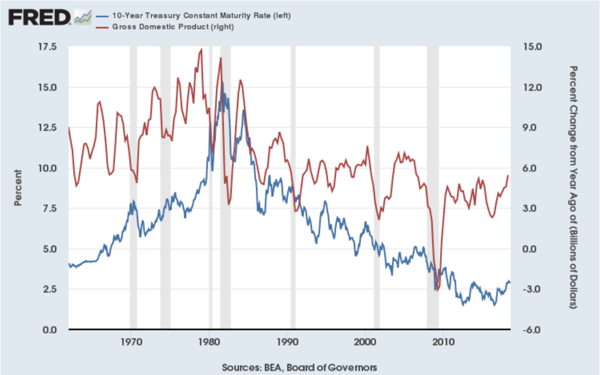
TIME SERIES ANALYSIS

M. SEO



Contents

1. Introduction

Time series data are data collected on the same observational unit at multiple time periods

- Aggregate consumption and GDP for a country (for example, 20 years of quarterly observations = 80 observations)
- Yen/\$, pound/\$ and Euro/\$ exchange rates (daily data for 1 year = 365 observations)
- Cigarette consumption per capita in California, by year (annual data)
- NASDAQ Composite Index, S&P 500, etc,

Uses of Time Series Data (e.g. the Federal Reserve Bank of St. Louis's FRED database)

- Forecasting
- Estimation of dynamic causal effects: e.g.
  - If the Fed increases the Federal Funds rate now, what will be the effect on the rates of inflation and unemployment in 3 months? In 12 months?

  - What is the effect over time on cigarette consumption of a hike in the cigarette tax?

- Modeling risks, which is used in financial markets (one aspect of this, modeling changing variances and "volatility clustering,")
- Applications outside of economics include environmental and climate modeling, engineering (system dynamics), computer science (network dynamics),...
- involves taking transformations of original series such as the differences (first, second-) $\Delta y_t = y_t - y_{t-1}$, natural logarithm $\ln y_t$, and the growth rates $100\left(\frac{\Delta y_t}{y_{t-1}}\right) = 100\left(\frac{y_t}{y_{t-1}} - 1\right)$. And

$$\frac{y_t}{y_{t-1}} - 1 \sim \log y_t - \log y_{t-1},$$

  when Δy_t is small or equivalently y_t/y_{t-1} is close to one.

- The number of observations s per year is called the frequency.
- "annualized" growth rate is the annual growth which would occur if the one-period growth rate is compounded for a full year. For a series with

frequency s the annualized growth rate is

$$100\left(\left(\frac{y_t}{y_{t-1}}\right)^s - 1\right) \sim 100s\left(\log y_t - \log y_{t-1}\right).$$

2. Stochastic Process

A sequence of observations $\{X_t, t \in \mathbb{T}\}$ indexed by time t is called a time series, where $\mathbb{T}$ is a time index set (for instance, $\mathbb{T}$=$\mathbb{Z}$, the integer set). It is often viewed as a *stochastic process* $X(t, \omega)$ with

$$X \; : \; \mathbb{T} \times \Omega \longrightarrow \mathbb{R}$$

so that $X(t, \cdot) = X_t(\cdot)$ a random variable. It follows that a realization $X(\cdot, \omega)$, $\omega \in \Omega$, becomes a numerical sequence, which is called a *sample path*.

Observations are often dependent each other. Dependency makes difficult statistical inferences based on a time series. Loosely, it implies reduced marginal information from an additional observation, and therefore the laws of large numbers and the central limit theorems may well break down. The dependence of the future on the present and past, however, makes a meaningful *prediction* feasible. Prediction is indeed one of the main themes of the time series analysis.

Example 2.1. Auto-Regressive Moving Average (ARMA) model: Introduce a lag operator L such that

$$Lx_t = x_{t-1}, L^2 x_t = L\left(Lx_t\right) = x_{t-2}, \text{ etc.}$$

Then, introduce

$$\begin{aligned}
\text{AR}(1): &\ \left(1 - \phi L\right) x_t = \varepsilon_t \\
\text{AR}(p): &\ \left(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p\right) x_t = \varepsilon_t \\
\text{MA}(1): &\ x_t = \left(1 + \theta L\right) \varepsilon_t \\
\text{MA}(q): &\ x_t = \left(1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q\right) \varepsilon_t,
\end{aligned}$$

where ε_t is a centered iid sequence. Also, define lag polynomials

$$\begin{aligned}
\phi\left(L\right) &= 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p \\
\theta\left(L\right) &= 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q
\end{aligned}$$

and rewrite an ARMA process in a more compact way:

$$\begin{aligned} \text{AR}: &\ \phi(L)\, x_t = \varepsilon_t \\ \text{MA}: &\ x_t = \theta(L)\, \varepsilon_t \\ \text{ARMA}: &\ \phi(L)\, x_t = \theta(L)\, \varepsilon_t. \end{aligned}$$

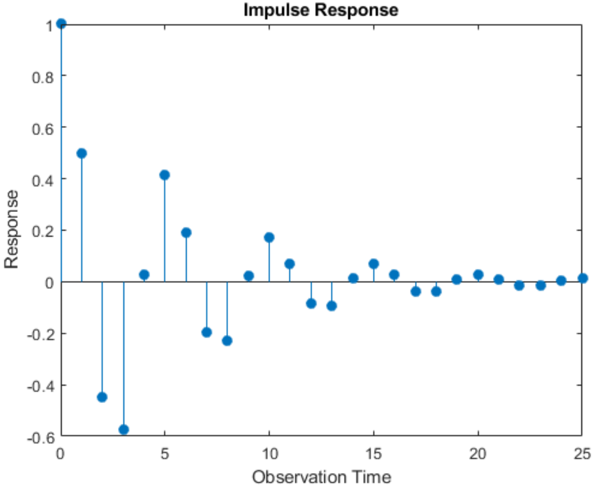
Sometimes, it is possible to rewrite an AR process to get an MA process

$$x_t = \phi(L)^{-1}\, \varepsilon_t,$$

i.e., under the *invertibility* of the lag polynomial $\phi(L)$. And this is an MA representation of x_t.

2.1. **Impulse Response.**

Example 2.2. The MA representation with shocks of unit variance is interpreted as an *impluse response function.*



2.2. **Invertibility.** Given a time series probability model, usually we can find multiple ways to represent it. Which representation to choose depends on our problem. For example, to study the impulse-response functions, MA representations maybe more convenient; while to estimate an ARMA model, AR representations maybe more convenient as usually x_t is observable while ε_t is not. However, not all ARMA processes can be inverted. In this section, we will consider under what conditions can we invert an AR model to an MA model and invert an MA model to an AR model. It turns out that *invertibility*, which means that the process can be inverted, is an important property of the model.

If we let 1 denotes the identity operator, i.e., $1y_t = y_t$, then the inversion operator $(1-\phi L)^{-1}$ is defined to be the operator so that

$$(1-\phi L)^{-1}(1-\phi L) = 1$$

For the AR(1) process, if we premulitply $(1-\phi L)^{-1}$ to both sides of the equation, we get

$$x_t = (1-\phi L)^{-1}\varepsilon_t$$

Is there any explicit way to rewrite $(1-\phi L)^{-1}$? Yes, and the answer just turns out to be $\theta(L)$ with $\theta_k = \phi^k$ for $|\phi| < 1$. To show this,

$$\begin{aligned}
& (1-\phi L)\,\theta(L) \\
=& (1-\phi L)\left(1+\theta_1 L+\theta_2 L^2+\ldots\right) \\
=& (1-\phi L)\left(1+\phi L+\phi^2 L^2+\ldots\right) \\
=& 1-\phi L+\phi L-\phi^2 L^2+\phi^2 L^2-\phi^3 L^3+\ldots \\
=& 1-\lim_{k\to\infty}\phi^k L^k \\
=& 1 \;\text{ for }\; |\phi|<1
\end{aligned}$$

We can also verify this result by recursive substitution,

$$\begin{aligned} x_t &= \phi x_{t-1} + \varepsilon_t \\ &= \phi^2 x_{t-2} + \varepsilon_t + \phi_{t-1} \\ &\vdots \\ &= \phi^k x_{t-k} + \varepsilon_t + \phi\varepsilon_{t-1} + \ldots + \phi^{k-1}\varepsilon_{t-k+1} \\ &= \phi^k x_{t-k} + \sum_{j=0}^{k-1} \phi^j \varepsilon_{t-j} \end{aligned}$$

With $|\phi| < 1$, we have that $\lim_{k\to\infty} \phi^k x_{t-k} = 0$, so again, we get the moving average representation with MA coefficient equal to ϕ^k. So the condition that $|\phi| < 1$ enables us to invert an AR(1) process to an MA(∞) process,

$$\text{AR}(1): \ (1 - \phi L)\, x_t = \varepsilon_t$$

$$\text{MA}(\infty): \ x_t = \theta(L)\, \varepsilon_t \ \text{ with } \ \theta_k = \phi^k.$$

We have got some nice results in inverting an AR(1) process to a MA(∞) process.

Lemma 2.3. *If $\{\varepsilon_t\}$ is an i.i.d. sequence with $E\left|\varepsilon_t\right| < \infty$ and $\sum_{k=0}^{\infty}\left|\theta_k\right| < \infty$, then the series*

$$x_t = \theta\left(L\right)\varepsilon_t = \sum_{k=0}^{\infty}\theta_k\varepsilon_{t-k}$$

exists a.s. And if $\sigma_{\varepsilon}^2 < \infty$, and $\sum\theta_k^2 < \infty$, then x_t exists in mean square.

Proof. Note that the monotone convergence theorem yields that

$$E\sum_{j=0}^{\infty}\left|c_j\varepsilon_{t-j}\right| = \sum_{j=0}^{\infty}E\left|c_j\varepsilon_j\right| = \mathrm{E}\left|\varepsilon_t\right| \ \sum_{j=0}^{\infty}\left|c_j\right| < \infty,$$

and thus $\sum_{j=0}^{\infty}\left|c_j\varepsilon_j\right| < \infty$ a.s. Thus, for any p

$$\left|\sum_{j=n}^{n+p}c_j\varepsilon_{t-j}\right| \leq \sum_{j=n}^{\infty}\left|c_j\varepsilon_{t-j}\right| \to 0,$$

a.s. as $n \to \infty$, which satisfies Cauchy's criterion to yield the existance of x_t a.s.

(See Appendix 3.A. in Hamilton): Recall the Cauchy criterion: a sequence $\{y_n\}$ converges in mean square if and only if $\left\|y_n - y_m\right\|_{L_2} \to 0$ as $n, m \to \infty$. In this problem, for $n > m > 0$, we want to show that

$$\begin{aligned} & E\left[\sum_{k=1}^{n} \theta_k \varepsilon_{t-k} - \sum_{k=1}^{m} \theta_k \varepsilon_{t-k}\right]^2 \\ = & \sum_{m \leq k \leq n} \theta_k^2 \sigma_\varepsilon^2 \\ \leq & \left[\sum_{k=m}^{\infty} \theta_k^2\right] \sigma_\varepsilon^2 \\ \to & 0 \ \text{ as } \ m \to \infty \end{aligned}$$

□

Furthermore, the two limits are the same (under the stronger condition of $\sum_{k=0}^{\infty} |\theta_k| < \infty$ to ensure the a.s. existance) because

$$\begin{aligned} E\left|x_t^1 - x_t^2\right| &= E\left|\lim_{n\to\infty}\sum_{j=0}^{n}\theta_j\varepsilon_{t-j} - x_t^2\right|. \\ &= E\liminf_n \left|\sum_{j=0}^{n}\theta_j\varepsilon_{t-j} - x_t^2\right| \\ &\leq \liminf_n E\left|\sum_{j=0}^{n}\theta_j\varepsilon_{t-j} - x_t^2\right| \\ &= 0, \end{aligned}$$

where we denote by x_t^1 and x_t^2 the a.s. and L_2 limits, respectively, and the first equality is definition, the second is by continuity, the inequality is due to Fatou's lemma, and the last by L_2-convergence.

Recall that absolutely summable implies square summable.

3. Dependence

3.1. **Stationarity and Ergodicity.** Roughly, stationarity implies *invariance under time shift.* In its strongest form, it implies the invariance of the distribution. Consider a time series $\{X_t\}$ and denote by $\Pr(\cdot, \ldots, \cdot)$ the joint distribution of any finite selection of X_t's.

