})\mathbf{Z}_{lt}^T(\boldsymbol{\theta}) + \mathbf{Z}_{st}(\boldsymbol{\theta})(\mathbf{I}_{N^2} + \mathbf{K}_{NN})\mathbf{Z}_{st}^T(\boldsymbol{\theta}) \quad (4.109)$$

where \mathbf{K}_{NN} is the commutation matrix of orders N and N , which transforms the vec of a square matrix \mathbf{A} of order N into the vec of its transpose:

$$\text{vec}(\mathbf{A}) = \mathbf{K}_{NN} \text{vec}(\mathbf{A}^T). \quad (4.110)$$

In effect, this means that the conditional expected value of the Hessian coincides with the expression for the conditional information matrix that we would obtain under the maintained assumption of normality. In the **univariate case, $\mathbf{K}_{11} = 1$, and we are left with:**

$$\frac{1}{\sigma^2(\theta_\infty)} \frac{\partial \mu(\theta_\infty)}{\partial \theta} \frac{\partial \mu(\theta_\infty)}{\partial \theta^T} + \frac{1}{2\sigma^4(\theta_\infty)} \frac{\partial \sigma^2(\theta_\infty)}{\partial \theta} \frac{\partial \sigma^2(\theta_\infty)}{\partial \theta^T} = \mathcal{A}(\theta_\infty, \rho_0) \quad (4.111)$$

The **expression for the conditional variance of the score is trickier, although it simplifies if we further assume that the conditional distribution of $\varepsilon_t(\theta^*)$ is *i.i.d.*, so that the dependence of x_t on its past is limited to its first two conditional moments**.

In particular, **if $\varepsilon_t^* \mid \mathbf{z}_t, I_{t-1}; \phi$ is *i.i.d.* $D(\mathbf{0}, \mathbf{I}_N, \varrho)$ with $\text{tr}[\mathcal{K}(\varrho)] < \infty$, then:**

$$\mathcal{B}_t(\theta_\infty, \rho_0) = V[\mathbf{s}_{\theta t}(\theta; \cdot) \mid X_{t-1}] = \mathbf{Z}_{dt}(\theta) \mathcal{K}(\rho) \mathbf{Z}_{dt}^T(\theta), \text{ where} \quad (4.112a)$$

$$\mathcal{K}(\rho) = V[\mathbf{e}_{dt}(\theta) \mid X_{t-1}] = \begin{bmatrix} \mathbf{I}_N & \Phi(\rho) \\ \Phi(\rho) & \Upsilon(\rho) \end{bmatrix}, \quad (4.112b)$$

where

$$\Phi(\rho) = E[\varepsilon_t^* \text{vec}^T(\varepsilon_t^* \varepsilon_t^{*'}) \mid \rho]; \text{ and} \quad (4.113a)$$

$$\Upsilon(\rho) = E[\text{vec}(\varepsilon_t^* \varepsilon_t^{*'} - \mathbf{I}_N) \text{vec}^T(\varepsilon_t^* \varepsilon_t^{*'} - \mathbf{I}_N) \mid \rho], \quad (4.113b)$$

depend on the multivariate third and fourth order cumulants of ε_t^* .

This expression simplifies if we make further assumptions, such as sphericity (e.g. a multivariate Student t), in which case:

$$\mathbf{Z}_{lt}(\boldsymbol{\theta})\mathbf{Z}_{lt}^T(\boldsymbol{\theta}) + \mathbf{Z}_{st}(\boldsymbol{\theta})[(\kappa + 1)(\mathbf{I}_{N^2} + \mathbf{K}_{NN}) + \kappa \text{vec}(\mathbf{I}_N) \text{vec}^T(\mathbf{I}_N)]\mathbf{Z}_{st}^T(\boldsymbol{\theta}), \quad (4.114)$$

where:

$$\kappa = \mathbb{E}\left(\frac{\varsigma_t^2}{(N(N+2))} - 1\right) \quad (4.115)$$

is the **multivariate excess kurtosis coefficient of the spherical distribution**.

Again, considerable simplification arises in the univariate case, when regardless of symmetry we obtain:

$$\mathbb{V}[s_t(\theta_\infty) | X_{t-1}; \rho_0] = \mathcal{B}_t(\theta_\infty, \rho_0) \implies \quad (4.116a)$$

$$\begin{aligned} \mathbb{V}[s_t(\theta_\infty) | X_{t-1}; \rho_0] &= \frac{\mathbb{V}[\varepsilon_t^*(\theta_\infty) | X_{t-1}; \rho_0]}{\sigma_t^2(\theta_\infty)} \frac{\partial \mu_t(\theta_\infty)}{\partial \theta} \frac{\partial \mu_t(\theta_\infty)}{\partial \theta^T} \\ &+ \frac{\mathbb{V}[\varepsilon_t^{*2}(\theta_\infty) | \rho_0]}{4\sigma_t^4(\theta_\infty)} \frac{\partial \sigma_t^2(\theta_\infty)}{\partial \theta} \frac{\partial \sigma_t^2(\theta_\infty)}{\partial \theta^T} \end{aligned} \quad (4.116b)$$

$$\begin{aligned} &+ \frac{\text{cov}[\varepsilon_t^*(\theta_\infty), \varepsilon_t^{*2}(\theta_\infty) | \rho_0]}{2\sigma_t^3(\theta_\infty)} \left[\frac{\partial \mu_t(\theta_\infty)}{\partial \theta} \frac{\partial \sigma_t^2(\theta_\infty)}{\partial \theta^T} + \frac{\partial \sigma_t^2(\theta_\infty)}{\partial \theta} \frac{\partial \mu_t(\theta_\infty)}{\partial \theta^T} \right] \implies \\ \mathbb{V}[s_t(\theta_\infty) | X_{t-1}; \rho_0] &= \frac{1}{\sigma_t^2(\theta_\infty)} \frac{\partial \mu_t(\theta_\infty)}{\partial \theta} \frac{\partial \mu_t(\theta_\infty)}{\partial \theta^T} + \frac{\kappa_0 - 1}{4\sigma_t^4(\theta_\infty)} \frac{\partial \sigma_t^2(\theta_\infty)}{\partial \theta} \frac{\partial \sigma_t^2(\theta_\infty)}{\partial \theta^T} \\ &+ \frac{\phi_0}{2\sigma_t^3(\theta_\infty)} \left[\frac{\partial \mu_t(\theta_\infty)}{\partial \theta} \frac{\partial \sigma_t^2(\theta_\infty)}{\partial \theta^T} + \frac{\partial \sigma_t^2(\theta_\infty)}{\partial \theta} \frac{\partial \mu_t(\theta_\infty)}{\partial \theta^T} \right]. \end{aligned} \quad (4.116c)$$

where ϕ and κ are the coefficients of skewness and kurtosis of the true distribution.

Of course, if the conditional distribution is not just i.i.d. but also normal, then the Gaussian pseudo-log likelihood becomes the true log-likelihood, and the conditional information matrix equality holds, which in turn implies that the usual unconditional information matrix equality will hold too. This can be easily checked from the previous expressions since $\phi = 0, \kappa = 3$ and $\varkappa = 0$ under normality.

4.3 Spectral Maximum Likelihood

4.3.1 Univariate Models

Let:

$$I_{yy}(\lambda) \equiv \frac{1}{2\pi T} \sum_{t=1}^T \sum_{s=1}^T (y_t - \pi)(y_s - \pi)^T e^{-i(t-s)\lambda} \quad (4.117)$$

denote the periodogram of y_t and

$$\lambda_j = 2\frac{\pi j}{T}, \quad j = 0, \dots, T-1 \quad (4.118)$$

the so-called Fourier frequencies.

Theorem 4.4 (Asymptotic Inconsistency & Unbiasedness of Periodogram). *$I_{yy}(\lambda)$ is an asymptotically unbiased non-parametric estimator of the spectral density at frequency λ . Still, it is not a consistent one because its sampling variance does not go to 0 when T increases without bound.*

There is a huge literature on non-parametric spectral density estimation, which achieves consistency by smoothing the periodogram using different windows. For example, we could average the periodograms of the Fourier frequencies closer to the one we are interested in. **Averaging reduces variance but it introduces bias.** Intuitively, to achieve consistency we must shrink the window width as the sample size increases to focus on those Fourier frequencies closer to the target one.

Weighted averages which give more weight to the closest frequencies and less weight to the relatively more distant ones typically work better than flat ones. To ensure that the **weights of these weighted averages are proper, we can rely on a density function or kernel**.^{4.2} The crucial element, though, is not the shape of the density function, but rather the rate at which its variance goes to 0 with the sample size. If this **smoothing parameter goes to 0 too fast, then we will under-smooth**. If, on the other hand, it

^{4.2}This is called the “kernel” density estimation.

goes too slowly, then we will over-smooth. Given the one-to-one relationship between spectral densities and autocovariance, the **smoothing procedures can be expressed in terms of weighted averages of the autocovariance**.

Remark (Newey-West HAC Procedure). *In fact, the so-called **Newey-West HAC procedure** that **estimates the spectral density function at frequency 0 coincides with the smoothed periodogram generated by the so-called Barlett kernel at that frequency**.*

Somewhat surprisingly, though, **the raw periodogram is all we need when we are interested in parametric spectral density estimation**. Whittle proposed a very clever way of obtaining maximum estimators of the model parameters for a stochastic process which is covariance stationary on a circle rather than the real line. Such a stochastic process is called a circulant one.

Definition 4.10 (Circulant Stochastic Process). A **circulant stochastic process** meets the following conditions:

$$\gamma(T-1) = \gamma(1); \quad (4.119a)$$

$$\gamma(T-2) = \gamma(2); \quad (4.119b)$$

and so on.

This imposes a very specific structure on the $T \times T$ autocovariance matrix of the observations, which becomes a (symmetric) circulant one. As a consequence, it **can be shown that its eigenvalues coincide with the spectral density at the Fourier frequencies, while its eigenvectors are given by a matrix which depends on T but not on the autocovariance properties of the process**. This matrix is the so-called **Fourier matrix**, and it exactly diagonalises the $T \times T$ autocovariance matrix of the observations.

Thus, Whittle **transformed a serially correlated but homoskedastic process** in the time domain into a **heteroskedastic but serially uncorrelated process in the frequency domain**. However, time series processes are not typically stationary on a circle even though they are stationary on the real line. Whittle's remarkable contribution was to **show that Fourier matrices asymptotically diagonalise such processes too**. In finite samples, though, it offers an approximation. The **discrete version of the (spectral) log-likelihood function** is:

$$-\frac{NT}{2} \ln(2\pi) - \frac{1}{2} \sum_{j=0}^{T-1} \ln |g_{yy}(\lambda_j)| - \frac{1}{2} \sum_{j=0}^{T-1} \frac{2\pi I_{yy}(\lambda_j)}{g_{yy}(\lambda_j)} \quad (4.120)$$

It is easy to prove that the **MLE of the mean of the process, which only enters through $I_{yy}(\lambda)$, is the sample mean, so in what follows we focus on demeaned variables**.