If

$$\Pr(x_{t_1}, \ldots, x_{t_n}) = \Pr(x_{t_1+s}, \ldots, x_{t_n+s})$$

for any choice of $t_i \in T$ and s for which $t_i + s \in T$, $i = 1, \ldots, n$, then we say that $\{X_t\}$ is *strictly* (or strongly) stationary.

We may impose a weaker form of invariance, such as the invariance of the first two moments. If a time series $\{X_t\}$ has finite first two moments satisfying

$$\mathrm{E}(X_t) = \mathrm{E}(X_{t+r}) \text{ and } \mathrm{E}(X_t X_s) = \mathrm{E}(X_{t+r} X_{s+r})$$

for all $t, s \in T$ and r such that $t + r, s + r \in T$, then it is called *weakly* (or second-order or covariance) stationary.

It is clear that a strictly stationary time series with finite first two moments is weakly stationary. However, strict stationarity in general does not imply weak stationarity, since a strictly stationary time series may have no first or second moment.

The *(auto)covariance function* of a weakly stationary time series $\{X_t\}$ is

$$\gamma(k) = \operatorname{cov}(X_t, X_{t-k})$$

which is a function of only k, due to the weak sationarity. It is easy to see that $\gamma(k) = \gamma(-k)$. The *(auto)correlation function* is then given by

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

Clearly, $-1 \leq \rho(k) \leq 1$.

A stationary time series is *ergodic* if $\gamma(k) \to 0$ as $k \to \infty$, loosely speaking. See e.g. Davidson 1994 for a precise definition.

Example 3.1. (Stationary and non-ergodic) Let $X_t = U_t + Z$, where $U_t \sim i.i.d.\text{Uniform}(-1,1)$ and $Z \sim N(0,1)$. Then X_t is stationary, as each set of observations exhibits the same joint distribution. However, this process is not ergodic, because

$$\gamma_X(h) = E(X_t X_{t+h}) = 1,$$

for any h. This means that the dependence is too persistent.

Example 3.2. (Random walk) Let S_t be a random walk $S_t = \sum_{s=0}^{t} X_s$ with $S_0 = 0$ and X_t is independent and identically distributed with mean zero and variance σ^2. Then for $h > 0$,

$$\begin{aligned} Cov(S_t, S_{t+h}) &= Cov\left(\sum_{i=1}^{t} X_i, \sum_{j=1}^{t+h} X_j\right) \\ &= Var\left(\sum_{i=1}^{t} X_i\right) \quad \text{since } \ Cov(X_i, X_j) = 0 \ for \ i \neq j \\ &= t\sigma^2 \end{aligned}$$

In this case, the autocovariance function depends on time t, therefore the random walk process S_t is not stationary.

Example 3.3. (Process with linear trend): Let $\varepsilon_t \sim iid\left(0, \sigma^2\right)$ and

$$X_t = \delta t + \varepsilon_t.$$

Then $E\left(X_t\right) = \delta t$, which depends on t, therefore a process with linear trend is not stationary.

Example 3.4. (WN): The time series ε_t is said to be a white noise (WN) with variance σ_ε^2, and written as

$$\varepsilon_t \sim WN\left(0, \sigma_\varepsilon^2\right),$$

if and only if ε_t has zero mean and covariance function as

$$\gamma_\varepsilon\left(h\right) = \left\{ \begin{array}{ccc} \sigma_\varepsilon^2 & if & h = 0 \\ 0 & if & h \neq 0 \end{array} \right\}$$

It is clear that a white noise process is stationary. Note that white noise assumption is weaker than identically independent distributed assumption.

Example 3.5. (ADL model): Autoregressive Distributed Lag model contains lags of the dependent variables as well as those of explanatory variables such that

$$\phi(L)\,y_t = \theta(L)\,x_t + \varepsilon_t,$$

where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = \theta_1 z + \cdots + \theta_q z^q$. e.g. Phillips curve.

We may want to distinguish the covariance stationarity from the conditional heteroskedasticity, which may still be stationary.

Example 3.6. (Engle's (1982) ARCH model) Consider a process

$$X_t = \sigma_t \varepsilon_t,$$

where $E(\varepsilon_t) = 0$, $E\left(\varepsilon_t^2\right) = 1$, and $E(\varepsilon_t \varepsilon_s) = 0$ for $t \neq s$ and

$$\sigma_t^2 = c + \rho X_{t-1}^2,$$

where $c > 0$, $0 < \rho < 1$. In this example, the conditional variance (volatility) of X_t changes over time

$$E_{t-1}\left(X_t^2\right) = E_{t-1}\left(c + \rho X_{t-1}^2\right) = c + \rho X_{t-1}^2.$$

However, the unconditional variance of X_t is constant, which is $\sigma^2 = c/\left(1-\rho\right).$ Therefore, this process is still stationary.

The stationarity and ergodicity are preserved under general transformations. The following two theorems are important in the time series analysis and stated without proof.

Theorem 3.7. *If $\{X_t\}$ is strictly stationary and ergodic and $Y_t = f\left(X_t, X_{t-1}, ...\right)$ is a random variable for each t, then $\{Y_t\}$ is strictly stationary and ergodic.*

Example 3.8. Let X_t be a strictly stationary process with $E\left(X_t^4\right) < \infty$. Let $Y_t = \sum_{j=0}^{\infty} \theta_j X_{t-j}$, where $\sum_{j=0}^{\infty} |\theta_j| < \infty$. Then Y_t is a strictly stationary process by Lemma 2.3 and Theorem 3.7. And $E\left|Y_t Y_s Y_i Y_j\right| < \infty$ for all t, s, i, and j.

We have the same LLN, known as the Ergodic theorem, as in iid case under the stationarity and ergodicity.

Theorem 3.9 (Ergodic Theorem)**.** *If* $\{X_t\}$ *is strictly stationary and ergodic and* $\mathrm{E}\,|X_t| < \infty$, *then as* $n \to \infty$

$$\frac{1}{n}\sum_{t=1}^{n} X_t \xrightarrow{p} \mathrm{E}\,(X_1)\,.$$

The ergodicity enables the LLN for time series data. Also,

Theorem 3.10. *An AR(p) model is stationary and ergodic if all roots of lag polynomial lie outside unit circle.*

We will often have to deal with a multiple time series $\{X_t\}$, where

$$X_t = (X_{1t}, \ldots, X_{mt})'.$$

All the concepts introduced above readily extend to this case of vector processes with some obvious modifications. In particular, the covariance function of a vector process $\{X_t\}$ whose mean is zero is given by

$$\Gamma(k) = \mathrm{E}\, X_t X_{t-k}'$$

In contrast to the scalar case, it is anti-symmetric, i.e.,

$$\Gamma(-k) = \Gamma(k)'$$

as one may easily see. The correlation function is given by

$$R(k) = D^{-1/2}\Gamma(k)D^{-1/2}$$

where D is a diagonal matrix consisting of the diagonal elements $\gamma_{ii}(0)$'s of $\Gamma(0)$.

The autocovariance functions can be estimated by the method of moments and their consistency can be shown by the ergodic theorem.

3.2. **Wold Decomposition.**

Theorem 3.11. *(Wold Decomposition) Any zero-mean covariance stationary process X_t can be reprepresented in the form*

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + V_t$$

where (*i*) $\psi_0 = 1$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ (*ii*) $\varepsilon_t \sim WN\left(0, \sigma_\varepsilon^2\right)$ (*iii*) $E\left(\varepsilon_t V_s\right) = 0, \ \forall s, t > 0$, (*iv*) ε_t *is the error in forecasting* X_t *on the basis of a linear function of lagged* X, (*v*) V_t *is a deterministic process and it can be predicted from a linear function of lagged* X*'s.*

3.3. **Martingales.**

Definition 3.12. Let $\{\Omega, \mathcal{F}, P\}$ be a probability space. Let $\mathcal{F}_n$ be an increasing sequence of sub-σ-fields of $\mathcal{F}$ and S_n is measurable wrt $\mathcal{F}_n$. Then, $\{S_n, \mathcal{F}_n\}$ is a martingale if $E\left|S_n\right| < \infty$ and

$$E\left(S_n|\mathcal{F}_{n-1}\right) = S_{n-1}. \tag{3.1}$$

And $\varepsilon_n = S_n - S_{n-1}$ is called a martingale difference sequence (MDS).

The condition (3.1) is equivalent to

$$E\left(S_{n+m}|\mathcal{F}_{n-1}\right) = S_{n-1}, \quad \text{for any } m \geq 0. \tag{3.2}$$

If $\mathcal{F}_n = \sigma\left(S_t, -\infty < t \leq n\right)$, i.e., $\mathcal{F}_n$ generated by current and all lagged observations of S_n, then it is called the natural filteration of S_n.
A simple example of martingale is a random walk.

Example 3.13. (Random walk) Let

$$S_t = S_{t-1} + \varepsilon_t, \;\; S_0 = 0, \;\; \varepsilon_t \sim i.i.d.\left(0, \sigma^2\right)$$

$$\mathcal{F}_t = \sigma\left\{\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_1\right\}.$$

Then we know that S_t is a martingale as $E\left|S_t\right| \leq \sum_{k=1}^{t} E\left|\varepsilon_k\right| < \infty$ and $E\left(S_t \mid \mathcal{F}_{t-1}\right) = S_{t-1}$.

Let $\{S_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be an adapted sequence. Two concepts that are related to martingales are *submartingales*, which means $S_t \leq E\left(S_{t+1} \mid \mathcal{F}_t\right)$ and *supermartingales*, which means $E\left(S_{t+1} \mid \mathcal{F}_t\right) \leq S_t$. An example of submartingale is the convex transformations of martingales, that is, by Jensen's inequality,

$$E\left(g\left(S_t\right)|\mathcal{F}_{t-1}\right) \geq g\left(E\left(S_t|\mathcal{F}_{t-1}\right)\right) = g\left(S_{t-1}\right),$$

where S_t is martingale with $\mathcal{F}_t$ and $g\left(\cdot\right)$ is a convex function. This also shows a well-known feature of the martingale that

$$E\left|S_t\right|^p \leq E\left(E\left(\left|S_{t+1}\right|^p|\mathcal{F}_t\right)\right) = E\left(\left|S_{t+1}\right|^p\right),$$

for $p \geq 1$. And thus,

$$\left\|S_t\right\|_p \leq \left\|S_{t+1}\right\|_p.$$

Being an mds is a stronger condition than being serially uncorrelated. If X_t is an mds, then we cannot forecast X_t as a linear or nonlinear function of its past realizations. However, it is a weaker condition than independence, since it does not rule out the possibility that higher moments such as $E\left(X_t^2 \mid \mathcal{F}_{t-1}\right)$ depends on lagged value of X_t. For instance, let $\varepsilon_t \sim i.i.d.\left(0, \sigma^2\right)$, then $X_t = \varepsilon_t \varepsilon_{t-1}$ is a an mds but not serially independent.

3.4. **Mixing.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\mathcal{G}$, $\mathcal{H}$ be σ subfields of $\mathcal{F}$, define

$$\alpha(\mathcal{G},\mathcal{H}) = \sup_{G\in\mathcal{G}, H\in\mathcal{H}} |P(G\cap H) - P(G)P(H)|, \tag{3.3}$$

$$\beta(\mathcal{G},\mathcal{H}) = \frac{1}{2}\sup_{I,J} \sum_{i\in I, j\in J} |P(G_i\cap H_j) - P(G_i)P(H_j)|, \tag{3.4}$$

where sup in (3.4) is taken over all finite partitions $(G_i)_{i\in I}$ and $(H_j)_{j\in J}$ that are $\mathcal{G}, \mathcal{H}$ measurable, and

$$\phi(\mathcal{G},\mathcal{H}) = \sup_{G\in\mathcal{G}, H\in\mathcal{H}; P(G)>0} |P(H \mid G) - P(H)|. \tag{3.5}$$

They are called *strong mixing, absolute regular, uniform mixing coefficient,* respectively. It is known that

$$2\alpha(\mathcal{G},\mathcal{H}) \leq \beta(\mathcal{G},\mathcal{H}) \leq \phi(\mathcal{G},\mathcal{H}),$$

e.g. Doukhan (1994).

The events in $\mathcal{G}$ and $\mathcal{H}$ are independent iff α and ϕ are zero.

For a sequence, $\{X_t\}_{-\infty}^{\infty}$, let

$$\mathcal{F}_{-\infty}^{t} = \sigma\left(\ldots X_{t-1}, X_t\right), \mathcal{F}_{t+m}^{\infty} = \sigma\left(X_{t+m}, X_{t+m+1}, \ldots\right).$$

Define the mixing coefficients

$$\alpha_m = \sup_t \alpha\left(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}\right), \quad \text{and} \quad \beta_m = \sup_t \beta\left(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}\right),$$

and the uniform mixing coefficient

$$\phi_m = \sup_t \phi\left(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}\right).$$

Then, the sequence is said to be α-mixing (strong mixing), β-mixng, and ϕ-mixing (uniform mixing) if $\lim_{m\to\infty} \alpha_m = 0$, $\lim_{m\to\infty} \beta_m = 0$, $\lim_{m\to\infty} \phi_m = 0$, respectively. Since $\alpha \leq \phi$, ϕ-mixing implies α-mixing.

Also, mixing coefficients get smaller for $g\left(X_t\right)$ for any g.

A mixing sequence is not necessarily stationary, and it could be hetergeneous.

A sequence is said to be α-mixing of size $-\gamma_0$ if $\alpha_m = O\left(m^{-\gamma}\right)$ for some $\gamma > \gamma_0$. If X_t is a α-mixing sequence of size $-\gamma_0$, and if $Y_t = g\left(X_t, X_{t-1}, \ldots, X_{t-k}\right)$ is a measurable function and k be finite, then Y is also α-mixing of size $-\gamma_0$. The same for other mixings.

When a sequence is stationary and mixing, then $Cov\left(X_1, X_m\right) \to 0$ as $m \to \infty$. Consider the ARMA processes. If it is MA(q), then the process must be mixing since any two events with time interval larger than q are independent, i.e., $\alpha\left(m\right) = \phi\left(m\right) = 0$ for $m > q$. Sufficient conditions for a MA(∞) to be strong or uniform mixing is more demanding but if the innovations are *i.i.d.* Gaussian, then absolute summability of the moving average coefficients yields strong mixing.

3.5. **Mixingales.** Mixingale is like asymptotic unpredictability in the sense of mds. It is a generalization of mds.

Definition 3.14. A sequence of random variables $\{X_t\}$ with $E(X_t) = 0$ is called a L^p mixingale ($p \geq 1$) with respect to $\{\mathcal{F}_t\}$ if there are sequences of nonnegative constants c_t and ξ_m, where $\xi_m \to 0$ as $m \to \infty$, such that

$$\|E(X_t \mid \mathcal{F}_{t-m})\|_p \leq c_t \xi_m \tag{3.6}$$

$$\|X_t - E(X_t \mid \mathcal{F}_{t+m})\|_p \leq c_t \xi_{m+1} \tag{3.7}$$

for all $t \geq 1$ and $m \geq 0$.

Condition (3.6) can then be written as

$$E|E(X_t \mid \mathcal{F}_{t-m})| \leq c_t \xi_m. \tag{3.8}$$

for $p = 1$. An mds is a special kind of mixingale and you can set $c_t = E|X_t|$ and set $\xi_0 = 1$ and $\xi_m = 0$ for $m \geq 1$. The mixingale X_t needs not be adapted to $\mathcal{F}_t$ as apparent in (3.7).

Example 3.15. Consider a linear process,

$$X_t = \sum_{j=-\infty}^{\infty} \theta_j \varepsilon_{t+j},$$

where ε_t is a stationary mds with $E\left|\varepsilon_t\right| < \infty$ and $\mathcal{F}_t$ is the natural filtration for ε_t. Then

$$E\left(X_t \mid \mathcal{F}_{t-m}\right) = \sum_{j=m}^{\infty} \theta_{-j} \varepsilon_{t-j}.$$

Take $c_t = E\left|\varepsilon_t\right|$ and take $\xi_m = \sum_{j=m}^{\infty} \left|\theta_{-j}\right|$. Then if the moving average coefficients are absolutely summable, i.e., $\sum_{j=-\infty}^{\infty} \left|\theta_j\right| < \infty$, then its tails has to go to zero, i.e., $\xi_m \to 0$. Then condition (3.8) is satisfied. Thus, X_t is an L_1-mixingale.

Unlike the strict stationarity and ergodicity, mixingale property is not preserved under transformation.

3.6. Near-epoch Dependent (NED).

Definition 3.16. For a stochastic sequence $\{V_t\}_{-\infty}^{\infty}$, possibly vector-valued, on a probability space $(\Omega, \mathscr{F}, P)$, let $\mathscr{F}_{t-m}^{t+m} = \sigma(V_{t-m}, \ldots, V_{t+m})$, such that $\{\mathscr{F}_{t-m}^{t+m}\}_{m=0}^{\infty}$ is an increasing sequence of σ-fields. If, for $p > 0$, a sequence of integrable r.v.s $\{X_t\}_{-\infty}^{+\infty}$ satisfies

$$\left\| X_t - E\left(X_t \mid \mathscr{F}_{t-m}^{t+m}\right)\right\|_p \leq d_t \nu_m, \tag{3.9}$$

where $\nu_m \to 0$, and $\{d_t\}_{-\infty}^{+\infty}$ is a sequence of positive constants, X_t will be said to be *near-epoch dependent in L_p-norm* (L_p-NED) on $\{V_t\}_{-\infty}^{+\infty}$. □

Many results in this literature are proved for the case $p = 2$ (Gallant and White, 1988, for example).

We will say that the sequence or array is L_p-NED of size $-\Phi_0$ when $\nu_m = O\left(m^{-\Phi}\right)$ for $\Phi > \Phi_0$. According to the Minkowski and conditional modulus inequalities,

$$\begin{aligned}\left\|X_t - E\left(X_t \mid \mathscr{F}_{t-m}^{t+m}\right)\right\|_p &\leq \|X_t - \mu_t\|_p + \left\|E\left(X_t - \mu_t \mid \mathscr{F}_{t-m}^{t+m}\right)\right\|_p \\ &\leq 2\|X_t - \mu_t\|_p,\end{aligned} \tag{3.10}$$

where $\mu_t = E(X_t)$. The role of the sequence $\{d_t\}$ in (3.9) is usually to account for the possibility of trending moments, and when $\|X_t - \mu_t\|_p$ is uniformly bounded, we should expect to set d_t equal to a finite constant for all t.

The usefulness of the near-epoch dependence concept is due largely to the next theorem and its permanance property thereafter.

Theorem 3.17. *Let $\{X_t\}_{-\infty}^{\infty}$ be an L_r-bounded zero-mean sequence, for $r > 1$.*

(i) Let $\{V_t\}$ be α-mixing of size $-a$. If X_t is L_p-NED of size $-b$ on $\{V_t\}$ for $1 \leq p < r$ with constants $\{d_t\}$, $\left\{X_t, \mathscr{F}_{-\infty}^{t}\right\}$ is an L_p-mixingale of size $-\min\{b, a(1/p - 1/r)\}$ with constants $c_t \ll \max\{\|X_t\|_r, d_t\}$.

(ii) Let $\{V_t\}$ be φ-mixing of size $-a$. If X_t is L_p-NED of size $-b$ on $\{V_t\}$ for $1 \leq p \leq r$ with constants $\{d_t\}$, $\left\{X_t, \mathscr{F}_{-\infty}^{t}\right\}$ is an L_p-mixingale of size $-\min\{b, a(1 - 1/r)\}$ with constants $c_t \ll \max\{\|X_t\|_r, d_t\}$.

Some permanance properties are collected below.

Theorem 3.18. *Let X_t and Y_t be L_p-NED on $\{V_t\}$ of respective sizes $-\Phi_X$ and $-\Phi_Y$. Then $X_t + Y_t$ is L_p-NED of size $-\min\{\Phi_X, \Phi_Y\}$.*

Recall that a variable that is L_q-NED is L_p-NED for $1 \leq p \leq q$.

Theorem 3.19. *Let X_t and Y_t be L_2-NED on $\{V_t\}$ of respective sizes $-\Phi_X$ and $-\Phi_Y$. Then, X_tY_t is L_1-NED of size $-\min\{\Phi_Y, \Phi_X\}$.*

For the preceding results we would like to be able to set $Y_t = X_{t+j}$ for some finite j.

Theorem 3.20. *If X_t is L_p-NED on $\{V_t\}$, so is X_{t+j} for $0 < j < \infty$.*

Putting the last two results together gives the following corollary.

Corollary 3.21. *If X_t and Y_t are L_2-NED of size $-\Phi_X$ and $-\Phi_Y$, X_tY_{t+k} is L_1-NED of size $-\min\{\Phi_Y, \Phi_X\}$.*

By considering $Z_t = X_{t-[k/2]}Y_{t+k-[k/2]}$, the L_1-NED numbers can be given here as

$$\nu_m^{'} = \begin{cases} \nu_0, & m \leq [k/2]+1 \\ \nu_{m-[k/2]-1}, & m > [k/2]+1 \end{cases}$$

where $\nu_m = \nu_m^Y + \nu_m^X + \nu_m^Y \nu_m^X$, and the constants are $4d_{t\ \ [k/2]}^X d_{t+k\ \ [k/2]}^Y$,assuming that d_t^X and d_t^Y are not smaller than the corresponding L_2 norms.

All these results extend to the array case as before, by simply including the extra subscript throughout.

Consider a uniform Lipschitz condition,

$$\left|\varphi_t\left(X^1\right) - \varphi_t\left(X^2\right)\right| \leq B_t \left|X^1 - X^2\right|_1 \text{ a.s.,} \tag{3.11}$$

where B_t is a finite constant.

Theorem 3.22. *Let X_{nt} be L_2-NED of size $-a$ on $\{V_t\}$ for $n = 1, \ldots, \nu$, with constants d_{nt}. If (3.11) holds, $\{\varphi_t(X_{nt})\}$ is also L_2-NED on $\{V_t\}$ of size $-a$, with constants a finite multiple of* $\max_n \{d_{nt}\}$.

4. Limit Theorems

4.1. **Prelimiaries.** Consider a centered covariance stationary process $\{X_n\}$, where $|\gamma(h)| < \infty$. Let us begin with its sample mean

$$\bar{X_n} = \frac{1}{n}\left(X_1 + \ldots + X_n\right).$$

It is unbiased, $E\left(\bar{X_n}\right) = E\left(X_t\right) = 0$, and

$$\begin{aligned}
\operatorname{var}\left(\bar{X_n}\right) &= \left(1/n^2\right)\sum_{i,j=1}^{n} E\left(X_i X_j\right) \\
&= \left(1/n^2\right)\sum_{i,j=1}^{n} \gamma_x\left(i-j\right) \\
&= \left(1/n\right)\left(\gamma_0 + 2\sum_{h=1}^{n-1}\left(1-\frac{h}{n}\right)\gamma\left(h\right)\right).
\end{aligned}$$

Note that

$$\begin{aligned} n\,\mathrm{var}\left(\bar{X_n}\right) &= \gamma\left(0\right)+\left(1-\frac{1}{n}\right)2\gamma\left(2\right)+\left(1-\frac{2}{n}\right)2\gamma\left(3\right)+\ldots+\left(1-\frac{m}{n}\right)2\gamma\left(m\right)+\ldots \\ &\leq \left|\gamma\left(0\right)\right|+2\left|\gamma\left(1\right)\right|+2\left|\gamma\left(2\right)\right|+\ldots \end{aligned}$$

If we assume the *absolute summability* of $\gamma\left(h\right)$,

$$\bar{X_n}\to 0,$$

in L_2, and

$$\lim_{n\to\infty} n\,\mathrm{var}\left(\bar{X_n}\right)=\sum_{h=-\infty}^{\infty}\gamma\left(h\right)=\gamma\left(0\right)+2\gamma\left(1\right)+2\gamma\left(2\right)+\cdots$$

exists by Kronecker lemma.

We summarize our results in the following theorem.

Theorem 4.1. *Let X_t be a zero-mean covariance stationary process with $E\left(X_tX_{t-h}\right)=\gamma\left(h\right)$ and absolutely summable autocovariances. Then,*

$$\bar{X_n}\to 0 \quad in \quad L_2,$$

and

$$\lim_{n\to\infty} nE\left(\bar{X}_n^2\right) = \sum\nolimits_{h=-\infty}^{\infty} \gamma(h).$$

If the process has population mean μ, then accordingly we have $\bar{X}_n \xrightarrow{p} \mu$ and the limit of $nE\left(\bar{X}_n - \mu\right)^2$ remains the same.