The **score w.r.t. all the remaining parameters** is:

$$\mathbf{s}_{\theta}(\theta) = \frac{1}{2} \sum_{j=0}^{T-1} \frac{\partial g_{yy}(\lambda_j)}{\partial \theta} M(\lambda_j) m(\lambda_j), \text{ where} \quad (4.121a)$$

$$m(\lambda) = 2\pi I_{yy}(\lambda) - g_{yy}(\lambda); \text{ and} \quad (4.121b)$$

$$M(\lambda_j) = g_{yy}^{-2}(\lambda_j). \quad (4.121c)$$

The **information matrix is block diagonal between the mean and the remaining parameters θ , with the (1,1)-element being $g_{yy}(0)$ and the (2,2)-block:**

$$\mathbf{Q} = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial g_{yy}(\lambda)}{\partial \theta} M(\lambda) \left\{ \frac{\partial g_{yy}(\lambda)}{\partial \theta} \right\}^* d\lambda \quad (4.122)$$

where $*$ denotes the conjugate transpose of a matrix.

A **consistent estimator** will be provided by

$$\Phi(\theta) = \frac{1}{2} \sum_{j=0}^{T-1} \frac{\partial g_{yy}(\lambda_j)}{\partial \theta} M(\lambda_j) \left\{ \frac{\partial g_{yy}(\lambda_j)}{\partial \theta} \right\}^*. \quad (4.123)$$

4.3.2 Multivariate Case

Let:

$$\mathbf{I}_{yy}(\lambda) = \frac{1}{2\pi T} \sum_{t=1}^T \sum_{s=1}^T (\mathbf{y}_t - \boldsymbol{\pi})(\mathbf{y}_s - \boldsymbol{\pi})^T e^{-i(t-s)\lambda} \quad (4.124)$$

denote the **periodogram matrix**.

The discrete version of the (spectral) log-likelihood function is

$$-\frac{NT}{2} \ln(2\pi) - \frac{1}{2} \sum_{j=0}^{T-1} \ln |\mathbf{G}_{yy}(\lambda_j)| - \frac{1}{2} \sum_{j=0}^{T-1} \text{tr} \{ \mathbf{G}_{yy}^{-1}(\lambda_j) [2\pi \mathbf{I}_{yy}(\lambda_j)] \} \quad (4.125)$$

Once again, the **MLE of the vector of means, which only enters through $\mathbf{I}_{yy}(\lambda)$, is the sample mean**.

The **score w.r.t. all the remaining parameters** is:

$$\mathbf{d}(\theta) = \frac{1}{2} \sum_{j=0}^{T-1} \frac{\partial \text{vec}^T [\mathbf{G}_{yy}(\lambda_j)]}{\partial \theta} \mathbf{M}(\lambda_j) \mathbf{m}(\lambda_j), \text{ where} \quad (4.126a)$$

$$\mathbf{m}(\lambda) = \text{vec} [2\pi \mathbf{I}_{yy}^T(\lambda) - \mathbf{G}_{yy}^T(\lambda)]; \text{ and} \quad (4.126b)$$

$$\mathbf{M}(\lambda) = [\mathbf{G}_{yy}^{-1}(\lambda) \otimes \mathbf{G}_{yy}'^{-1}(\lambda)]. \quad (4.126c)$$

The **information matrix** is:

$$\mathbf{Q} = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial \text{vec}^T [\mathbf{G}_{yy}(\lambda)]}{\partial \theta} \mathbf{M}(\lambda) \left[\frac{\partial \text{vec} [\mathbf{G}_{yy}(\lambda)]}{\partial \theta^T} d\lambda \right]^*. \quad (4.127)$$

Consistent estimators will be provided either by the outer product of the score or by:

$$\Phi(\theta) = \frac{1}{2} \sum_{j=0}^{T-1} \frac{\partial \text{vec}^T [\mathbf{G}_{yy}(\lambda_j)]}{\partial \theta} \mathbf{M}(\lambda_j) \left[\frac{\partial \text{vec} [\mathbf{G}_{yy}(\lambda_j)]}{\partial \theta^T} \right]^*. \quad (4.128)$$

4.3.3 Spectral Maximum Likelihood: Pros and Cons

- ☞ Spectral maximum likelihood is **very convenient for covariance stationary processes that have a linear representation**.
- ☞ As we will see **when we discuss hypothesis tests, the log-likelihood scores have a very simple interpretation**.
- ☞ In addition, it **offers the non-trivial advantage of providing “closed-form” expressions for the information matrix which do not depend on the data, only on the model parameters**, whereby “closed-form” I mean “up to a definite integral”.
- ☞ In fact, the **required integrals can be arbitrarily well approximated by their “sample” counterparts in a “fictitious” sample of size KT “generated” from the parametric model, where K is very large**.

- ✍ Given that $\Phi(\theta)$ does not depend on the data, it is not even necessary to generate the sample in the first place.
- ✍ Unfortunately, the **sandwich formula does not work, in the sense that the plim of the covariance matrix of the scores coincides with the information matrix regardless of the normality of the process.**

4.4 Sequential Estimation

Let:

$$\theta = (\theta_1^T, \theta_2^T)^T \quad (4.129)$$

where the joint estimator is:

$$\hat{\theta}_T = \arg \min_{\theta \in \Theta} Q_T(\theta_1, \theta_2). \quad (4.130)$$

Sometimes we **can consistently estimate θ_1 as:**

$$\check{\theta}_{1T} = \arg \min_{\theta_1} Q_T(\theta_1, \theta_2^*) \quad (4.131)$$

by **specifying some specific value for θ_2 , say θ_2^* .** The most obvious example is an **estimator of the conditional mean and variance parameters based on a Gaussian pseudo-log-likelihood function**, in which we effectively set the parameters that characterise the shape of the true distribution to 0. Having obtained $\check{\theta}_{1T}$ in this way, we can then estimate θ_2 as

$$\check{\theta}_{2T} = \arg \min_{\theta_2} Q_T(\check{\theta}_{1T}, \theta_2). \quad (4.132)$$

The **asymptotic distribution of the first block is easy to obtain using the usual expansion:**

$$\sqrt{T}(\check{\theta}_{1T} - \theta_{1\infty}) \xrightarrow{d} \mathcal{N}(0, \mathcal{A}_{11}^{-1} \Upsilon \mathcal{A}_{11}^{-1}) \quad (4.133)$$

where

$$\mathcal{A}_{11} = p \lim \left\{ T^{-1} \left[\sum_t \mathbf{h}_{\theta_1 \theta_1 t}(\theta_{1\infty}, \theta_2^*) \right] \right\}; \text{ and} \quad (4.134a)$$

$$\Upsilon = \mathbb{V} \left[T^{\frac{1}{2}} \sum_t \mathbf{s}_{\theta_1 t}(\theta_{1\infty}, \theta_2^*) \right]. \quad (4.134b)$$

As for the second block, I use Taylor expansion:

$$\begin{aligned} T^{1/2} \sum_{\theta_2^t} (\check{\theta}_{1T}, \theta_{2T}) &= T^{1/2} \sum_t \mathbf{s}_{\theta_2 t}(\theta_{1\infty}, \theta_{2\infty}) + \sum_t \frac{\partial \mathbf{s}_{\theta_2 t}(\theta_{1\infty}, \theta_{2\infty})}{\partial \theta_1^T} \sqrt{T}(\check{\theta}_{1T} - \theta_{1\infty}) + \\ &+ \sum_t \frac{\partial \mathbf{s}_{\theta_2 t}(\theta_{1\infty}, \theta_{2\infty})}{\partial \theta_2^T} \sqrt{T}(\check{\theta}_{2T} - \theta_{2\infty}) \end{aligned} \quad (4.135)$$

where **we can then plug the expansion of $\sqrt{T}(\check{\theta}_{1T} - \theta_{1\infty})$.**

The **sequential estimator of θ_2 is generally inefficient relative to the joint estimator**, but there are two sufficient conditions when it is not:

1. when $\sum_t \partial \mathbf{s}_{\theta_2 t}(\theta_{1\infty}, \theta_{2\infty}) / \partial \theta_1^T$ **converges in probability to 0**, in which case the asymptotic distribution of the sequential estimator is not affected by the sampling variability in the first stage estimator of θ_1 ; and
2. when the **first stage estimator of θ_1 is in fact efficient.**

4.5 Hypothesis Testing

The **null hypothesis** is $H_0 : \theta \in \Theta_0 \subseteq \Theta$. The **alternative hypothesis** is $H_1 : \theta \in \Theta_1 \subseteq \Theta$ ($\Theta_0 \cap \Theta_1 = \emptyset$, while $\Theta_0 \cup \Theta_1 = \Theta$ not required).

Both null and alternative hypotheses **can be simple or composite**. In the case in which the null is composite, we will usually look at the “least favourable” case. The decision theory implies the “Gun to your head” reject/don’t reject answer. Unfortunately, we will always make mistakes.

Definition 4.11 (Size of a Test). The **size of a test** is the probability of rejecting the null when it is true:

$$\alpha_T(\theta) = \mathbb{P}[(x_1, \dots, x_T) \in C_T \mid \theta \in \Theta_0]. \quad (4.136)$$

Definition 4.12 (Asymptotic Size). The **asymptotic size** is:

$$\alpha(\theta) = \lim_{T \rightarrow \infty} \alpha_T(\theta). \quad (4.137)$$

Definition 4.13 (Power of a Test). The **power of a test is the probability of rejecting the null** when it is not true:

$$\pi_T(\theta) = \mathbb{P}[(x_1, \dots, x_T) \in C_T \mid \theta \in \Theta_1]. \quad (4.138)$$

Definition 4.14 (Asymptotic Power). The **asymptotic power** is:

$$\pi(\theta) = \lim_{T \rightarrow \infty} \pi_T(\theta). \quad (4.139)$$

There is a **clear trade-off**.

1. To **minimise the probability of the type I error, we should never reject**.
2. To **minimise the probability of the type II error, we should always reject**.

Definition 4.15 (Test Consistence). A **consistent test** is one in which:

$$\lim_{T \rightarrow \infty} \pi_T(\theta) = 1. \quad (4.140)$$

This is a rather weak requirement, but it makes the comparison of alternative tests of the same null hypothesis complicated. There are **two standard solutions**:

1. **Pitman’s local alternatives**: Make the alternative closer and closer to the null (e.g., $H_l : \theta = \theta_1 + \delta/\sqrt{T}, \delta^T \delta \neq 0$; and
2. **Badahur’s approximate slope**: Divide the test statistic by a function of the sample size so that it converges in probability to a constant.

4.5.1 Classical Hypothesis Testing

A **null hypothesis is in the implicit form** if Θ_0 is defined by the **implicit equation** $h(\theta) = 0$, where $h(\cdot)$ is a vector of r functions. We will often assume that $\text{rank}\left(\frac{\partial h(\theta)}{\partial \theta^T}\right) = r$ to avoid redundant restrictions.