4.2. **Convergences for MG.**

We begin with introducing some basic inequalities. The first is an analogue of Kolmogorov inequality.

Theorem 4.2. *If $\{S_t, \mathscr{F}_t, 1 \le t \le n\}$ is a submartingale, then for each real λ,*

$$\lambda P\left(\max_{t\le n} S_t > \lambda\right) \le E\left[S_n I\left(\max_{t\le n} S_t > \lambda\right)\right].$$

Proof. Define

$$B = \left\{\max_{t\le n} S_t > \lambda\right\} = \bigcup_{t=1}^{n} \left\{S_t > \lambda; \max_{1\le j<t} S_t \le \lambda\right\} = \bigcup_{t=1}^{n} B_t,$$

say. The events B_t are $\mathscr{F}_t$-measurable and disjoint. Then

$$\begin{aligned}
\lambda P(B) &\leq \sum_t E[S_t I(B_t)] \\
&\leq \sum_t E[E(S_n \mid \mathscr{F}_t) I(B_t)] \quad \text{(submartingale property)} \\
&= \sum_t E[E(S_n I(B_t) \mid \mathscr{F}_t)] \\
&= \sum_t E[S_n I(B_t)] \\
&= E[S_n I(B)].
\end{aligned}$$

□

If $\{S_t, 1 \leq t \leq n\}$ is a martingale, then $\{|S_t|^p, 1 \leq t \leq n\}$ is a submartingale with $p \geq 1$. By applying the ablve to this submartingale we obtain

Corollary 4.3. *If $\{S_t, \mathscr{F}_t, 1 \le t \le n\}$ is a martingale, then for each $p \ge 1$ and $\lambda > 0$,*

$$\lambda^p P\left(\max_{t \le n} |S_t| > \lambda\right) \le E\left|S_n\right|^p .$$

It has an application in another direction, which yields the following result.

Theorem 4.4. *(Doob's inequality) If $\{S_t, \mathscr{F}_t, 1 \le t \le n\}$ is a martingale, then for $p > 1$,*

$$\left\|S_n\right\|_p \le \left\|\max_{t \le n} |S_t|\right\|_p \le q \left\|S_n\right\|_p ,$$

where $p^{-1} + q^{-1} = 1$.

Proof. The left-hand inequality is trivial. To prove the right-hand one we note that by integral-by-parts, Theorem 4.2 and Hölder's inequality,

$$\begin{aligned}
E\left(\max_{t\leq n}|S_t|^p\right) &= p\int_0^\infty x^{p-1}P\left(\max_{t\leq n}|S_t|>x\right)dx \\
&\leq p\int_0^\infty x^{p-2}E\left[|S_n|\,I\left(\max_{t\leq n}|S_t|>x\right)\right]dx \\
&= pE\left[|S_n|\int_0^{\max_{t\leq n}|S_t|}x^{p-2}dx\right] \\
&= qE\left[|S_n|\left(\max_{t\leq n}|S_t|^{p-1}\right)\right] \\
&\leq q\left(E|S_n|^p\right)^{1/p}\left(E\left(\max_{t\leq n}|S_t|^p\right)\right)^{1/q},
\end{aligned}$$

which gives the desired result. □

For $-\infty<a<b<\infty$, let $\nu=\nu(a,b,n)$ denote the number of times that the sequence $\{S_t, 1\leq t\leq n\}$ crosses from a value $\leq a$ to one $\geq b$; ν is called the *number of upcrossings* of $[a,b]$ by $\{S_t\}$. The following upcrossing inequality provides a bound for the mean number of the upcrossings by the submartingale $\{S_t, \mathscr{F}_t, 1 \leq t \leq n\}$ across an interval $[a, b]$.

Theorem 4.5. *Let ν denote the number of upcrossings of the compact interval $[a, b]$ by the submartingale $\{S_t, \mathscr{F}_t, 1 \leq t \leq n\}$. Then*

$$(b-a) E(\nu) \leq E(S_n - a)^+ - E(S_1 - a)^+ .$$

The next result is a useful alternative to Theorem 4.2 which utilizes the upcrossing inequality.

Theorem 4.6. *If $\{S_t, \mathscr{F}_t, 1 \leq t \leq n\}$ is a zero-mean martingale, then for each $\lambda > 0$,*

$$\begin{aligned} \lambda P\left(\max_{t \leq n} |S_t| > 2\lambda\right) &\leq \lambda P(|S_n| > \lambda) + E[(|S_n| - 2\lambda) I(|S_n| \geq 2\lambda)] \\ &\leq E[|S_n| I(|S_n| > \lambda)] . \end{aligned}$$

One of the fundamental results in martingale theory is the martingale convergence theorem, aka Doob's convergence theorem.

Theorem 4.7. *(Doob's Convergence Theorem) If $\{S_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ is an L_1-bounded submartingale, then $S_n \rightarrow_{a.s.} S$, where $E|S| < \infty$. If $\{S_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ is an L_p-bounded martingale for some $p > 1$, then S_n converges in L_p as well as a.s.*

Proof. We first show that

$$P\{\limsup S_n = \liminf S_n\} = 1,$$

or equivalently

$$P\{\liminf S_n < a < b < \limsup S_n\} = 0$$

for any $a < b$. Suppose not. Then, for some $a < b$ the number of upcrossing of the interval $[a, b]$ would be infinity with positive probability. However, the expected number of the upcrossing is finite for L_1 bounded submartingales, yielding contradiction.

Next, by Fatou's lemma, $E(\lim |S_n|) \leq \liminf E(|S_n|) < \infty$. □

This is an existence theorem and it tells that S_n converges to S, but it does not tell what S is. But martingale convergence theorem is still a very powerful result.

Example 4.8. Let $\varepsilon_t \sim mds\left(0, \sigma_t^2\right)$ with $\sup_t \sigma_t^2 = M < \infty$. Define

$$S_n = \sum\nolimits_{t=1}^{n} \frac{\varepsilon_t}{t},$$

then S_n is a martingale with $E\left(S_n^2\right) = \sum\nolimits_{t=1}^{n} \sigma_t^2/t^2$. Since $\sup_n E\left(\left|S_n\right|^2\right) \leq \sup_t \sigma_t^2 \sum\nolimits_{t=1}^{n} \left(1/t^2\right) < \infty$, $S_n = \sum\nolimits_{t=1}^{n} \varepsilon_t/t$ converges a.s. by Theorem 4.7. Furthermore,

$$\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t \overset{a.s.}{\to} 0,$$

by Kronecker Lemma as well as in L_2 since

$$E\left(\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t\right)^2 = \frac{1}{n^2}\sum_{t=1}^{n} \sigma_t^2 \leq \frac{M}{n} \to 0.$$

4.3. **LLN for MDS.**

4.3.1. *WLLN.*

Theorem 4.9. *Let* $\{S_n = \sum_{t=1}^n X_t, \mathscr{F}_n, n \geq 1\}$ *be a martingale and* $\{b_n\}$ *a sequence of positive constants with* $b_n \uparrow \infty$ *as* $n \to \infty$*. Then, writing* $X_{nt} = X_t I\left(|X_t| \leq b_n\right), 1 \leq t \leq n$*, we have that* $b_n^{-1} S_n \overset{\text{p}}{\to} 0$ *as* $n \to \infty$ *if*

(i) $\sum_{t=1}^n P\left(|X_t| > b_n\right) \to 0,$
(ii) $b_n^{-1} \sum_{t=1}^n E\left(X_{nt} \mid \mathscr{F}_{t-1}\right) \overset{\text{p}}{\to} 0,$ *and*
(iii) $b_n^{-2} \sum_{t=1}^n \left\{E X_{nt}^2 - E\left[E\left(X_{nt} \mid \mathscr{F}_{t-1}\right)\right]^2\right\} \to 0.$

To derive the law of large numbers for L_1-mixingales, we need the notion of *uniformly integrable.*

Definition 4.10. *A sequence* $\{X_t\}$ *is said to be uniformly integrable if for every* $\varepsilon > 0$ *there exists a number* $c > 0$ *for all* t *such that*

$$E\left(|X_t| \mathbf{1}\left\{|X_t| > c\right\}\right) < \varepsilon$$

Some sufficient conditions are given below.

Lemma 4.11. *A sequence $\{X_t\}$ is uniformly integrable if there is $r > 1$ such that $E\left(|X_t|^r\right) < M$ for all t.*

Example 4.12. Let $\{S_n = \sum_{t=1}^n X_t, \mathscr{F}_n, n \geq 1\}$ be a martingale and Suppose that the sequence $\{|X_n|^p, n \geq 1\}$ is uniformly integrable for $1 \leq p < 2$. Then, $n^{-1/p} S_n \xrightarrow{\text{p}} 0$.

We shall verify the conditions of Theorem 4.9, with $b_n = n^{1/p}$, to prove that $n^{-1/p} S_n \xrightarrow{\text{p}} 0$ as $n \to \infty$.

First, note that

$$\sum_{t=1}^n P\left(|X_t| > n^{1/p}\right) \leq n^{-1} \sum_{t=1}^n E\left[|X_t|^p I\left(|X_t| > n^{1/p}\right)\right] \to 0,$$

due to the uniform integrability. This verifies (i).

Second, as X_t is mds,

$$\begin{aligned} & n^{-1/p} E\left\{\sum_{t=1}^{n} \left|E\left[X_{nt} \mid \mathscr{F}_{t-1}\right]\right|\right\} \\ = \quad & n^{-1/p} E\left\{\sum_{t=1}^{n} \left|E\left[X_t I\left(|X_t| > n^{1/p}\right) \mid \mathscr{F}_{t-1}\right]\right|\right\} \\ \leq \quad & n^{-1} \sum_{t=1}^{n} E\left[|X_t|^p \, I\left(|X_t| > n^{1/p}\right)\right] \to 0, \end{aligned}$$

where we apply Jensen's inequality, the law of iterated expectation, and Markov inequality for the inequality and then the uniform integrability for the convergence. This verifies condition (ii).

Condition (iii) will hold if we show that

$$n^{-2/p} \sum_{t=1}^{n} E\left[X_t^2 I\left(|X_t| \leq n^{1/p}\right)\right] \to 0.$$

For any $0 < \epsilon < 1$, the left-hand side above equals

$$
\begin{aligned}
& n^{-2/p} \sum_{t=1}^{n} E\left[X_t^2 I\left(|X_t| \leq \epsilon n^{1/p}\right) + X_t^2 I\left(\epsilon n^{1/p} \leq |X_t| \leq n^{1/p}\right)\right] \\
\leq & n^{-2/p}\left(\epsilon n^{1/p}\right)^{2-p} \sum_{t=1}^{n} E\left|X_t\right|^p + n^{-2/p}\left(n^{1/p}\right)^{2-p} \sum_{t=1}^{n} E\left[\left|X_t\right|^p I\left(|X_t| > \epsilon n^{1/p}\right)\right] \\
\leq & \epsilon^{2-p}\left(\max_{t \leq n} E\left|X_t\right|^p\right) + n^{-1} \sum_{t=1}^{n} E\left[\left|X_t\right|^p I\left(|X_t| > \epsilon n^{1/p}\right)\right],
\end{aligned}
$$

which can be made arbitrarily small by choosing ϵ sufficiently small and then n sufficiently large. Therefore (i), (ii), and (iii) hold, and so $n^{-1/p} S_n \xrightarrow{\text{p}} 0$.

In fact, a stronger result holds.

Theorem 4.13. *Let* $\{S_n = \sum_{t=1}^{n} X_t, \mathscr{F}_n, n \geq 1\}$ *be a martingale and let* $1 \leq p < 2$. *If* $\{|X_n|^p, n \geq 1\}$ *is uniformly integrable, then* $n^{-1} E|S_n|^p \to 0$.

Also,

Theorem 4.14. *Let $\{X_t\}$ be an L_1-mixingale with corresponding constants $\{c_t\}$. If $\{X_t\}$ is uniformly integrable and*

$$\lim_{n\to\infty} (1/n) \sum_{t=1}^{n} c_t < \infty,$$

then $n^{-1} S_n \xrightarrow{p} 0$.