Definition 4.16 (Unrestricted Estimator). The **unrestricted estimator** is:

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} Q_T(\theta). \quad (4.141)$$

Definition 4.17 (Restricted Estimator). The **restricted estimator** is:

$$\tilde{\theta} = \text{argmin}_{\theta \in \Theta_0} Q_T(\theta). \quad (4.142)$$

The **triad of classical tests** is:

1. **Wald**;
2. **Likelihood Ratio** (the distance metric); and
3. **Lagrange Multiplier** (*Kuhn-Tucker* Multiplier).

Proposition 4.4.1 (Wald Test). *The **Wald test** can be derived by a straightforward application of the delta method to $h(\theta)$ evaluated at the unrestricted estimator $\hat{\theta}$:*

$$W_T = T h^T(\hat{\theta}_T) \left[\frac{\partial h(\theta)}{\partial \theta^T} \mathcal{I}^{-1} \frac{\partial h^T(\theta)}{\partial \theta} \right]^{-1} h(\hat{\theta}_T). \quad (4.143)$$

Under H_0 :

$$W_T \xrightarrow{d} \chi_r^2. \quad (4.144)$$

Under H_l .^{4.3}

$$W_T \xrightarrow{d} \chi_r^2(\psi), \quad (4.145)$$

where the non-centrality parameter ψ depends on the direction of departure δ . *Under H_1 :*

$$W_T \rightarrow \infty. \quad (4.146)$$

Proposition 4.4.2 (Likelihood Ratio). *The **likelihood ratio** is given by:*

$$LR_T = 2T \left[Q_T(\tilde{\theta}_T) - Q_T(\hat{\theta}_T) \right]. \quad (4.147)$$

Under H_0 and H_l :

$$W_T - LR_T \xrightarrow{p} 0. \quad (4.148)$$

Under H_1 :

$$LR_T \rightarrow \infty. \quad (4.149)$$

Proposition 4.4.3 (Lagrange Multiplier). *The **Lagrange multiplier test** can also be computed as a score test:*

$$LM_T = T q_T^T(\tilde{\theta}_T) \mathcal{I}^{-1} q_T(\tilde{\theta}_T) = T \tilde{\lambda}_T^T \frac{\partial h(\tilde{\theta}_T)}{\partial \theta^T} \mathcal{I}^{-1} \frac{\partial h^T(\tilde{\theta}_T)}{\partial \theta} \tilde{\lambda}_T. \quad (4.150)$$

Under H_0 and H_l :

$$W_T - LM_T \xrightarrow{p} 0. \quad (4.151)$$

Under H_1 :

$$LM_T \rightarrow \infty. \quad (4.152)$$

Wald and Lagrange multiplier can be made robust to misspecification of the log-likelihood.

Example 4.8. For simplicity, consider the case in which we can partition θ as $(\theta_1^T, \theta_2^T)^T$ and $H_0 : \theta_1 = \theta_1^*$.

In this case:

$$W_T = T \hat{\theta}_{1T}^T \mathcal{C}_{11}^{-1} \hat{\theta}_{1T}; \text{ and} \quad (4.153a)$$

$$LM_T = \left[\frac{\sqrt{T}}{T} \sum s_{\theta_{1t}}^T(\tilde{\theta}_T) \right] [\mathcal{A}^{11} \mathcal{C}_{11}^{-1} \mathcal{A}^{11}] \left[\frac{\sqrt{T}}{T} \sum s_{\theta_{1t}}^T(\tilde{\theta}_T) \right]. \quad (4.153b)$$

^{4.3}This is the so-called “local” alternative. This is when the restrictions do not hold but are becoming closer and closer to holding as T increases.

where \mathcal{A}^{11} is the $(1, 1)$ block of the inverse of \mathcal{A} and \mathcal{C}_{11} is the corresponding block for:

$$\mathcal{C} = \mathcal{A}^{-1} \mathcal{B} \mathcal{A}^{-1}. \quad (4.154)$$

Under H_0 :

$$W_T \xrightarrow{d} \chi_r^2; \text{ and} \quad (4.155a)$$

$$W_T - LM_T \xrightarrow{p} 0. \quad (4.155b)$$

Under H_l :

$$W_T \xrightarrow{d} \chi_r^2(\psi); \text{ and} \quad (4.156a)$$

$$W_T - LM_T \xrightarrow{p} 0. \quad (4.156b)$$

Under H_1 :

$$W_T \rightarrow \infty; \text{ and} \quad (4.157a)$$

$$LM_T \rightarrow \infty. \quad (4.157b)$$

LR will generally have a non-standard distribution in that case (Imhof distribution: quadratic form in normal variables).

4.6 Specification Tests

Sometimes we are interested in whether the estimated model is correctly specified. There are many ways in which we could do this, but we will only consider in some detail three approaches:

1. score tests;
2. Hausman tests; and
3. information matrix tests.

Score tests will artificially nest the estimated model into a more general model and then will test the parametric restrictions that reduce the more general model to the estimated model using LM statistics.

Example 4.9. Estimated model is $x_t \mid x_{t-1}, \dots$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, nesting models:

$$\overbrace{x_t \mid x_{t-1}, \dots \sim \mathcal{N}(\alpha + \rho x_{t-1}, \sigma^2), H_0 : \rho = 0;}^{\text{=Breusch-Godfrey Test}} \quad (4.158a)$$

$$\overbrace{x_t \mid x_{t-1}, \dots \sim \mathcal{N}(\mu, \omega + \alpha (x_{t-1} - \mu)^2), H_0 : \alpha = 0;}^{\text{=Engel's ARCH Test}} \text{ and} \quad (4.158b)$$

$$x_t \mid x_{t-1}, \dots \sim t(\mu, \sigma^2, \eta), H_0 : \eta = 0 \text{ (one-sided)}. \quad (4.158c)$$

4.6.1 Serial Correlation Tests in Frequency Domain

Suppose that **the model under the null is an AR(2) process**, while the **model under the alternative is**:

$$\underbrace{(1 - \psi_{x1}L)(1 - \alpha_{x1}L - \alpha_{x2}L^2)}_{\text{=AR(3)}} x_t = f_t. \quad (4.159)$$

Therefore, **the null is** $H_0 : \psi_{x1} = 0$. This is a multiplicative alternative. Instead, we could test $H_0 : \alpha_{x3} = 0$ **in the additive alternative**:

$$(1 - \alpha_{x1}L - \alpha_{x2}L^2 - \alpha_{x3}L^3)x_t = f_t. \quad (4.160)$$

In that case, it would be more convenient to re-parametrise the model in terms of partial autocorrelations. We stick to multiplicative alternatives, which cover MA terms too.

Under **the alternative, the spectral density of x_t is**:

$$g_{xx}(\lambda \mid \sigma_f^2, \alpha_{x1}, \alpha_{x2}, \psi_{x1}) = \frac{1}{1 + \psi_x^2 - 2\psi_x \cos \lambda} \cdot g_{xx}(\lambda \mid \sigma_f^2, \alpha_{x1}, \alpha_{x2}, 0) \quad (4.161)$$

where

$$g_{xx}(\lambda \mid \sigma_f^2, \alpha_{x1}, \alpha_{x2}, 0) = \frac{\sigma_f^2}{1 + \alpha_{x1}^2 + \alpha_{x2}^2 - 2\alpha_{x1}(1 - \alpha_{x2})\cos \lambda - 2\alpha_{x2}\cos 2\lambda}. \quad (4.162)$$

The **derivative of $g_{xx}(\lambda \mid \sigma_f^2, \alpha_{x1}, \alpha_{x2}, \psi_{x1})$ w.r.t. ψ_{x1} at H_0 is**:

$$\frac{\partial g_{xx}(\lambda \mid \sigma_f^2, \alpha_{x1}, \alpha_{x2}, 0)}{\partial \psi_{x1}} = 2 \cos \lambda \cdot g_{xx}(\lambda \mid \sigma_f^2, \alpha_{x1}, \alpha_{x2}, 0) \quad (4.163)$$

Hence, the **spectral version of the score w.r.t. ψ_{x1} under H_0 is**:

$$\sum_{j=0}^{T-1} \cos \lambda_j g_{xx}^{-1}(\lambda_j) [2\pi I_{xx}(\lambda_j) - g_{xx}(\lambda_j)] = \sum_{j=0}^{T-1} \cos \lambda_j [2\pi I_{ff}(\lambda_j)]. \quad (4.164)$$

Given that:

$$I_{ff}(\lambda_j) = \hat{\gamma}_{ff}(0) + 2 \sum_{k=1}^{T-1} \hat{\gamma}_{ff}(k) \cos(k\lambda_j) \quad (4.165)$$

the **spectral version of the score becomes**:

$$\sum_{j=0}^{T-1} \cos \lambda_j [2\pi I_{ff}(\lambda_j)] = T [\hat{\gamma}_{ff}(1) + \hat{\gamma}_{ff}(T-1)] \quad (4.166)$$

In turn, the **time domain version of the score will be**:

$$\sum_t (x_t - \alpha_{x1}x_{t-1} - \alpha_{x2}x_{t-2})(x_{t-1} - \alpha_{x1}x_{t-2} - \alpha_{x2}x_{t-3}) = \sum_t f_t f_{t-1} \quad (4.167)$$

which is (almost) identical as $\hat{\gamma}_{ff}(T-1) = T^{-1}f_T f_1 = o_p(1)$.

Remark. *Therefore, the spectral test of AR(2) versus AR(3) is simply checking that the first sample autocorrelation of f_t coincides with its theoretical value of 0 under H_0 .*

4.6.2 (Durbin-Wu-)Hausman Tests

(Durbin-Wu-)Hausman tests **compare the estimator of θ obtained under the maintained assumptions with another estimator obtained under generally weaker assumptions.**

For example, we may **compare the estimator of the mean obtained with a Student t log-likelihood with a Gaussian pseudo ML estimator.** Gaussian pseudo-maximum likelihood (PML) estimators remain consistent for the conditional mean and variance parameters of conditionally heteroskedastic dynamic regression models irrespective of the degree of asymmetry and kurtosis of the conditional distribution of the observed variables, so long as the first two moments are correctly specified and the fourth moments are bounded.

However, **empirical researchers are often interested in quantiles or tail correlations, which are necessary for the computation of commonly used risk management measures such as V @ R, and recently proposed systemic risk measures such as CoV@R.**

For that reason, they frequently specify a **non-Gaussian parametric distribution for the standardised innovations, which they use to estimate the conditional mean and variance parameters by maximum likelihood (ML), either fixing the parameters that characterise the shape of the assumed distribution to some reasonable values or jointly estimating them.** The dominant commercially available econometric packages have responded to this demand. Eviews and Stata support Student t and GED in univariate models. Stata also allows for Student t innovations in multivariate ones. A non-trivial advantage of these procedures is that they deliver more efficient estimators of the conditional mean and variance parameters.

Remark. *However, non-Gaussian ML estimators often achieve efficiency gains under correct specification at the risk of returning inconsistent parameter estimators under distributional misspecification. Consequently, it would be highly advisable to test the distributional assumptions behind non-Gaussian ML estimators.*

Consider a **multivariate dynamic model in discrete time** in which the $N \times 1$ vector of variables, \mathbf{y}_t , is assumed to be generated as:

$$\mathbf{y}_t = \boldsymbol{\mu}_t(\boldsymbol{\theta}_0) + \boldsymbol{\Sigma}_t^{1/2}(\boldsymbol{\theta}_0) \boldsymbol{\varepsilon}_t^* \quad (4.168a)$$

$$\boldsymbol{\mu}_t(\boldsymbol{\theta}) = \boldsymbol{\mu}(I_{t-1}; \boldsymbol{\theta}); \text{ and} \quad (4.168b)$$

$$\boldsymbol{\Sigma}_t(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(I_{t-1}; \boldsymbol{\theta}). \quad (4.168c)$$

where $\boldsymbol{\mu}()$ and $\text{vech}[\boldsymbol{\Sigma}()]$ are $N \times 1$ and $N(N+1)/2 \times 1$ vector functions known up to the $p \times 1$ vector of mean and variance parameters $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}_t^*$ is such that:

$$\mathbb{E}(\boldsymbol{\varepsilon}_t^* | I_{t-1}; \boldsymbol{\theta}_0) = \mathbf{0}; \text{ and} \quad (4.169a)$$

$$\mathbb{V}(\boldsymbol{\varepsilon}_t^* | I_{t-1}; \boldsymbol{\theta}_0) = \mathbb{I}_N \quad (4.169b)$$

so that **both the conditional mean vector and covariance matrix are correctly specified.**

To complete the model, we must specify the conditional distribution of $\boldsymbol{\varepsilon}_t^*$. Given the options in the commercially available software, we work with:

$$\boldsymbol{\varepsilon}_t^* | I_{t-1}; \boldsymbol{\theta}, \boldsymbol{\eta} \sim \text{i.i.d. } s(\mathbf{0}, \mathbf{I}_N, \boldsymbol{\eta}), \quad (4.170)$$

where $\boldsymbol{\eta}$ are some q **additional parameters that determine the shape of the distribution of $\zeta_t = \boldsymbol{\varepsilon}_t^* \boldsymbol{\varepsilon}_t^*$.** We **identify $\boldsymbol{\eta} = \mathbf{0}$ with normality.**

Example 4.10 (Multivariate Student- t). For example, STATA assumes that:

$$\boldsymbol{\varepsilon}_t^* | I_{t-1}; \boldsymbol{\theta}, \boldsymbol{\eta} \sim \text{i.i.d. } t(\mathbf{0}, \mathbf{I}_N, \boldsymbol{\eta}) \quad (4.171)$$

where $\boldsymbol{\eta} = \frac{1}{\nu}$, ν being the degrees of freedom of the t distribution.

Since the **multivariate t approaches the multivariate normal as $\nu \rightarrow \infty$, $0 \leq \boldsymbol{\eta} < \frac{1}{2}$.** Such inequality constraints on $\boldsymbol{\eta}$ sometimes play a crucial role. An **important moment is the population coefficient of multivariate excess kurtosis, which is defined as:**

$$\kappa = \mathbb{E} \left\{ \frac{\zeta^2}{N(N+2)} \right\} - 1 \implies \quad (4.172a)$$

$$\kappa = \frac{2}{\nu - 4}. \quad (4.172b)$$

for the Student t and 0 under normality.

Example 4.11 (Multivariate Market Model). Let \mathbf{r}_t denote the excess returns on a vector of N assets. Another rather popular model is:

$$\mathbf{r}_t = \mathbf{a} + \mathbf{b}r_{Mt} + \Omega^{\frac{1}{2}}\varepsilon_t^*. \quad (4.173)$$

In this case,

$$\mu_t(\boldsymbol{\theta}) = \mathbf{a} + \mathbf{b}r_{Mt}; \text{ and} \quad (4.174a)$$

$$\Sigma_t(\boldsymbol{\theta}) = \Omega = \Omega^{1/2}\Omega^{1/2} \quad (4.174b)$$

The parameters of interest are $\boldsymbol{\theta}^T = (\mathbf{a}^T, \mathbf{b}^T, \omega^T)$, where $\omega = \text{vech}(\Omega)$.

Estimators of $\boldsymbol{\theta}$ They are as follows:

1. the restricted (or infeasible) maximum likelihood (RML):

$$\hat{\boldsymbol{\theta}}_T(\bar{\boldsymbol{\eta}}) = \arg \max_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}, \bar{\boldsymbol{\eta}}) \quad (4.175)$$

where the infeasible parametric efficiency bound is $\mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\phi_0)$ assuming $\bar{\boldsymbol{\eta}} = \boldsymbol{\eta}_0$;

2. the joint (or unrestricted) maximum likelihood (UML):

$$(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\eta}}_T) = \arg \max_{\boldsymbol{\theta}, \boldsymbol{\eta}} L_T(\boldsymbol{\theta}, \boldsymbol{\eta}) \quad (4.176)$$

and the feasible parametric efficiency bound is:

$$\mathcal{P}(\phi_0) = [\mathcal{I}^{\boldsymbol{\theta}\boldsymbol{\theta}}(\phi_0)]^{-1} = \mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\phi_0) - \mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\eta}}(\phi_0)\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}(\phi_0)\mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\eta}}^T(\phi_0); \text{ and} \quad (4.177)$$

3. the Gaussian Pseudo ML (PML):

$$\tilde{\boldsymbol{\theta}}_T = \arg \max_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}, \mathbf{0}). \quad (4.178)$$

Assuming $\varepsilon_t^* \mid I_{t-1}; \phi \sim \text{i.i.d.} \mathcal{D}(\mathbf{0}, \mathbf{I}_N, \boldsymbol{\varrho})$ with bounded fourth moments:

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \rightarrow \mathcal{N}[\mathbf{0}, \mathcal{C}(\phi_0)]; \quad (4.179a)$$

$$\mathcal{C}(\phi) = \mathcal{A}^{-1}(\phi)\mathcal{B}(\phi)\mathcal{A}^{-1}(\phi); \quad (4.179b)$$

$$\mathcal{A}(\phi) = \mathbb{E}[\mathcal{A}_t(\phi) \mid \phi]; \quad (4.179c)$$

$$\mathcal{B}(\phi) = \mathbb{E}[\mathcal{B}_t(\phi) \mid \phi]; \quad (4.179d)$$

$$\mathcal{A}_t(\phi) = -\mathbb{E}[\mathbf{h}_{\boldsymbol{\theta}\boldsymbol{\theta}t}(\boldsymbol{\theta}; \mathbf{0}) \mid I_{t-1}; \phi]; \text{ and} \quad (4.179e)$$

$$\mathcal{B}_t(\phi) = \mathbb{V}[\mathbf{s}_{\boldsymbol{\theta}t}(\boldsymbol{\theta}; \mathbf{0}) \mid I_{t-1}; \phi]. \quad (4.179f)$$

4.6.3 The Matryoshka Dolls' Relative Efficiency of the Estimators

Proposition 4.4.4. *If $\varepsilon_t^* \mid I_{t-1}; \phi_0$ is i.i.d. $\mathcal{D}(\mathbf{0}, \mathbf{I}_N, \boldsymbol{\eta}_0)$ with $\kappa_0 < \infty$, then*

$$\begin{aligned} \lim_{T \rightarrow \infty} \mathbb{V} \left[\sqrt{T} \begin{pmatrix} \hat{\boldsymbol{\theta}}_T(\boldsymbol{\eta}_0) - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 \\ \tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 \end{pmatrix} \right] & \begin{matrix} \leftarrow \text{RML} \\ \leftarrow \text{UML} \\ \leftarrow \text{PML} \end{matrix} \\ & = \begin{bmatrix} \mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\phi_0) & \mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\phi_0) & \mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\phi_0) \\ \mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\phi_0) & \mathcal{P}^{-1}(\phi_0) & \mathcal{P}^{-1}(\phi_0) \\ \mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\phi_0) & \mathcal{P}^{-1}(\phi_0) & \mathcal{C}(\phi_0) \end{bmatrix}. \end{aligned} \quad (4.180)$$

Therefore, each estimator is “efficient” relative to all the others below.^{4.4}

^{4.4}UML \equiv unrestricted ML, RML \equiv restricted ML, and PML \equiv gaussian pseudo-ML.

This result **justifies up to three pairwise comparisons of different estimators using Durbin-Wu-Hausman (DWH) tests**. In each of them, the asymptotic covariance matrix of the difference between the estimators involved coincides with the difference in their respective covariance matrices.

Simultaneous Comparisons of Several Estimators: A priori, it is not clear which comparison researchers should focus on. The **early literature on DWH tests explicitly studied multiple comparisons for some well-known examples**:

- ▣ Full sample vs first subsample vs second subsample in Chow tests;
- ▣ GLS vs within-groups vs between-groups in panel data; and
- ▣ Tobit vs probit vs truncated regressions.

In those examples, though, all possible comparisons give rise to the same DWH test, but this does not happen in our case. However, there is no reason to choose just one such pair because the Matryoshka dolls' covariance structure implies that non-overlapping pairwise comparisons give rise to asymptotically independent test statistics. Therefore, generalised DWH tests that simultaneously look at multiple comparisons are extremely simple thanks to this convenient additive decomposition.

4.6.4 Information Matrix Tests

- ✓ The **information matrix (IM) test introduced by White (1982) constitutes a rather general procedure for examining the specification of models estimated by maximum likelihood (ML)**.
- ✓ It directly assesses the **IM equality, which states that the sum of the Hessian matrix and the outer product of the score vector should be zero in the expected value when the estimated model is correctly specified**.
- ✓ In other words, **IM tests compare the matrices \mathcal{A} and \mathcal{B}** . But how?
- ✓ Consider a **parametric model that fully characterises \mathbf{y} , a random vector of dimension M , as a function of ϕ , a p -dimensional vector of parameters**, with p finite, using its probability distribution in the discrete case or its density in the continuous one, both of which we will simply call $f(\mathbf{y}; \phi)$ henceforth.
- ✓ Assuming for simplicity that sampling is random, the log-likelihood function of a sample of size N on \mathbf{y} will be given by:

$$L_T(\phi) = \sum_{t=1}^T \ln f(\mathbf{y}_t; \phi) = \sum_{t=1}^T l_t(\phi) \quad (4.181)$$

Otherwise, we can use the prediction error decomposition. Consequently, the **average score and Hessian of this model will be given by**:

$$\bar{\mathbf{s}}_T(\phi) = \frac{1}{T} \frac{\partial L_T(\phi)}{\partial \phi} = \frac{1}{T} \sum_{t=1}^T \frac{\partial l_t(\phi)}{\partial \phi} = \frac{1}{T} \sum_{t=1}^T \mathbf{s}_t(\phi); \text{ and} \quad (4.182a)$$

$$\bar{\mathbf{h}}_T(\phi) = \frac{1}{T} \frac{\partial^2 L_T(\phi)}{\partial \phi \partial \phi} = \frac{1}{T} \sum_{t=1}^T \frac{\partial^2 l_t(\phi)}{\partial \phi \partial \phi} = \frac{1}{T} \sum_{t=1}^T \mathbf{h}_t(\phi), \quad (4.182b)$$

respectively. If we **call $\hat{\phi}_T$ the unrestricted maximum likelihood estimators of the parameters of interest, we will have that $\bar{\mathbf{s}}_T(\hat{\phi}_T) = \mathbf{0}$ and $\bar{\mathbf{h}}_T(\hat{\phi}_T)$ negative definite**.