Example 4.15. Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be an mds with $E\left|\varepsilon_t\right|^r < M$ for some $r > 1$ and $M < \infty$ (i.e. ε_t is L_r-bounded). Let $X_{nt} = (t/n)\,\varepsilon_t$. Then, $\{X_{nt}\}$ is a uniformly integrable L_1-mixingale with $c_{nt} = \sup_t E\left|\varepsilon_t\right|$, $\xi_0 = 1$ and $\xi_m = 0$ for $m > 0$. Then applying LLN for L_1-mixingales, we have $\bar{X}_n \xrightarrow{p} 0$.

4.3.2. *SLLN.* The SLLN demands stronger conditions.

Theorem 4.16. *Let $\{S_n = \sum_{t=1}^{n} X_t, \mathscr{F}_n, n \geq 1\}$ be a zero-mean, square-integrable martingale. Then S_n converges a.s. on the set $\left\{\sum_{t=1}^{\infty} E\left(X_t^2 \mid \mathscr{F}_{t-1}\right) < \infty\right\}$.*

Theorem 4.17. *Let* $\{S_n = \sum_{t=1}^{n} X_t, n \geq 1\}$ *be a sequence of r.v. and* $\{\mathscr{F}_n, n \geq 1\}$ *an increasing sequence of* σ*-fields such that* S_n *is* $\mathscr{F}_n$*-measurable. Let* c *be a positive constant. Then* S_n *converges a.s. on the set where*

(*i*) $\sum_{t=1}^{\infty} P\left(|X_t| \geq c \mid \mathscr{F}_{t-1}\right) < \infty,$
(*ii*) $\sum_{t=1}^{\infty} E\left[X_t I\left(|X_t| \leq c\right) \mid \mathscr{F}_{t-1}\right]$ *converges, and*
(*iii*) $\sum_{t=1}^{\infty} \left\{E\left[X_t^2 I\left(|X_t| \leq c\right) \mid \mathscr{F}_{t-1}\right] - \left[E\left(X_t I\left(|X_t| \leq c\right) \mid \mathscr{F}_{t-1}\right)\right]^2\right\} < \infty$ *hold.*

The following LLN (McLeish (1975)) applies to mixing sequences.

Theorem 4.18. *Let* $\{X_t\}$ *be strong mixing with size* $-r/\left(r-1\right)$ *for some* $r > 1$, *with finite means* $\mu_t = E\left(X_t\right).$ *If for some* δ, $0 < \delta \leq r$,

$$\sum_{t=1}^{\infty} \left(\frac{E\left|Z_t - \mu_t\right|^{r+\delta}}{t^{r+\delta}}\right)^{1/r} < \infty, \tag{4.1}$$

then $\bar{X_n} - \bar{\mu_n} \rightarrow_{a.s.} 0.$

Note the trade-off between the mixing size and moment requirement. And for a mixingale,

Theorem 4.19. *If $\{X_n, \mathscr{F}_n\}$ is a mixingale and $\{b_n\}$ is a sequence of positive constants increasing to ∞ such that*

$$\sum_{n=1}^{\infty} b_n^{-2} c_n^2 < \infty \text{ and } \xi_n = O\left(n^{-1/2} (\log n)^{-2}\right) \text{ as } n \to \infty,$$

then

$$b_n^{-1} \sum_{t=1}^{n} X_t \overset{\text{a.s.}}{\to} 0.$$

4.4. **CLT for MDS.**

Definition 4.20. If $\{Y_n\}$ is a sequence of r.v. on a probability space $(\Omega, \mathscr{F}, P)$ converging in distribution to an r.v. Y, we say that the convergence is *stable* if for all continuity points y of Y and all events $E \in \mathscr{F}$, the limit

$$\lim_{n\to\infty} P\left(\{Y_n \le y\} \cap E\right) = Q_y\left(E\right)$$

exists, and if $Q_y\left(E\right) \to P\left(E\right)$ as $y \to \infty$. (Clearly Q_y, if it exists, is a probability measure on $(\Omega, \mathscr{F})$.) We designate the convergence by writing

$$Y_n \overset{\mathrm{d}}{\to} Y \quad \text{(stably)}.$$

Theorem 4.21. *Let $\{S_{nt}, \mathscr{F}_{nt}, 1 \le t \le k_n, n \ge 1\}$ be a zero-mean, square-integrable martingale array with differences X_{nt}, and let η^2 be an a.s. finite r.v. Suppose that*

$$\max_t |X_{nt}| \overset{\mathrm{p}}{\to} 0, \tag{4.2}$$

$$\sum_t X_{nt}^2 \xrightarrow{\mathrm{p}} \eta^2, \tag{4.3}$$

$$E\left(\max_t X_{nt}^2\right) \quad \textit{is bounded in } n, \tag{4.4}$$

and

$$\mathscr{F}_{n,t} \subseteq \mathscr{F}_{n+1,t} \quad for \quad 1 \le t \le k_n, \quad n \ge 1. \tag{4.5}$$

Then,

$$S_{nk_n} = \sum_t X_{nt} \xrightarrow{d} Z \quad (\textit{stably}),$$

where the r.v. Z has characteristic function $E\exp\left(-\frac{1}{2}\eta^2 t^2\right)$.

Note that

$$\begin{aligned} P\left\{\max_t |X_t| > \varepsilon\right\} &= P\left\{\sum_t |X_t| \, 1\left\{|X_t| > \varepsilon\right\} > \varepsilon\right\} \\ &= P\left\{\sum_t X_t^2 1\left\{|X_t| > \varepsilon\right\} > \varepsilon^2\right\}, \end{aligned}$$

which motivates an alternative version:

Corollary 4.22. *If (4.2) and (4.4) are replaced by the conditional Lindeberg condition:*

$$for\ \ all\ \ \varepsilon > 0,\ \ \sum_t E\left[X_{nt}^2 1\left(|X_{nt}| > \varepsilon\right) \mid \mathscr{F}_{n,t-1}\right] \xrightarrow{\text{p}} 0,$$

and (4.3) is replaced by an analogous condition on the conditional variance:

$$V_{nk_n}^2 = \sum E\left(X_{nt}^2 \mid \mathscr{F}_{n,t-1}\right) \xrightarrow{\text{p}} \eta^2,$$

and if (4.5) holds, then the conclusion of Theorem 2 *remains true.*

Under further conditions of stationarity and ergodicity,

Corollary 4.23. *Let $\{X_t\}$ be stationary and ergodic martingale difference sequences with $E\left(X_t^2\right) = \sigma^2 < \infty$, then*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \to N\left(0, \sigma^2\right). \tag{4.6}$$

4.5. **CLT's for Correlated Seq.** First, consider the linear process.

Theorem 4.24. *Let*

$$X_t = \mu + \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}.$$

where ε_t *is i.i.d. with* $E\left(\varepsilon_t^2\right) < \infty$ *and* $\sum_{j=0}^{\infty} j \cdot |c_j| < \infty$. *Then*

$$\sqrt{n}\left(\bar{X_n} - \mu\right) \to_d N\left(0, \sum_{h=-\infty}^{\infty} \gamma\left(h\right)\right).$$

Proof. The proof uses Beverage-Nelson Decomposition and Phillips-Solo Device. Let

$$u_t = C\left(L\right)\varepsilon_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}. \tag{4.7}$$

The BN-decomposition tells that we could rewrite the lag operator as

$$C\left(L\right) = C\left(1\right) + \left(L-1\right)\tilde{C}\left(L\right),$$

where

$$C(1) = \sum_{j=0}^{\infty} c_j, \quad \tilde{C}(L) = \sum_{j=0}^{\infty} \tilde{c}_j L^j,$$

and $\tilde{c_j} = \sum_{j+1}^{\infty} c_k$. Since we assume that $\sum_{j=0}^{\infty} j \cdot |c_j| < \infty$, we have $\sum_{j=0}^{\infty} |\tilde{c_j}| < \infty$. When $C(1) \neq 0$, we can rewrite u_t as

$$\begin{aligned} u_t &= \left(C(1) + (L-1)\tilde{C}(L)\right)\varepsilon_t \\ &= C(1)\varepsilon_t - \tilde{C}(L)(\varepsilon_t - \varepsilon_{t-1}) \\ &= C(1)\varepsilon_t - (\tilde{u}_t - \tilde{u}_{t-1}). \end{aligned}$$

Therefore,

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t = C(1) \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t - \frac{1}{\sqrt{n}} (\tilde{u_n} - \tilde{u_0}).$$

First, $C(1) \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t \overset{d}{\to} N\left(0, C(1)^2 \sigma_\varepsilon^2\right)$. Next, since $\tilde{c_j}$ is absolutely summable, then $\tilde{u_n} - \tilde{u_0}$ is bounded in probability, hence

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t = C(1) \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t + o_p(1) \to N\left(0, C(1)^2 \sigma_\varepsilon^2\right). \tag{4.8}$$

You can verify that $\sum_{h=-\infty}^{\infty} \gamma_x (h) = \left(\sum_{j=0}^{\infty} c_j \right)^2 \sigma_\varepsilon^2 = C(1)^2 \sigma_\varepsilon^2$.

This result also applies when ε_t is a martingale difference sequence satisfying certain moment conditions (Phillips and Solo 1992). □

4.6. **HAC.** We develop a consistent estimator of

$$\omega^2 = \sum_{h=-\infty}^{\infty} \gamma_x(h),$$

where $\gamma(h)$ denotes the autocovariance function of X_t. It is known as a heteroskedasticity-autocorrelation consistent (HAC) estimator. See Newey and West (1987) and Andrews (1991) Andrews (1991).

We may begin with $X_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}$, where $\sum_{j=0}^{\infty} |\theta_j| < \infty$ and ε_t is *mds* with $E|\varepsilon_t|^r < \infty$ for some $r > 2$. First, by LLN (e.g. Theorem 4.14)

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} X_t X_{t+h} \xrightarrow{p} \gamma(h)$$

for each h.

However, it is known that

$$\lim_{n\to\infty} \sum_{h=-n+1}^{n-1} \widehat{\gamma}(|h|) \neq \omega^2$$

with a positive probability, see e.g. Kiefer and Vogelsang (2002), and thus we need to consider,

$$\hat{\omega}^2 = \sum_{h=-n+1}^{n-1} k\left(\frac{h}{b_n}\right)\widehat{\gamma}\left(|h|\right),$$

with $b_n = o\left(n\right)$. Basically, the standard error arising from adding up all the sample autocovariance functions is too large and thus we need to control the s.e. by trimming or downweighting $\hat{\gamma}\left(h\right)$ with larger h at the expense of increased biases. (much like in the nonparametric kernel estimation).

Suppose $\{X_t : t \geq 1\}$ is a sequence of random n-vectors generated by the linear process $X_t = \sum_{l=0}^{\infty} C_l \epsilon_{t-l}$ and consider estimation of

$$\Omega = \lim_{n\to\infty} n^{-1}\sum_{t=1}^{n}\sum_{s=1}^{n} E\left(X_t X_s^{'}\right),$$

the long-run covariance matrix of X_t. Let $\|\cdot\|$ denote the Euclidean norm. This part comes from Jansson (2002)Jansson (2002).

Assumption 4.25. $\sum_{l=0}^{\infty} \|C_l\| < \infty$.