The **information matrix test can be regarded as a moment test based on the following influence functions**:

$$\text{vech} [\mathbf{h}_t(\phi) + \mathbf{s}_t(\phi) \mathbf{s}_t^T(\phi)] = \mathbf{D}^+ \text{vec} [\mathbf{h}_t(\phi) + \mathbf{s}_t(\phi) \mathbf{s}_t^T(\phi)], \quad (4.183)$$

where \mathbf{D}^+ is the **Moore-Penrose inverse of the duplication matrix**.

In practice, we **need to evaluate these influence functions at $\hat{\phi}_T$, so we need to compute the asymptotic covariance matrix of**:

$$\frac{\sqrt{T}}{T} \sum_{t=1}^T \text{vech} [\mathbf{h}_t(\hat{\phi}_T) + \mathbf{s}_t(\hat{\phi}_T) \mathbf{s}_t^T(\hat{\phi}_T)] \quad (4.184)$$

We could **employ a standard first-order expansion of the influence functions, which requires the calculation of third-order derivatives of $l_t(\phi)$** . The above is based on the following Taylor expansion:

$$\frac{\sqrt{T}}{T} \sum_t m_t(\hat{\phi}_T) = \frac{\sqrt{T}}{T} \sum_t m_t(\phi_0) + \frac{1}{T} \sum_t \frac{\partial m_t}{\partial \phi^T}(\phi_0) \sqrt{T} (\hat{\phi}_T - \phi_0) + o_p(1); \text{ and} \quad (4.185a)$$

$$0 = \frac{\sqrt{T}}{T} \sum_t s_t(\hat{\phi}_T) = \frac{\sqrt{T}}{T} \sum_t s_t(\phi_0) + \frac{1}{T} \sum_t \frac{\partial s_t}{\partial \phi^T}(\phi_0) \sqrt{T} (\hat{\phi}_T - \phi_0) \implies \quad (4.185b)$$

$$\sqrt{T} (\hat{\phi}_T - \phi_0) = \left(-\frac{1}{T} \sum_t \frac{\partial s_t}{\partial \phi^T}(\phi_0) \right)^{-1} \frac{\sqrt{T}}{T} \sum_t s_t(\phi_0) + o_p(1) \implies \quad (4.185c)$$

$$\frac{\sqrt{T}}{T} \sum_t m_t(\hat{\phi}_T) = \frac{\sqrt{T}}{T} \sum_t m_t(\phi_0) + \frac{1}{T} \sum_t \frac{\partial m_t}{\partial \phi^T} \left(-\frac{1}{T} \sum_t \frac{\partial s_t}{\partial \phi^T} \right) \frac{\sqrt{T}}{T} \sum_t s_t(\phi_0) + o_p(1), \quad (4.185d)$$

where:

$$m_t(\phi) = \text{vech} [h_t + s_t(\phi) s_t^T(\phi)]. \quad (4.186)$$

However, we **can also use the generalised information matrix equality to obtain the expected value of the Jacobian of the influence functions above with respect to ϕ from the covariance matrix between them and $s_t(\phi)$ evaluated at the true values of the parameters, ϕ_0** . In effect, our maintained correct specification assumption means that we simply need to compute the residual covariance matrix from the least squares projection of the influence functions underlying the IM test onto the linear span of $\mathbf{s}_t(\phi_0)$, which is given by

$$\mathcal{R}(\phi_0) - \mathcal{U}(\phi_0) \mathcal{I}^{-1}(\phi_0) \mathcal{U}^T(\phi_0), \quad (4.187)$$

where

$$\begin{bmatrix} \mathcal{R}(\phi_0) & \mathcal{U}(\phi_0) \\ \mathcal{U}^T(\phi_0) & \mathcal{I}(\phi_0) \end{bmatrix} = \mathbb{V} \left\{ \begin{bmatrix} \text{vech} [\mathbf{h}_t(\phi_0) + \mathbf{s}_t(\phi_0) \mathbf{s}_t^T(\phi_0)] \\ \mathbf{s}_t(\phi_0) \end{bmatrix} \right\}. \quad (4.188)$$

Therefore, the **infeasible IM test statistic will be given by the following quadratic form**:

$$\begin{aligned} & T \left\{ \frac{1}{T} \sum_{t=1}^T \text{vech}^T [\mathbf{h}_t(\hat{\phi}_T) + \mathbf{s}_t(\hat{\phi}_T) \mathbf{s}_t^T(\hat{\phi}_T)] \right\} [\mathcal{R}(\phi_0) - \mathcal{U}(\phi_0) \mathcal{I}^{-1}(\phi_0) \mathcal{U}(\phi_0)]^+ \\ & \times \left\{ \frac{1}{T} \sum_{t=1}^T \text{vech} [\mathbf{h}_t(\hat{\phi}_T) + \mathbf{s}_t(\hat{\phi}_T) \mathbf{s}_t^T(\hat{\phi}_T)] \right\} \end{aligned} \quad (4.189)$$

where we **have relied on a Moore-Penrose generalised inverse because some of the influence functions may be an exact linear combination of $s_t(\phi_0)$ or appear multiple times**.

☞ A feasible version of the IM statistic is given by T times the R^2 in the regression of a vector of T ones onto $s_t(\hat{\phi}_T)$ and $\text{vech} [\mathbf{h}_t(\hat{\phi}_T) + \mathbf{s}_t(\hat{\phi}_T) \mathbf{s}_t^T(\hat{\phi}_T)]$ using an OLS routine robust to multicollinearity.

Effectively, the inclusion of $\mathbf{s}_t(\hat{\phi}_T)$ as additional regressors makes the test statistic robust to the fact that the influence functions (96) are evaluated at $\hat{\phi}_T$.

Nevertheless, this outer product regression has very poor finite sample properties, so in practice, it is better to work with the parametric bootstrap applied to a feasible version of infeasible test statistics which evaluates the theoretical expression for their variance at the MLE $\hat{\phi}_T$.

Example 4.12. Assume that y_t is normally distributed with mean μ and variance σ^2 so that, in terms of the notation above, we would have $\phi = (\mu, \sigma^2)^T$:

$$\mathbf{s}_t(\phi) = \begin{bmatrix} \frac{(y_t - \mu)}{\sigma^2} \\ \frac{(y_t - \mu)^2}{(2\sigma^4)} - \frac{1}{\sigma^2} \end{bmatrix} \quad (4.190)$$

and

$$\mathbf{h}_t(\phi) = - \begin{bmatrix} 1/\sigma^2 & (y_t - \mu)/\sigma^4 \\ (y_t - \mu)/\sigma^4 & (y_t - \mu)^2/(2\sigma^6) - 1/(2\sigma^4) \end{bmatrix}. \quad (4.191)$$

Hence, it is easy to prove that the IM test for a univariate normal random variable simply checks that the third- and fourth-order Hermite polynomials of the standardised variable, namely:

$$H_3(\varepsilon^*) = \varepsilon^{*3} - 3\varepsilon^*; \text{ and} \quad (4.192a)$$

$$H_4(\varepsilon^*) = \varepsilon^{*4} - 6\varepsilon^{*2} + 3, \quad (4.192b)$$

have zero means in the population.

In contrast, the sample average of the first and second polynomials:

$$H_1(\varepsilon^*) = \varepsilon^*; \text{ and} \quad (4.193a)$$

$$H_2(\varepsilon^*) = \varepsilon^{*2} - 1 \quad (4.193b)$$

will be 0 from the FOC of the MLE estimators of the mean and variance parameters.

5 Estimating Time Series Regression Models

5.1 Autoregressive Models

5.1.1 Case I

Consider the **model**:

$$x_t = \alpha x_{t-1} + u_t. \quad (5.1)$$

Assume the DGP is:

$$x_t \mid x_{t-1}, \dots \sim \mathcal{N}(\alpha x_{t-1}, \sigma^2). \quad (5.2)$$

This model is perfectly specified! The MLE are:

$$\hat{\alpha}_T = \frac{T^{-1} \sum_t x_t x_{t-1}}{T^{-1} \sum_t x_{t-1}^2}; \text{ and} \quad (5.3a)$$

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_t (x_t - \hat{\alpha}_T x_{t-1})^2. \quad (5.3b)$$

To obtain their asymptotic distribution we first need:

$$\begin{pmatrix} s_{\alpha t}(\alpha, \sigma^2) \\ s_{\sigma^2 t}(\alpha, \sigma^2) \end{pmatrix} = \begin{bmatrix} \frac{(x_t - \alpha x_{t-1})}{\sigma} \times \frac{x_{t-1}}{\sigma} \\ \frac{1}{2} \left\{ \frac{(x_t - \alpha x_{t-1})^2}{\sigma^2} - 1 \right\} \frac{1}{\sigma^2} \end{bmatrix} \quad (5.4)$$

which means that:

$$\mathbb{V}_{t-1} \left(\frac{\varepsilon_t^* \frac{x_{t-1}}{\sigma}}{\frac{[(\varepsilon_t^*)^2 - 1]}{\sigma^2}} \right) = \mathbb{E}_{t-1} \left(\frac{(\varepsilon_t^*)^2 \frac{x_{t-1}^2}{\sigma^2}}{\frac{[(\varepsilon_t^*)^2 - 1](\varepsilon_t^*)x_{t-1}}{2\sigma^3}} \quad \frac{\frac{[(\varepsilon_t^*)^2 - 1](\varepsilon_t^*)x_{t-1}}{2\sigma^3}}{\frac{1}{4} \left[(\varepsilon_t^*)^2 - 1 \right]^2 \frac{1}{\sigma^4}} \right) \Rightarrow \quad (5.5a)$$

$$\mathbb{V}_{t-1} \left(\frac{\varepsilon_t^* \frac{x_{t-1}}{\sigma}}{\frac{[(\varepsilon_t^*)^2 - 1]}{\sigma^2}} \right) = \begin{pmatrix} \frac{x_{t-1}}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}, \quad (5.5b)$$

which means that we arrive to:

$$I(\theta) = \mathbb{E} \{ \mathbb{E} [-\mathbf{h}_t(\theta) \mid x_{t-1}, \dots] \} = \begin{pmatrix} (1 - \alpha^2)^{-1} & 0 \\ 0 & (2\sigma^4)^{-1} \end{pmatrix} \quad (5.6)$$

so

$$\sqrt{T} \begin{pmatrix} \hat{\alpha}_T - \alpha_0 \\ \hat{\sigma}_T^2 - \sigma_0^2 \end{pmatrix} \xrightarrow{d} \mathcal{N} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 - \alpha_0^2 & 0 \\ 0 & 2\sigma_0^4 \end{pmatrix} \right] \quad (5.7)$$

Note that the **assumptions of the classical linear regression model do not quite hold in this model because the dynamic nature of the AR(1) implies that we cannot treat the regressors as fixed in repeated samples**. This means that the OLS estimator of the autoregressive coefficient will be biased (downwards) in finite samples, and the t-statistics will only be asymptotically valid.

In particular,

$$\underbrace{\mathbb{E}(\hat{\alpha}) = \alpha_0 - 2\frac{\alpha_0}{T} + O(T^{-2})}_{=\text{Biased}}, \quad (5.8)$$

which **means that it gets worse as α_0 increases**. The OLS-reported standard errors will continue to be valid, though, because $\hat{\sigma}_T^2$ is consistent.

5.1.2 Case II

Assume the DGP is

$$x_t \mid x_{t-1}, \dots \sim \mathcal{D}(\alpha x_{t-1}, \sigma^2) \quad (5.9)$$

but not necessarily Gaussian. The MLE are:

$$\hat{\alpha}_T = \frac{T^{-1} \sum_t x_t x_{t-1}}{T^{-1} \sum_t x_{t-1}^2}; \text{ and} \quad (5.10a)$$

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_t (x_t - \hat{\alpha}_T x_{t-1})^2. \quad (5.10b)$$

To obtain their asymptotic distribution we need to use the sandwich formula. We already know that:

$$\mathbb{E} \{ \mathbb{E} [-\mathbf{h}_t(\theta) \mid x_{t-1}, \dots] \} = \begin{pmatrix} (1 - \alpha^2)^{-1} & 0 \\ 0 & (2\sigma^4)^{-1} \end{pmatrix} \quad (5.11)$$

As for the variance of the score Ω :

$$\mathbb{V} \left[\begin{pmatrix} s_{\alpha t}(\alpha, \sigma^2) \\ s_{\sigma^2 t}(\alpha, \sigma^2) \end{pmatrix} \right]. \quad (5.12)$$

We are interested in:

$$\mathbb{E} \left[\frac{x_t - \alpha x_{t-1}}{\sigma} \frac{x_{t-1}}{\sigma} \right]^2 \quad (5.13a)$$

$$\mathbb{E} \left[\frac{1}{2} \left\{ \frac{(x_t - \alpha x_{t-1})^2}{\sigma^2} - 1 \right\} \frac{1}{\sigma^2} \right]^2, \text{ and} \quad (5.13b)$$

$$\mathbb{E} \left[\frac{x_t - \alpha x_{t-1}}{\sigma} \frac{x_{t-1}}{\sigma} \left\{ \frac{(x_t - \alpha x_{t-1})^2}{\sigma^2} - 1 \right\} \frac{1}{2\sigma^2} \right]. \quad (5.13c)$$

By LIE, the **previous expression reduces** to:

$$\mathbb{V} \left[\begin{pmatrix} s_{\alpha t}(\alpha, \sigma^2) \\ s_{\sigma^2 t}(\alpha, \sigma^2) \end{pmatrix} \right] = \begin{pmatrix} \mathbb{E} \left[\frac{x_{t-1}^2}{\sigma^2} \right] & \frac{1}{2} \mathbb{E} \left[\frac{x_{t-1} \phi_t}{\sigma} \right] \\ \frac{1}{2} \mathbb{E} \left[\frac{x_{t-1} \phi_t}{\sigma} \right] & \frac{1}{4} \mathbb{E} \left[\frac{(\kappa_t - 1)}{\sigma^4} \right] \end{pmatrix}, \quad (5.14)$$

so

$$\sqrt{T} \begin{pmatrix} \hat{\alpha}_T - \alpha_0 \\ \hat{\sigma}_T^2 - \sigma_0^2 \end{pmatrix} \xrightarrow{d} \mathcal{N} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 - \alpha^2 & \frac{1}{2} \mathbb{E} \left[\frac{x_{t-1} \phi_t}{\sigma} \right] \\ \frac{1}{2} \mathbb{E} \left[\frac{x_{t-1} \phi_t}{\sigma} \right] & \mathbb{E} [(\kappa_t - 1) \sigma^4] \end{pmatrix} \right] \quad (5.15)$$

where

$$\phi_t = \mathbb{E} \left[\frac{x_t - \alpha x_{t-1}}{\sigma} \left\{ \frac{(x_t - \alpha x_{t-1})^2}{\sigma^2} - 1 \right\} \middle| x_{t-1}, \dots \right] \quad (5.16)$$

and

$$\kappa_t = \mathbb{E} \left[\frac{(x_t - \alpha x_{t-1})^4}{\sigma^4} \middle| x_{t-1}, \dots \right]. \quad (5.17)$$

□ **In the special case in which $\phi_t = \phi \forall t$ then the unconditional covariance of the scores is 0 because x_t has zero unconditional mean.**

□ **If in addition $\kappa_t = \kappa \forall t$ then the variance of the second component of the score simplifies too.**

□ **In any case, if we are only interested in the autoregressive coefficient**, the asymptotic distribution is the same as in the case of a correctly specified model, and the related discussion remains valid.

5.2 LLNs and CLTs with Dependent Observations

Let y_t with:

$$\mathbb{E}(y_t) = \mu; \text{ and} \quad (5.18a)$$

$$\Gamma_j = \text{cov}(y_t, y_{t-j}) \quad (5.18b)$$

such that $\sum_{j=0}^{\infty} |\Gamma_j| < \infty$ (stronger than stationarity), then under certain assumptions on higher order dependence and moments of y_t :

$$\bar{y}_T \xrightarrow{2} \mu; \text{ and} \quad (5.19a)$$

$$\sqrt{T}(\bar{y}_T - \mu) \xrightarrow{d} \mathcal{N}[0, \Psi_{yy}(1)] \quad (5.19b)$$

where $\Psi_{yy}(1)$ is the **long-run covariance matrix**.

Definition 5.1 (Long-Run Covariance Matrix). $\Psi_{yy}(1)$ is the **long-run covariance matrix**:

$$\Psi_{yy}(1) = \sum_{j=-\infty}^{\infty} \Gamma_j. \quad (5.20)$$

5.3 Autoregressive Models Cont'd

5.3.1 Case III

Assume the DGP is:

$$\mathbb{E}[x_t \mid x_{t-1}, \dots] = \alpha x_{t-1}; \quad (5.21)$$

but:

$$\mathbb{V}[x_t \mid x_{t-1}, \dots] = \underbrace{\gamma + \delta (x_{t-1} - \alpha x_{t-2})^2}_{=\text{ARCH}(1)} \quad (5.22)$$

Thus, while the conditional mean remains correctly specified, the conditional variance is now misspecified. As a result:

$$\varepsilon_t^*(\alpha, \sigma^2) = \frac{(x_t - \alpha x_{t-1})}{\sigma} \quad (5.23)$$

will no longer be conditionally standardised.

Still, the **score with respect to α will continue to be a martingale difference sequence because $\varepsilon_t^*(\alpha, \sigma^2)$ is a martingale difference sequence.** But the **score with respect to σ^2 will not be because $\varepsilon_t^{*2}(\alpha, \sigma^2) - 1$ will no longer be a martingale difference sequence.** Therefore, the variance of the second element of the average score would be proportional to the autocovariance generating function of $\varepsilon_t^{*2}(\alpha, \sigma^2) - 1$ evaluated at 1.

In addition, **in computing the variance of the first element of the score, we have to take into account the fact that the conditional variance of $\sigma^{-1}x_{t-1}\varepsilon_t^*(\alpha, \sigma^2)$ is no longer x_{t-1}^2 , so the usual standard errors for $\hat{\alpha}$ will be wrong, and we would need to resort to heteroskedasticity robust standard errors.** As for the Hessian, its unconditional expected value will be the same as in the previous two cases, even though the conditional expected value will be different.

5.3.2 Case IV

Assume the **DGP is MA(1) with i.i.d. but not necessarily Gaussian innovations but we estimate an AR(1).** Specifically:

$$x_t = u_t - \beta u_{t-1}, \quad |\beta| < 1; \text{ and} \quad (5.24a)$$

$$u_t \mid x_{t-1}, x_{t-2}, \dots \sim \mathcal{D}(0, 1). \quad (5.24b)$$

We know the **binding function for α is:**

$$\alpha(\beta) = -\frac{\beta}{1 + \beta^2}. \quad (5.25)$$

We also **know that the score with respect to α is proportional to:**

$$m_t(\beta) = \left(x_t + \frac{\beta}{1 + \beta^2} x_{t-1} \right) x_{t-1}. \quad (5.26)$$

The most tedious thing is to **compute the autocovariance of this score evaluated at the pseudo-true value.**

In that regard, it is **convenient to write:**

$$m_t(\beta_0) = \left(1 + \frac{\beta_0 L}{1 + \beta_0^2} \right) x_t \cdot x_{t-1} = \left(1 + \frac{\beta_0 L}{1 + \beta_0^2} \right) (1 - \beta_0 L) u_t \cdot (1 - \beta_0 L) u_{t-1} \implies \quad (5.27a)$$

$$m_t(\beta_0) = \frac{1}{1 + \beta_0^2} (1 + \beta_0^2 - \beta_0^3 L - \beta_0^2 L^2) u_t \cdot (1 - \beta_0 L) u_{t-1} \implies \quad (5.27b)$$

$$m_t(\beta_0) = \frac{1}{1 + \beta_0^2} [(1 + \beta_0^2) u_t u_{t-1} - (1 + \beta_0^2) \beta_0 u_t u_{t-2} - \beta_0^3 u_{t-1}^2 - (1 - \beta_0^2) \beta_0^2 u_{t-1} u_{t-2} + \beta_0^3 u_{t-2}^2]. \quad (5.27c)$$

Given that $m_t(\beta_0)$ has the autocorrelation structure of an MA(2) at most, we will have that:

$$\mathbb{E}[m_t(\beta)m_{t-j}(\beta) \mid \beta] = 0 \quad \forall j \geq 3. \quad (5.28)$$

Therefore, we simply need to **compute a few moments**.