Assumption 4.26. *$E_{t-1}(\epsilon_t) \overset{\text{a.s.}}{=} 0$, $E_{t-1}\left(\epsilon_t \epsilon_t^{'}\right) \overset{\text{a.s.}}{=} I_p$, and $\|\epsilon_t\|^2$ is uniformly integrable.*

Defining

$$\Gamma = \lim_{n\to\infty} n^{-1} \sum_{t=2}^{n} \sum_{s=1}^{t-1} E\left(X_t X_s^{'}\right)$$

and

$$\Sigma = \lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} E\left(X_t X_t^{'}\right),$$

yields

$$\Omega = \Gamma + \Gamma^{'} + \Sigma.$$

It is assumed that Γ is estimated by a kernel estimator of the form

$$\hat{\Gamma}_n = n^{-1} \sum_{t=2}^{n} \sum_{s=1}^{t-1} k\left(\frac{|t-s|}{b_n}\right) X_t X_s^{'},$$

where $k(\cdot)$ is a (measurable) kernel function and $\{b_n : n \geq 1\}$ is a sequence of bandwidth parameters. The corresponding estimator of Ω is

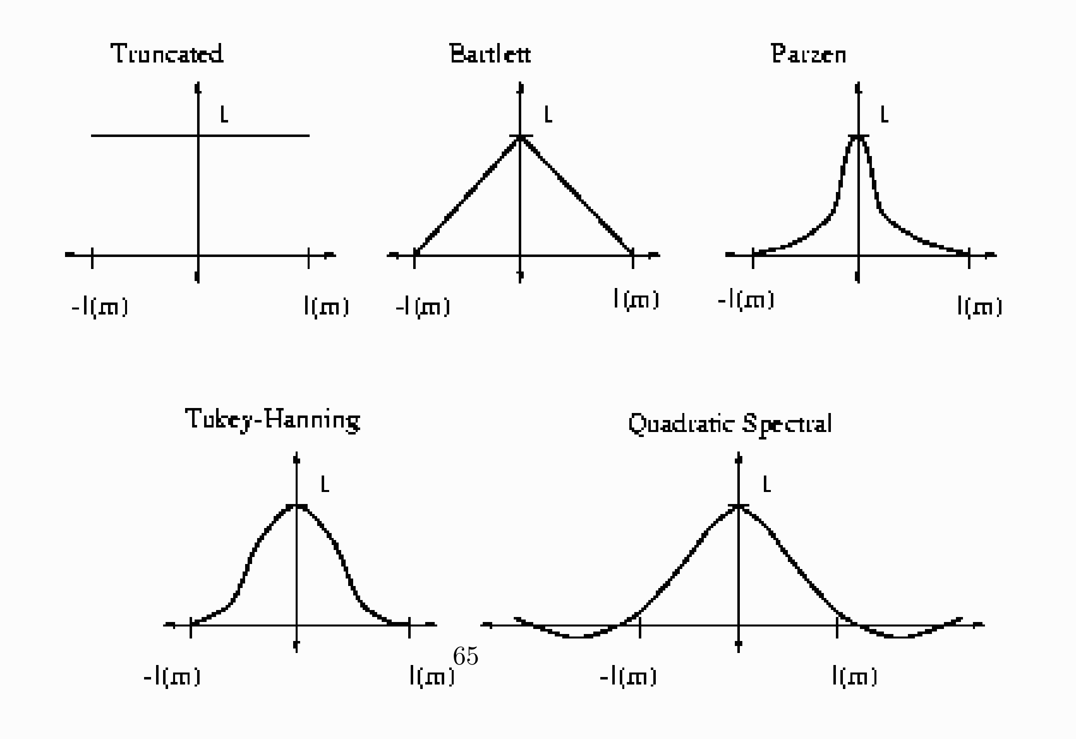
$$\begin{aligned}\hat{\Omega}_n &= n^{-1} \sum_{t=1}^{n} \sum_{s=1}^{n} k\left(\frac{|t-s|}{b_n}\right) X_t X_s^{'}, \\ &= \hat{\Gamma}_n + \hat{\Gamma}'_n + \hat{\Sigma}_n,\end{aligned}$$

where $\hat{\Sigma}_n = n^{-1} \sum_{t=1}^{n} X_t X_t^{'}$. Because $\hat{\Sigma}_n \rightarrow_p \Sigma$ under Assumption 4.25 and 4.26, $\hat{\Omega}_n$ is a consistent estimator of Ω whenever $\hat{\Gamma}_n$ is a consistent estimator of Γ.

Consider the following assumptions on $k(\cdot)$ and $\{b_n\}$.

Assumption 4.27. *(i)* $k(0) = 1$, $k(\cdot)$ *is continuous at zero and* $\sup_{x \geq 0} |k(x)| < \infty$.

(ii) $\int_{[0,\infty)} \bar{k}(x)\, dx < \infty,$ *where* $\bar{k}(x) = \sup_{y \geq x} |k(y)|$.



Truncated

Bartlett

Parzen

Tukey-Hanning

Quadratic Spectral

Assumption 4.28. $\{b_n\} \subseteq (0, \infty)$ *and* $\lim_{n\to\infty} \left(b_n^{-1} + n^{-1/2} b_n\right) = 0.$

An important implication of Assumption 4.27 (ii) is the following lemma.

Lemma 4.29. *Suppose* $k(\cdot)$ *satisfies Assumption 4.27* (*ii*) *and suppose* $\{b_n\} \subseteq (0, \infty)$. *Then*

$$\overline{\lim}_{n\to\infty} \sup_{0<\alpha\leq\alpha_u} b_n^{-1} \sum_{t=1}^{n-1} \left| k\left(\frac{t}{\alpha b_n}\right) \right| < \infty \;\; for \;\; any \;\; 0 < \alpha_u < \infty.$$

The main result of the paper is the following theorem.

Theorem 4.30. *Suppose Assumption 4.25- Assumption 4.28 hold. Then* $\hat{\Gamma}_n \to_p \Gamma$ *and* $\hat{\Omega}_n \to_p \Omega$.

4.6.1. *Proof.* Using change of variables, $\hat{\Gamma}_n$ can be written as

$$\hat{\Gamma}_n = \sum_{t=1}^{n-1} k\left(\frac{t}{b_n}\right) \left(n^{-1} \sum_{j=1}^{n-t} X_{j+t} X_j^{'} \right).$$

Lemma 4.31. *Suppose Assumption 4.25 and Assumption 4.26 hold. Then*

$$E\left\|n^{-1}\sum_{j=1}^{n-t}\left[X_{j+t}X_j^{'}-E\left(X_{j+t}X_j^{'}\right)\right]\right\|\leq \beta_t\psi_n+n^{-1/2}\eta_n,\quad 0\leq t\leq n-1,$$

where $\{\beta_t : t\geq 0\}$ *and* $\{\psi_n,\eta_n : n\geq 1\}$ *are nonnegative sequences with* $\sum_{t=1}^{\infty}\beta_t<\infty$, $\lim_{n\to\infty}\psi_n=0$, *and* $\overline{\lim}_{n\to\infty}\eta_n<\infty$.

It is worthwhile to note that ψ_n and η_n do not depend on t.

Proof of Lemma 4.31 Because

$$\begin{aligned} X_{j+t}X_j^{'} &= \left(\sum_{l=0}^{\infty}C_l\epsilon_{j+t-l}\right)\left(\sum_{m=0}^{\infty}C_m\epsilon_{j-m}\right)^{'} \\ &= \sum_{m=0}^{\infty}C_{m+t}\epsilon_{j-m}\epsilon_{j-m}^{'}C_m^{'}+\sum_{l=0}^{\infty}\sum_{m\neq l-t}C_l\epsilon_{j+t-l}\epsilon_{j-m}^{'}C_m^{'}, \end{aligned}$$

it follows that

$$n^{-1}\sum_{j=1}^{n-t}\left(X_{j+t}X_{j}^{'}-E\left(X_{j+t}X_{j}^{'}\right)\right)$$

$$=\sum_{m=0}^{\infty}C_{m+t}\left(n^{-1}\sum_{j=1}^{n-t}\left(\epsilon_{j-m}\epsilon_{j-m}^{'}-I_{n}\right)\right)C_{m}^{'}+\sum_{l=0}^{\infty}\sum_{m\neq l-t}C_{l}\left(n^{-1}\sum_{j=1}^{n-t}\epsilon_{j+t-l}\epsilon_{j-m}^{'}\right)C_{m}^{'}$$

under Assumption 4.25 and Assumption 4.26. Using subadditivity of $\|\cdot\|$ and the fact that $\|AB\| \leq \|A\| \cdot \|B\|$ for conformable A and B, this expression can be bounded as follows:

$$\begin{aligned}
&\left\| n^{-1} \sum_{j=1}^{n-t} \left(X_{j+t} X_j^{'} - E \left(X_{j+t} X_j^{'} \right) \right) \right\| \\
\leq & \sum_{m=0}^{\infty} \left\| n^{-1} \sum_{j=1}^{n-t} \left(\epsilon_{j-m} \epsilon_{j-m}^{'} - I_n \right) \right\| \left\| C_m \right\| \left\| C_{m+t} \right\| \\
+ & \sum_{l=0}^{\infty} \sum_{m \neq l-t} \left\| n^{-1} \sum_{j=1}^{n-t} \epsilon_{j+t-l} \epsilon_{j-m}^{'} \right\| \left\| C_l \right\| \left\| C_m \right\| .
\end{aligned}$$

Therefore, $E\left\|n^{-1}\sum_{j=1}^{n-t}\left[X_{j+t}X_j^{'}-E\left(X_{j+t}X_j^{'}\right)\right]\right\|\leq\beta_t\psi_n+n^{-1/2}\eta_n$, where

$$\begin{aligned}\beta_t &= \sum_{m=0}^{\infty}\|C_m\|\,\|C_{m+t}\|\,,\\ \psi_n &= \sup_{m\geq 0}\max_{0\leq t\leq n-1}E\left\|n^{-1}\sum_{j=1}^{n-t}\left(\epsilon_{j-m}\epsilon_{j-m}^{'}-I_n\right)\right\|,\\ \eta_n &= \left(\max_{0\leq t\leq n-1}\sup_{l,m\geq 0}E\left\|1\left\{l\neq m+t\right\}n^{-1/2}\sum_{j=1}^{n-t}\epsilon_{j+t-l}\epsilon_{j-m}^{'}\right\|\right)\left(\sum_{l=0}^{\infty}\|C_l\|\right)^2.\end{aligned}$$

By Assumption 4.25,

$$\sum_{t=1}^{\infty}\beta_t=\sum_{t=1}^{\infty}\sum_{m=0}^{\infty}\|C_m\|\,\|C_{m+t}\|\leq\left(\sum_{m=0}^{\infty}\|C_m\|\right)^2<\infty.$$

Each element of $\left\{\epsilon_{j-m}\epsilon_{j-m}^{'}-I_n:j\geq 1\right\}$ is a uniformly integrable martingale difference sequence under Assumption 4.26. As a consequence, for any $\epsilon>0$

there is a finite constant λ_ϵ (independent of t and m) such that

$$\begin{aligned} E\left\|n^{-1}\sum_{j=1}^{n-t}\left(\epsilon_{j-m}\epsilon_{j-m}^{'}-I_n\right)\right\| &\leq (n-t)^{1/2}\,n^{-1}\lambda_\epsilon+(n-t)\,n^{-1}\epsilon \\ &\leq n^{-1/2}\lambda_\epsilon+\epsilon, \end{aligned}$$

where the first inequality is obtained by proceeding as in the proof of Theorem 4.13 (HH, Theorem 2.22). Therefore, $\overline{\lim}_{n\to\infty}\psi_n \leq \epsilon$ for any $\epsilon > 0$, so $\psi_n \to 0$.