Specifically, we need:

$$\mathbb{E}[m_t^2(\beta) \mid \beta] = \frac{\beta_0^8 + (2\kappa - 3)\beta_0^6 + 4\beta_0^4 + 3\beta_0^2 + 1}{(1 + \beta_0^2)^2} \quad (5.29)$$

where $\kappa = \mathbb{E}(u_t^4) = 3$, because

$$\begin{aligned} & ((1 + \beta^2) u_t u_{t-1} - (1 + \beta^2) \beta u_t u_{t-2} - \beta^3 u_{t-1}^2 - (1 - \beta^2) \beta^2 u_{t-1} u_{t-2} + \beta^3 u_{t-2}^2)^2 \\ &= (1 + \beta^2)^2 u_t^2 u_{t-1}^2 + (1 + \beta^2)^2 \beta^2 u_t^2 u_{t-2}^2 + \beta^6 u_{t-1}^4 + (\beta^4 - 4\beta^2 + 1) \beta^4 u_{t-1}^2 u_{t-2}^2 \\ & \quad + \beta^6 u_{t-2}^4 - 2(1 + \beta^2)^2 \beta u_t^2 u_{t-1} u_{t-2} - 2(1 + \beta^2) \beta^3 u_t u_{t-1}^3 \\ & \quad + 2\beta^2 (2\beta^4 + \beta^2 - 1) u_t u_{t-1}^2 u_{t-2} + 2(1 + \beta^2) (2 - \beta^2) \beta^3 u_t u_{t-1} u_{t-2}^2 \\ & \quad - 2(1 + \beta^2) \beta^4 u_t u_{t-2}^3 + 2(1 - \beta^2) \beta^5 u_{t-1}^3 u_{t-2} - 2(1 - \beta^2) \beta^5 u_{t-1} u_{t-2}^3 \end{aligned} \quad (5.30)$$

By the same token:

$$\mathbb{E}[m_t(\beta)m_{t-1}(\beta) \mid \beta] = \frac{(2 - \kappa)\beta_0^6 - \beta_0^2}{(1 + \beta_0^2)^2} \quad (5.31)$$

because

$$\begin{aligned} & ((1 + \beta^2) u_t u_{t-1} - (1 + \beta^2) \beta u_t u_{t-2} - \beta^3 u_{t-1}^2 - (1 - \beta^2) \beta^2 u_{t-1} u_{t-2} + \beta^3 u_{t-2}^2) \\ & \times ((1 + \beta^2) u_{t-1} u_{t-2} - (1 + \beta^2) \beta u_{t-1} u_{t-3} - \beta^3 u_{t-2}^2 - (1 - \beta^2) \beta^2 u_{t-2} u_{t-3} + \beta^3 u_{t-3}^2) \\ &= u_{t-1} u_{t-2}^2 \beta^8 + (u_{t-3}^2 u_{t-1} u_{t-2} - 2u_{t-3} u_{t-1}^2 u_{t-2} + u_{t-3} u_{t-2}^3 - u_t u_{t-3} u_{t-2}^2 \\ & \quad - u_{t-1} u_{t-2}^3) \beta^7 + (-u_{t-2}^4 + u_t u_{t-2}^3 + (u_{t-3}^2 - 3u_{t-1} u_{t-3} + 2u_{t-1}^2) u_{t-2}^2 \\ & \quad + u_t (2u_{t-1} u_{t-3} - u_{t-3}^2) u_{t-2} + u_{t-3} u_{t-1}^3 - u_{t-1}^2 u_{t-3}^2) \beta^6 \\ & \quad + ((2u_{t-1} - u_{t-3}) u_{t-2}^3 - 2u_t u_{t-1} u_{t-2}^2 + (u_{t-3} u_{t-1}^2 - u_{t-1}^3 - u_{t-3}^2 u_{t-1}) u_{t-2} \\ & \quad - u_{t-1} u_{t-1}^2 u_{t-3} + u_{t-3}^2 u_t u_{t-1}) \beta^5 \\ & \quad + (u_t u_{t-2}^3 + u_t (u_{t-1}^2 - u_{t-3}^2 + 2u_{t-3} u_{t-1}) u_{t-2} + u_{t-3} u_{t-1}^3) \beta^4 \\ & \quad + (u_{t-1} u_{t-2}^3 + u_t (u_{t-3} - 3u_{t-1}) u_{t-2}^2 - u_{t-2} (2u_{t-3} u_{t-1}^2 - u_{t-3}^2 u_{t-1})) \beta^3 \\ & \quad + (u_{t-3} u_{t-1}^2 - u_{t-1}^3) u_{t-2} - u_{t-3} + u_{t-1} u_{t-2}^2) \beta + u_t u_{t-1}^2 u_{t-2}. \end{aligned} \quad (5.32)$$

Similarly:

$$\mathbb{E}[m_t(\beta)m_{t-2}(\beta) \mid \beta] = 0$$

because

$$\begin{aligned} & ((1 + \beta^2) u_t u_{t-1} - (1 + \beta^2) \beta u_t u_{t-2} - \beta^3 u_{t-1}^2 - (1 - \beta^2) \beta^2 u_{t-1} u_{t-2} + \beta^3 u_{t-2}^2) \\ & \times ((1 + \beta^2) u_{t-2} u_{t-3} - (1 + \beta^2) \beta u_{t-2} u_{t-4} - \beta^3 u_{t-3}^2 - (1 - \beta^2) \beta^2 u_{t-3} u_{t-4} + \beta^3 u_{t-4}^2) \\ &= u_{t-3} u_{t-4} u_{t-1} u_{t-2} \beta^8 + (u_{t-4} (u_{t-3} - u_{t-1}) u_{t-2}^2 + (u_{t-1} (u_{t-4}^2 - u_{t-3}^2) - u_{t-4} u_{t-3} u_t \\ & \quad + (-su_{t-2}^3 + (u_{t-4}^2 - u_{t-3}^2 + u_t u_{t-4} + u_{t-1} u_{t-3}) u_{t-2}^2 \\ & \quad + (u_{t-4} u_{t-1}^2 - 2u_{t-4} u_{t-3} u_{t-1} - u_t (u_{t-4}^2 - u_{t-3}^2)) u_{t-2} + u_{t-4} u_{t-3} u_t u_{t-1} - u_{t-1}^2 \\ & \quad + (u_{t-3} u_{t-2}^3 - u_{t-3} (u_{t-4} + u_t) u_{t-2}^2 - (u_{t-3} u_{t-1}^2 + u_{t-1} (u_{t-4}^2 - u_{t-3}^2) + u_{t-4} u_t u_{t-1} \\ & \quad + u_{t-4} u_{t-3} u_{t-1}^2 + u_t (u_{t-4}^2 - u_{t-3}^2) u_{t-1} \\ & \quad + (-u_{t-4} u_{t-2}^3 + 2u_{t-4} u_t u_{t-2}^2 + (u_{t-4} u_{t-1}^2 - u_t (u_{t-4}^2 - u_{t-3}^2) + u_{t-4} u_{t-3} u_{t-1} + u_{t-3} u_t \\ & \quad + (u_{t-3} u_{t-2}^3 + (u_{t-4} u_{t-1} - 2u_{t-3} u_t) u_{t-2}^2 + (u_{t-4} u_{t-3} u_t - 2u_{t-4} u_t u_{t-1} - u_{t-3} u_{t-1}^2) u_t \\ & \quad + ((u_{t-4} u_t - u_{t-3} u_{t-1}) u_{t-2}^2 + 2u_{t-3} u_t u_{t-1} u_{t-2} - u_{t-4} u_{t-3} u_t u_{t-1}) \beta^2 \\ & \quad - u_t (u_{t-3} u_{t-2}^2 + u_{t-4} u_{t-1} u_{t-2}) \beta + u_{t-3} u_t u_{t-1} u_{t-2}. \end{aligned} \quad (5.33)$$

so in fact, $m_t(\beta)$ has the autocorrelation structure of an MA(1).

As a result:

$$\lim \mathbb{V} \left(\sqrt{T} \bar{m}_T(\beta_0) \right) = \frac{\beta_0^8 + (2\kappa - 3)\beta_0^6 + 4\beta_0^4 + 3\beta_0^2 + 1}{(1 + \beta_0^2)^2} + 2 \frac{(2 - \kappa)\beta_0^6 - \beta_0^2}{(1 + \beta_0^2)^2} \implies \quad (5.34a)$$

$$\lim \mathbb{V} \left(\sqrt{T} \bar{m}_T(\beta_0) \right) = \frac{\beta_0^8 + \beta_0^6 + 4\beta_0^4 + \beta_0^2 + 1}{(1 + \beta_0^2)^2}. \quad (5.34b)$$

which does not depend on the kurtosis coefficient κ .

Remark. *This is the asymptotic variance of the first sample autocorrelation coefficient of an MA(1) process with i.i.d. innovations.*

5.4 Unit Roots

5.4.1 Driftless Random Walk

Consider

$$x_t = \alpha x_{t-1} + u_t, u_t \mid x_{t-1}, \dots \sim \mathcal{N}(0, \sigma^2) \quad (5.35)$$

for $t = 2, \dots, T$ and x_1 treated as fixed. Our previous analysis suggests that when $\alpha_0 = 1$:

$$\sqrt{T}(\hat{\alpha}_T - 1) \xrightarrow{d} \mathcal{N}(0, 0) \quad (5.36)$$

which is pretty useless if we want to conduct inferences about α . It simply means that

$$\mathbb{P} \left[\left| \sqrt{T}(\hat{\alpha}_T - 1) \right| > \varepsilon \right] \rightarrow 0, \forall \varepsilon > 0 \text{ as } T \rightarrow \infty \quad (5.37)$$

The intuition is as follows. $\hat{\alpha}_T$ is not longer root-T consistent but superconsistent i.e. $\hat{\alpha}_T$ approaches α_0 at a rate faster than $T^{1/2}$. So what can we do to make inferences?

To start with, let's consider the simple case in which we focus on the sample mean $T^{-1} \sum_t x_t$. Standard **CLT does not apply when** $\text{cov}(x_t, x_s) = |t - s|\sigma^2$. In any case, they typically apply to averages scaled by \sqrt{T} whose variance is finite, while the **variance of the mean in this random walk case becomes**:

$$\mathbb{V} \left(\frac{1}{T} \sum_t x_t \right) = \frac{\sigma^2}{T^2} \left[\sum_{\tau=1}^T \tau + 2 \sum_{\tau=1}^T \sum_{s=\tau}^T (t - s) \right] = O(T). \quad (5.38)$$

So we need an alternative approach. Notice that:

$$\frac{1}{T^{3/2}} \sum_t x_t = \frac{1}{T} \sum_t \frac{x_t}{\sqrt{T}} = \int_0^1 Y_T(r) dr \quad (5.39)$$

where

$$Y_T(r) = \frac{x_{[Tr]}}{\sqrt{T}}. \quad (5.40)$$

It is important to **realise that the Riemann-Stieljes integral above is a random variable, and therefore stochastic**. Specifically, it will take different values in different realisations of x_t .