Finally,

$$
\begin{aligned}
&\left\| 1\left\{ l \neq m+t \right\} n^{-1/2} \sum_{j=1}^{n-t} \epsilon_{j+t-l} \epsilon_{j-m}^{'} \right\|^2 \\
=& 1\left\{ l \neq m+t \right\} \operatorname{tr} \left[\left(n^{-1/2} \sum_{j_1=1}^{n-t} \epsilon_{j_1+t-l} \epsilon_{j_1-m}^{'} \right)^{'} \left(n^{-1/2} \sum_{j_2=1}^{n-t} \epsilon_{j_2+t-l} \epsilon_{j_2-m}^{'} \right) \right] \\
=& n^{-1} \sum_{j_1=1}^{n-t} \sum_{j_2=1}^{n-t} 1\left\{ l \neq m+t \right\} \epsilon_{j_1-m}^{'} \epsilon_{j_2-m} \epsilon_{j_1+t-l}^{'} \epsilon_{j_2+t-l}.
\end{aligned}
$$

By Assumption 4.26,

$$
E\left(1\left\{ l \neq m+t \right\} \epsilon_{j_2-m}^{'} \epsilon_{j_1-m} \epsilon_{j_1+t-l}^{'} \epsilon_{j_2+t-l} \right) = p^2 1\left\{ j_1 = j_2 \right\} 1\left\{ l \neq m+t \right\},
$$

because, e.g.,

$$
E\left(\epsilon_{j_2-m}^{'} \epsilon_{j_1-m} \epsilon_{j_1+t-l}^{'} \epsilon_{j_2+t-l} \right) = E\left[E_{j_2-m}\left(\epsilon_{j_1-m}^{'} \right) \epsilon_{j_2-m} \epsilon_{j_1+t-l}^{'} \epsilon_{j_2+t-l} \right] = 0
$$

when $j_1 > j_2$ and $l > m+t$, whereas

$$E\left(\epsilon_{j_2-m}^{'}\epsilon_{j_1-m}\epsilon_{j_1+t-l}^{'}\epsilon_{j_2+t-l}\right) = E\left[E_{j+t-l}\left(\epsilon_{j-m}^{'}\epsilon_{j-m}\right)\epsilon_{j+t-l}^{'}\epsilon_{j+t-l}\right] = p^2$$

when $j_1 = j_2 = j$ and $l > m+t$. Therefore,

$$\begin{aligned}
&E\left\|1\left\{l \neq m+t\right\}n^{-1/2}\sum_{j=1}^{n-t}\epsilon_{j+t-l}\epsilon_{j-m}^{'}\right\| \\
\leq & E\left(\left\|1\left\{l \neq m+t\right\}n^{-1/2}\sum_{j=1}^{n-t}\epsilon_{j+t-l}\epsilon_{j-m}^{'}\right\|^2\right)^{1/2} \\
= & \left(n^{-1}\sum_{j_1=1}^{n-t}\sum_{j_2=1}^{n-t}n^2 1\left\{j_1 = j_2\right\}1\left\{l \neq m+t\right\}\right)^{1/2} \\
= & \left[1\left\{l \neq m+t\right\}p^2 n^{-1}\left(n-t\right)\right]^{1/2} \\
\leq & p,
\end{aligned}$$

where the first inequality uses the Cauchy-Schwarz inequality. In particular,

$$\overline{\lim}_{n\to\infty}\eta_n \leq p\cdot\left(\sum_{l=0}^{\infty}\|C_l\|\right)^2 < \infty,$$

as was to be shown. ■

Lemma 4.32. *Suppose Assumption 4.25 and Assumption 4.26 hold. Moreover, suppose* $k(\cdot)$ *satisfies Assumption 4.27* (*i*) *and suppose* $\{b_n\} \subseteq (0,\infty)$ *with* $\lim_{n\to\infty} b_n^{-1} = 0$*. Then*

$$\lim_{n\to\infty}\sup_{\alpha\geq\alpha_l}\left\|E\left[\hat{\Gamma}_n(\theta_0,\alpha b_n)\right]-\Gamma\right\| = 0 \;\; for \;\; any \;\; 0<\alpha_l<\infty.$$

Proof of Lemma 4.32 Under Assumption 4.25 and Assumption 4.26,

$$E\left[\hat{\Gamma}_n(\theta_0,\alpha b_n)\right] = \sum_{t=1}^{\infty} 1\{t\leq n-1\}\,k\left(\frac{t}{\alpha b_n}\right)\frac{n-t}{n}E\left(X_{1+t}X_1^{'}\right)$$

for any $\alpha > 0$, whereas $\Gamma = \sum_{t=1}^{\infty} E\left(X_{1+t}X_1^{'}\right)$. Therefore, by subadditivity of $\|\cdot\|$,

$$\begin{aligned}
&\left\|E\left[\hat{\Gamma}_n\left(\theta_0, \alpha b_n\right)\right] - \Gamma\right\| \\
&\leq \sum_{t=1}^{I}\left|1\left\{t \leq n-1\right\}k\left(\frac{t}{\alpha b_n}\right)\frac{n-t}{n} - 1\right| \cdot \left\|E\left(X_{1+t}X_1^{'}\right)\right\| \\
&+ \sum_{t=I+1}^{\infty}\left|1\left\{t \leq n-1\right\}k\left(\frac{t}{\alpha b_n}\right)\frac{n-t}{n} - 1\right| \cdot \left\|E\left(X_{1+t}X_1^{'}\right)\right\| \\
&\leq \max_{1\leq t\leq I}\left|1\left\{t \leq n-1\right\}k\left(\frac{t}{\alpha b_n}\right)\frac{n-t}{n} - 1\right| \cdot \sum_{t=1}^{I}\left\|E\left(X_{1+t}X_1^{'}\right)\right\| \\
&+ \left(\sup_{x\geq 0}\left|k\left(x\right)\right| + 1\right) \cdot \sum_{t=I+1}^{\infty}\left\|E\left(X_{1+t}X_1^{'}\right)\right\|
\end{aligned}$$

for any $I \geq 1$. Because $\sum_{t=0}^{\infty} \left\| E\left(X_{1+t} X_1^{'}\right)\right\| < \infty$,

$$\left(\sup_{x\geq 0} |k(x)| + 1\right) \cdot \sum_{t=I+1}^{\infty} \left\| E\left(X_{1+t} X_1^{'}\right)\right\|$$

can be made arbitrarily small (under Assumption 4.27 (i)) by taking I large enough. For any given I,

$$\begin{aligned}
&\sup_{\alpha\geq\alpha_l} \max_{1\leq t\leq I} \left| 1\{t \leq n-1\}\, k\left(\frac{t}{\alpha b_n}\right) \frac{n-t}{n} - 1\right| \\
&= \sup_{\alpha\geq\alpha_l} \max_{1\leq t\leq I} \left| k\left(\frac{t}{\alpha b_n}\right) \frac{n-t}{n} - 1\right| \\
&\leq \sup_{\alpha\geq\alpha_l} \max_{1\leq t\leq I} \left(\left| k\left(\frac{t}{\alpha b_n}\right) - 1\right| + n^{-1} t \left| k\left(\frac{t}{\alpha b_n}\right)\right| \right) \\
&\leq \sup_{0\leq x\leq I/(\alpha_l b_n)} |k(x) - 1| + n^{-1} I \sup_{x\geq 0} |k(x)|
\end{aligned}$$

whenever $0 < \alpha_l < \infty$ and $n \geq I + 1$. Lemma 4.32 now follows because the expression on the last line tends to zero whenever Assumption 4.27 (i) holds and $\lim_{n\to\infty} b_n^{-1} = 0$. ■

Proof of Theorem 4.30 By subadditivity of $\|\cdot\|$,

$$\begin{aligned}\left\|\hat{\Gamma}_n - \Gamma\right\| &\leq \left\|\hat{\Gamma}_n - E\left(\hat{\Gamma}_n\right)\right\| + \left\|E\left(\hat{\Gamma}_n\right) - \Gamma\right\|, \\ \left\|\hat{\Omega}_n - \Omega\right\| &\leq \left\|\hat{\Omega}_n - E\left(\hat{\Omega}_n\right)\right\| + \left\|E\left(\hat{\Omega}_n\right) - \Omega\right\|.\end{aligned}$$

Now, $E\left(\hat{\Sigma}_n\right) = \Sigma$, so $\left\|E\left(\hat{\Omega}_n\right) - \Omega\right\| \leq 2 \cdot \left\|E\left(\hat{\Gamma}_n\right) - \Gamma\right\| \to 0$ by Lemma 4.32. Moreover,

$$\left\|\hat{\Omega}_n - E\left(\hat{\Omega}_n\right)\right\| \leq 2 \cdot \left\|\hat{\Gamma}_n - E\left(\hat{\Gamma}_n\right)\right\| + \left\|\hat{\Sigma}_n - E\left(\hat{\Sigma}_n\right)\right\|.$$

By Lemma 4.31,

$$E\left\|\hat{\Sigma}_n - E\left(\hat{\Sigma}_n\right)\right\| = E\left\|n^{-1}\sum_{j=1}^{n}\left[X_j X_j^{'} - E\left(X_j X_j^{'}\right)\right]\right\| \to 0.$$

In particular, $\hat{\Sigma}_n - E\left(\hat{\Sigma}_n\right) = o_p(1)$. The proof of Theorem 4.30 can be completed by showing that $E\left\|\hat{\Gamma}_n - E\left(\hat{\Gamma}_n\right)\right\| \to 0$. By subadditivity of $\|\cdot\|$,

$$\left\|\hat{\Gamma}_n - E\left(\hat{\Gamma}_n\right)\right\| \leq \sum_{t=1}^{n-1}\left|k\left(\frac{t}{b_n}\right)\right| \cdot \left\|n^{-1}\sum_{j=1}^{n-t}\left[X_{j+t}X_j^{'} - E\left(X_{j+t}X_j^{'}\right)\right]\right\|.$$

Using Lemmas 4.29 and 4.31 and the notation from Lemma 4.31,

$$\begin{aligned}
E\left\|\hat{\Gamma}_n - E\left(\hat{\Gamma}_n\right)\right\| &\leq \sum_{t=1}^{n-1}\left|k\left(\frac{t}{b_n}\right)\right| \cdot E\left\|n^{-1}\sum_{j=1}^{n-t}\left[X_{j+t}X_j^{'} - E\left(X_{j+t}X_j^{'}\right)\right]\right\| \\
&\leq \bar{k}(0)\left(\sum_{t=1}^{n-1}\beta_t\right)\psi_n + n^{-1/2}\eta_n\sum_{t=1}^{n-1}\left|k\left(\frac{t}{b_n}\right)\right| \\
&\leq \bar{k}(0)\left(\sum_{t=1}^{\infty}\beta_t\right)\psi_n + \left(n^{-1/2}b_n\eta_n\right)\left(b_n^{-1}\sum_{t=1}^{n-1}\left|k\left(\frac{t}{b_n}\right)\right|\right) \\
&\to 0. \blacksquare
\end{aligned}$$

4.6.2. *Bandwidth.* The choice of bandwidth is an important practical issue. There is a large literature, including Andrews (1991)Andrews (1991), Sun, Phillips, Jin (2008)Sun, Phillips, and Jin (2008), and many others. Andrews' optimal bandwidth is given as $O\left(T^{1/(1+2q)}\right)$, where q stands for the order of kernel used.

4.6.3. *Fixed Bandwidth.* Kiefer and Vogelsang (2002) and Kiefer and Vogelsang (2005)'s idea is to fix

$$\frac{b_n}{n} = b \in (0,1].$$

This results in inconsistency of HAC but the test statistic itself is asymptotically pivotal, albeit having a non-standard asymptotic distribution. They claim that this asymptotic distribution reflects the influence of the bandwidth selection on the finite sample distributions. See Lazarus et al. (2018)Lazarus et al. (2018) for more recent survey. We'll revisit this after studying the FCLT.

The intuition has grown from the observation that

$$\sum_i \sum_j \bar{v}_i \bar{v}_j \left(1 - \frac{|i-j|}{n}\right) = 2\frac{1}{n}\sum_{t=1}^{n}\left(\sum_{i=1}^{t}\bar{v}_i\right)^2,$$

where $\bar{v}_i = v_i - \bar{v}$.

4.7. **ARMA.** The autoregressive model can be estimated by OLS (or Yule-Walker estimation). For an AR(p) process

$$\phi(L)\, y_t = \varepsilon_t,$$

with autocovariance function $\gamma(h)$, premultiply $(y_{t-1}, ..., y_{t-p})'$ on both sides of $[y_t - \varepsilon_t = (y_{t-1}, ..., y_{t-p})\phi]$, and take expectation. Then, we have

$$\Gamma\phi = \gamma,$$

where $\Gamma = [\gamma(i-j)]_{i,j=1,\ldots,p}$ and $\gamma = (\gamma(1), ..., \gamma(p))'$. Plugging in the sample autocovariances

$$\hat{\gamma}(j) = n^{-1} \sum_{t=|j|+1}^{n} (y_t - \bar{y}_n)\left(y_{t-|j|} - \bar{y}_n\right),$$

and solving the equation for ϕ yields the Yule-Walker estimation.

It is known that the plug-in estimate $\hat{\Gamma}$ is non-singular as long as $\hat{\gamma}(0) > 0$. On the other hand, if the scaling n^{-1} is replaced by $(n-j)^{-1}$ then the resulting $\hat{\Gamma}$ may not be ≥ 0.

It is straightforward to show that the OLS estimator is consistent and asymptotically normal using the ergodic theorem and the MDS CLT provided that $\mathrm{E}\left(\varepsilon_t^4\right) < \infty$.

The ARMA model is commonly estimated by MLE assuming the normality of the innovations $\{\varepsilon_t\}$. The asymptotic analysis of the estimator is more demanding. It is plausible to use a sample analogue to estimate the ARMA coefficients but not efficient. See Brockwell and Davis for more details.