From topic 1, we know that

$$Y_T(r) \rightarrow \sigma W(r) \quad (5.41)$$

Then, if we use the **Continuous Mapping Theorem**:

$$\int_0^1 Y_T(r) dr \xrightarrow{p} \sigma \int_0^1 W(r) dr \sim \mathcal{N} \left(0, \frac{\sigma^2}{3} \right) \text{ as } T \rightarrow \infty. \quad (5.42)$$

This expression provides us with a valid asymptotic distribution from which we can make inferences about the sample mean of x_t . We **could've assumed that u_t was i.i.d. Gaussian, derive the final sample distribution of the sample mean of x_t , which would be Gaussian, and take limits in distribution.** The advantage of the “unit root” approach above is its generality.

In particular, **we can show that if $x_t \sim I(1)$ and $\Delta x_t \sim I(0)$ but not WN :**

$$\frac{1}{T^{\frac{3}{2}}} \sum_t x_t \xrightarrow{p} \mathcal{N} \left[0, \frac{\psi_{\Delta x \Delta x}(1)}{3} \right]. \quad (5.43)$$

Going back to the **regression of x_t on x_{t-1} , by playing around with the numerator and the denominator of the OLS coefficient we can show that:**

$$T(\hat{\alpha}_T - 1) = \frac{T^{-1} \sum x_t u_t}{T^{-2} \sum x_t^2} \xrightarrow{p} \frac{1}{2} \frac{W^2(1) - 1}{\int_0^1 W^2(r) dr}, \quad (5.44)$$

where the **numerator follows a χ^2 distribution with 1 degree of freedom while the denominator is another stochastic integral.** This is a non-standard distribution.

However, we can tabulate it by simulation to any arbitrary degree of precision by **exploiting the convergence of a random walk to a Wiener process.** More generally, if $x_t \sim I(1)$ and $\Delta x_t \sim I(0)$ but not WN :

$$T(\hat{\alpha}_T - 1) = \frac{T^{-1} \sum x_t u_t}{T^{-2} \sum x_t^2} \xrightarrow{p} \frac{1}{2} \frac{\psi_{\Delta x \Delta x}(1) [W^2(1) - 1]}{\psi_{\Delta x \Delta x}^2(1) \int_0^1 W^2(r) dr}. \quad (5.45)$$

5.4.2 Unit Root Tests

Definition 5.2 (Dickey-Fuller Test). **Regress x_t on x_{t-1} (we want 1). Regress Δx_t on x_{t-1} (we want 0).**

There are some issues. They only work with **Δx_t being WN .** There are different tables for those cases in which the regression includes a constant or a time trend even though the true model is a driftless random walk.

The alternatives are as follows:

🔔 **Augmented DF Test.** Regress Δx_t on x_{t-1} and $\Delta x_{t-1}, \Delta x_{t-2}, \dots$ it works well with ARs.^{5.1}

🔔 **Phillip-Perron test.** Consider:

$$\frac{\hat{\alpha}_T - 1}{\sqrt{\frac{[T^{-1} \sum_t (x_t - \hat{\alpha}_T x_{t-1})^2]}{[T^{-1} \sum_t x_{t-1}^2]}}} \quad (5.46)$$

where the residual variance in the denominator is replaced by a consistent estimator of $\psi_{uu}(1) \dots$ it works well with MAs.

There is a **discontinuity in the AR(1) asymptotic theory.**

When $\alpha_0 = 1$ **we get unusual asymptotics, but when $\alpha_0 = .99$ we get standard asymptotics.** It gets even worse because we also get **standard asymptotics even though $\alpha_0 = 1$ if the true DGP is a random walk with a non-zero drift.** Common sense says that this discontinuity is a problem in the approximations, not in the finite sample distributions. We would expect the finite sample distributions for $\alpha_0 = .99$ to be closer to the finite sample distribution when $\alpha_0 = 1$ than when $\alpha_0 = .1$. Similarly, we would expect the finite sample

^{5.1}This doesn't work very well when the residual serial correlation in MA(1). If there is an issue with that, use the Phillips-Perron test.

distribution when the drift is very small to closely resemble the finite sample distribution of a driftless random walk. The possible solution is **the local to unity asymptotics**:

$$x_t = \left(1 - \frac{\delta}{T}\right) x_{t-1} + u_t. \quad (5.47)$$

5.5 Cointegration Tests

Let

$$y_t = \gamma x_t + u_t \quad (5.48)$$

with both $x_t, y_t \sim I(1)$ and we are interested in testing whether $y_t - \gamma x_t \sim I(0)$. In the late 80s, unit root test on u_t or \hat{u}_t . If γ (e.g. log-income and log-consumption) is known, then DF, ADF or PP. If not, $\hat{\gamma}_T$ is superconsistent then test the OLS residuals i.e. $y_t - \hat{\gamma}_T x_t$ on $y_{t-1} - \hat{\gamma}_T x_{t-1}$ (e.g. Engle and Granger (1987)).

We could also use the **rank test** (Johansen, early 90s). consider the VECM form of topic 3:

$$\begin{pmatrix} \Delta x_t \\ \Delta y_t \end{pmatrix} = -\Pi \begin{pmatrix} x_t \\ y_t \end{pmatrix} + \sum_{j=1}^{k-1} \beta \begin{pmatrix} \Delta x_{t-j} \\ \Delta y_{t-j} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}. \quad (5.49)$$

He obtained the (non-standard) **distribution of estimators of the rank of Π , which in turn are based on the rank of the distribution of the minimum singular values of $\hat{\Pi}$** . Consider the **following error vector correction model representation of a cointegrated VAR(2)**:

$$\Delta X_t = -HG^T X_{t-1} - A_2 \Delta X_{t-1} + u_t \quad (5.50)$$

If we assume for estimation purposes that:

$$u_t \mid X_{t-1}, \dots \sim \mathcal{N}(0, \Omega) \quad (5.51)$$

the **log-likelihood function of ΔX_t can be written using the expressions in the previous topic**. Nevertheless, since the conditional mean of the VECM is non-linear in H and G , it's worthwhile employing a few tricks. Specifically, if **H and G were known, we could estimate A_2 and Ω from the multivariate regression of $\Delta X_t + HG^T X_{t-1}$ on ΔX_{t-1}** . The same would apply to higher order VAR(p) models. Therefore, in practice, we can concentrate on estimating H and G . To do so, it is convenient to run the following auxiliary regressions:

$$\Delta X_t = B \Delta X_{t-1} + v_t; \text{ and} \quad (5.52a)$$

$$X_{t-1} = C \Delta X_{t-1} + w_t \quad (5.52b)$$

Let:

$$\hat{S}_{vv} = \frac{1}{T} \sum_{t=2}^T \hat{v}_t \hat{v}_t^T, \quad (5.53a)$$

$$\hat{S}_{ww} = \frac{1}{T} \sum_{t=2}^T \hat{w}_t \hat{w}_t^T, \text{ and} \quad (5.53b)$$

$$\hat{S}_{vw} = \frac{1}{T} \sum_{t=2}^T \hat{v}_t \hat{w}_t^T \quad (5.53c)$$

denote the **covariance matrices of the corresponding OLS residuals**. It can be proved that:

$$\hat{H} = -\hat{S}_{vw} \hat{G} \left(\hat{G}^T \hat{S}_{ww} \hat{G} \right)^{-1} \quad (5.54)$$

where \hat{G} are the eigenvectors associated to the generalised eigenvalues $\hat{\lambda}_1, \dots, \hat{\lambda}_k$ that solve:

$$\left| \hat{\lambda}_j \hat{S}_{ww} - \hat{S}_{vw}^T \hat{S}_{vv}^{-1} \hat{S}_{vw} \right| = 0. \quad (5.55)$$

In practice, it is convenient to impose the implicit normalisation restrictions $\hat{G}' \hat{S}_{ww} \hat{G} = I_k$ to pin down a unique basis for the k -dimensional linear subspace of cointegrating vectors. The asymptotic distribution of the Gaussian PML estimators of A_2 and Ω is \sqrt{T} Gaussian, and therefore standard. In contrast, the asymptotic distribution of the estimators of $\hat{H}^{5.2}$ and \hat{G} is a function of Brownian motions, and therefore non-standard. The maximum value of the log-likelihood function is (ignoring the constant of integration):

$$-\frac{T}{2} \ln |\hat{S}_{vv}| - \frac{T}{2} \sum_{j=1}^k \ln (1 - \hat{\lambda}_j) \quad (5.56)$$

Johansen's cointegration test compares this criterion function to the log-likelihood function of the stationary model

$$X_t = A_1 X_{t-1} + A_2 X_{t-2} + u_t \quad (5.57)$$

in which there are N cointegrating relationships.

Thus, the LR version of his test will be:

$$-T \sum_{j=k+1}^N \ln (1 - \hat{\lambda}_j). \quad (5.58)$$

There is also an asymptotically equivalent trace version, which exploits the approximation:

$$\ln (1 - \hat{\lambda}_j) \simeq -\hat{\lambda}_j. \quad (5.59)$$

As in the case of the Dickey-Fuller tests, the common distribution of those tests is non-standard.

5.6 Dynamic Regression Models in Practice: Exchange Rate Example

Let s_t be the log-exchange rate $\$/\mathcal{L}$, $r_t^{\$}$ and $r_t^{\mathcal{L}}$ be the interest rate of short term investments (say, money market account) in dollars and euros, respectively. Let also $\Delta_2 s_t = s_t - s_{t-2}$; then **the uncovered interest rate parity states that**:

$$\mathbb{E}_{t-1} [\Delta_2 s_{t+1}] = p_{2,t-1} \quad (5.60)$$

where:

$$p_{2,t} = r_t^{\$} - r_t^{\mathcal{L}}. \quad (5.61)$$

How do we test this theory? $H_0 : \alpha = 0, \beta = 1$ in

$$\Delta_2 s_{t+1} = \alpha + \beta p_{2,t-1} + u_{t+1} \quad (5.62)$$

The issue is that u_{t+1} has serial correlation since $\Delta_2 s_{t+1} = \Delta s_t + \Delta s_{t+1}$ (two forecasting errors there, $MA(1)$ structure in this example). More generally, for q steps-ahead forecast, $MA(q-1)$ structure in the forecasting errors.

OLS is still consistent, however, **an inference could be misleading** $\nabla(\hat{\beta})$ could be estimated **inconsistently**, as you already should know (last two lectures).

^{5.2}The null hypothesis is that the order of cointegration is equal to the rank of H .

Unfortunately, GLS is generally inconsistent because the regressor is not strictly exogenous. The **first solution** is to **compute an OLS standard error which is robust to serial correlation and possible conditional heteroskedasticity**. The **second solution is to apply ML** to a full model that tries to capture the actual DGP, say a joint VAR for $s_t - s_{t-1}$ and $r_t^{\$} - r_t^{\mathcal{L}}$ in the previous example, and then test the **Rational Expectation-type cross-equation restrictions implied by the Uncovered Interest Rate parity theory**. The first approach has the advantage that it makes fewer assumptions, but it is inefficient. Perhaps more importantly, the HAC estimators don't work well when the degree of overlap is very high. The second approach is more efficient and doesn't require robust standard errors, but it can give rise to misleading conclusions if the joint model is incorrectly specified.