It is also useful to note that one can estimate the $AR(\infty)$ model instead of the ARMA model, i.e., $AR(p)$ with $p \to \infty$. Setting p suitablely is an issue, see e.g. Berk (1974).

4.7.1. *Lag order selection.* In practice, the lag order p is unknown and needs to be chosen based on the data. This is a model selection issue. The most commonly used method is to use the information criteria such as AIC or BIC, that is, choose the model that minimizes the following criterion functions, e.g.,

$$AIC\left(p\right)=\log\hat{\sigma}^{2}\left(p\right)+2\frac{p}{n},$$

or

$$BIC\left(p\right)=\log\hat{\sigma}^{2}\left(p\right)+\log\left(n\right)\frac{p}{n},$$

where $\hat{\sigma}^{2}\left(p\right)$ is the residual variance estimated from an $AR\left(p\right)$. The probability that the true p, say, p^{*} is chosen,

$$\Pr\left\{BIC\left(p\right)-BIC\left(p^{*}\right)\geq0,\forall p\right\},$$

needs to converge to 1, for which the selection rule is called ‘consistent’. BIC is consistent, while AIC is not. Usually, we evaluate its complement’s probability

$$\begin{aligned} &\Pr\left\{BIC\left(p\right)-BIC\left(p^{*}\right)<0,\exists p\right\} \\ &\leq \sum_{p}\Pr\left\{n\left(\log\hat{\sigma}^{2}\left(p\right)-\log\hat{\sigma}^{2}\left(p^{*}\right)\right)<-\log n\left(p-p^{*}\right)\right\}\rightarrow 0. \end{aligned}$$

4.7.2. *Testing.* We may perform a test for remaining serial correlation in the error. For instance, consider a model

$$\begin{aligned} y_t &= \phi_0+\phi_1 y_{t-1}+\cdots+\phi_p y_{t-p}+u_t \\ u_t &= \rho u_{t-1}+\varepsilon_t, \end{aligned}$$

where $\{\varepsilon_t\}$ is a sequence of *iid* random variables. And you want to test for the missing serial correlation in u_t, that is, test the hypothesis

$$H_0:\rho=0.$$

This is equivalent to testing $H_0 : \phi_{p+1} = 0$ in the regression

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \phi_{p+1} y_{t-p-1} + \varepsilon_t.$$

4.7.3. *Machine learning approach.* Another alternative is the lasso type machine learning approach. Refer to Wang, Li, and Tsai (2007).

4.7.4. *Numerical Optimization.* Widely used methods include the grid search, Mixed Integer Optimization (MIO), and gradient methods, which serves different situations.

(1) The grid search is used when the criterion function is not smooth.
(2) Gradient method: recall that

$$\begin{aligned} 0 &= z_n\left(\hat{\theta}\right) \\ &= z_n(\theta) + H_n\left(\tilde{\theta}\right)\left(\hat{\theta}-\theta\right). \end{aligned}$$

This suggest iteration based on

$$\theta_{i+1} = \theta_i - H_n(\theta_i)^{-1} z_n(\theta_i),$$

with some initial value θ_0. This is the *Newton-Raphson* method.

- Drawbacks: (*i*) rank condition for H_n; (*ii*) step size

(3) *Gauss-Newton* method: in case of MLE, the hessian $H_n(\theta)$ can replaced by

$$\frac{1}{n}\sum_{i=1}^{n} z_i(\theta)\, z_i(\theta)',$$

and thus we do not have to compute the second derivative of M_n.

(4) The MIO is more recent development for discontinuous criterions, see e.g. Bertsimas, King, and Mazumder (2016). Recall tht

$$\min_p BIC(p) = \log \hat{\sigma}^2(p) + \log(n)\frac{p}{n},$$

and reformulate it as

$$\min_{\phi, d_j \in \{0,1\}} n^{-1}\sum_{t=1}^{n}\left(y_t - y_{t-1}\phi_1 d_1 - \cdots - y_{t-k}\phi_k d_k\right)^2 + \lambda_n \sum_{j=1}^{k} d_j.$$

(5) The computational complexity motivates Tibshirani (1996)'s lasso approach

$$\min_{\phi} n^{-1} \sum_{t=1}^{n} \left(y_t - y_{t-1}\phi_1 - \cdots - y_{t-k}\phi_k\right)^2 + \lambda_n \sum_{j=1}^{k} \left|\phi_j\right|,$$

which is a convex optimization problem.

5. Unit Root

We noted in the previous section that the stationarity requires to exclude unit roots in the AR models. However, the presence of such a root is prevalent in the economic data and it distorts the standard inference procedure. Spurious regression is noted by an experiment performed by Granger and Newbold (1974) . An interesting real example is given by Hendry who showed that when the money supply in England is regressed to rainfall in Mongola the t-statistic is highly significant if we based our inference on the standard Normal distribution.

5.1. **FCLT.**

As the asymptotic behavior of the sample means of a process with a unit root is different from that of a stationary process, we need to introduce a more general convergence concept known as the functional central limit theorem (FCLT) or weak convergence of random functions. And the asymptotic distribution is characterized by a stochastic process called the Brownian motion (BM).

Definition 5.1. The univariate standard Brownian Motion (or Wiener process) W is a stochastic process with continuous sample path satisfying the following three properties:
(i) $W(0) = 0$ a.s.
(ii) (*Independent Increment*) For $0 < r_1 < r_2 < \cdots < 1$

$$W(r_1), W(r_2) - W(r_1), W(r_3) - W(r_2), \ldots$$

are independet.
(iii) (*Gaussianity*) For $r < s$

$$W(s) - W(r) \sim \mathcal{N}(0, s - r)$$

We have in particular that $W(r) \sim \mathcal{N}(0, r)$. Also, $W(r)$ and $W(s)$ are jointly normal with covariance $r \wedge s$. We may generalize this to any finite selection of $W(r)$'s.

We can view the stochastic process as a random mapping from the sample space Ω to a function space defined on the interval $[0, 1]$. In case of the Brownian motion, the space consists of continuous functions. As for a time series, it can also viwed as the mapping from the product space $\Omega \times [0,1]$ to $\mathbb{R}$. Thus, for each $r \in [0,1]$ W is a random variable and for each $\omega \in \Omega$, W is a function, which is called sample path.

We generalize the sample mean to the *partial sum process* on $r \in [0,1]$, which is defined as

$$W_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} \varepsilon_t,$$

for a sequence of random variables $\{\varepsilon_t\}$. Note that $W_n(1)$ is the sample mean of $\{\varepsilon_t\}$. The weak convergence of the partial sum process is signified by " $\Rightarrow$ ".

Definition 5.2. (weak convergence $\Rightarrow$) we write

$$W_n \Rightarrow W$$

if

$$\mathrm{E}^* f(W_n) \to \mathrm{E} f(W) \ \forall \ f \in \mathcal{U}\left(\ell^\infty([0,1])\right),$$

where $\mathcal{U}\left(\ell^{\infty}\left([0,1]\right)\right)$ is the collection of all real functions f that are defined on $\ell^{\infty}\left([0,1]\right)$ and bounded and uniformly continuous, while $\ell^{\infty}\left([0,1]\right)$ denotes the space of bounded functions on $[0,1]$.

Conceptually, weak convergence is the same as convergence in distribution but defined in more general spaces.

Theorem 5.3. *Let $W_n\left(r\right)$ be a partial sum process with $\{\varepsilon_t\}$, which is a centered iid sequence with variance σ^2. Then*

$$W_n\left(r\right) \Rightarrow \sigma W\left(r\right).$$

This is a generalization of the standard CLT. For each fixed r, the CLT yields that

$$\frac{1}{\sqrt{[nr]}}\sum_{t=1}^{[nr]}\varepsilon_t \xrightarrow{d} \mathcal{N}\left(0,\sigma^2\right).$$

Compare this with

$$W_n(r) = \frac{\sqrt{[nr]}}{\sqrt{n}} \left(\frac{1}{\sqrt{[nr]}} \sum_{t=1}^{[nr]} \varepsilon_t \right) \xrightarrow{d} \sqrt{r}\mathcal{N}\left(0, \sigma^2\right) = \mathcal{N}\left(0, r\sigma^2\right),$$

We focus on how this theorem can be used to obtain the representation of the asymptotic distribution of sample means of AR processes with a unit root. The following theorem sheds some light on it.

Hereafter, $\{\varepsilon_t\}$ is a centered iid sequence with variance σ^2. The following is some fundamental convergence theorems for commonly used sample moments.

Theorem 5.4. *Let $y_t = y_{t-1} + \varepsilon_t$ and $y_0 = 0$. Then,*

$$\begin{aligned} \frac{1}{n^{1+k/2}} \sum_{t=1}^{n} y_{t-1}^k &\Rightarrow \sigma^k \int_0^1 W^k(r)\,dr \\ \frac{1}{n} \sum_{t=1}^{n} y_{t-1}\varepsilon_t &\Rightarrow (1/2)\,\sigma^2 \left(W(1)^2 - 1\right). \end{aligned}$$

Proof. Sketch) Note that

$$\frac{1}{n^{1+k/2}}\sum_{t=1}^{n} y_{t-1}^{k} = \int_0^1 \left(\frac{y_{[nr]}}{\sqrt{n}}\right)^k dr \Rightarrow \sigma^k \int_0^1 W^k\left(r\right) dr,$$

by the CMT. Furthermore, since

$$y_t^2 = y_{t-1}^2 + \varepsilon_t^2 + 2y_{t-1}\varepsilon_t,$$

we have

$$\begin{aligned} 2\frac{1}{n}\sum_{t=1}^{n} y_{t-1}\varepsilon_t &= \frac{1}{n}\sum_{t=1}^{n}\left(y_t^2 - y_{t-1}^2\right) - \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 \\ &= \frac{1}{n}y_n^2 - \frac{1}{n}y_0^2 - \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 \\ &\Rightarrow \sigma^2\left(W\left(1\right)^2 - 1\right). \end{aligned}$$

□

Consider an AR(1) process

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

and the OLS estimator

$$\hat{\phi} = \phi + \left(\sum_{t=2}^{n} y_{t-1}^2 \right) \sum_{t=2}^{n} y_{t-1} \varepsilon_t .$$

We saw in the previous section that $\hat{\phi}$ is consistent and asymptotically normal when $|\phi| < 1$.

If $\phi = 1$, then $\hat{\phi}$ is still consistent but not asymptotically normal. From the proposition and the CMT,

$$n \left(\hat{\phi} - \phi \right) \Rightarrow \left(\int_0^1 W^2 \left(r \right) dr \right)^{-1} \left(W \left(1 \right)^2 - 1 \right) \frac{1}{2}.$$

Due to the so-called Ito's lemma, we can write

$$\left(W \left(1 \right)^2 - 1 \right) \frac{1}{2} = \int_0^1 W dW,$$

which is known as a stochastic integral.

A word on Ito's lemma: the Brownian motion's sample paths are special, continuous and no-where differentiable and

$$\sum \left(W\left(t_i\right) - W\left(t_{i-1}\right)\right)^2$$

has a non-degenrate (probability or L_2) limit. (quadratic variation $[W]_t = t$) An implication of this on the taylor series expansion is that

$$\begin{aligned} df\left(W\left(t\right), t\right) &\sim f_W dW + f_t dt + 2^{-1} f_{WW} \left(dW\right)^2 \\ &\sim f_W dW + f_t dt + 2^{-1} f_{WW} dt. \end{aligned}$$

For Ito's formula with $f\left(W, t\right) = W\left(t\right)^2/2$ yields the above.

In practice, we always include the constant 1 in the regression. Then, by the FWL theorem,

$$n\left(\hat{\phi} - 1\right) = \left(\frac{1}{n^2}\left(\sum_{t=2}^{n} y_{t-1}^2 - n\bar{y}^2\right)\right)^{-1}\left(\frac{1}{n}\left(\sum_{t=2}^{n} y_{t-1}\varepsilon_t - n\bar{y}\bar{\varepsilon}\right)\right).$$

And from the FCLT and CLT

$$\begin{aligned}\frac{\bar{y}}{\sqrt{n}} &= \frac{1}{n^{3/2}}\sum_{t=1}^{n} y_t \Rightarrow \sigma \int_0^1 W, \\ \sqrt{n}\bar{\varepsilon} &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n} \varepsilon_t \xrightarrow{d} \sigma W(1).\end{aligned}$$

Thus,

$$n\left(\hat{\phi}-1\right) \Rightarrow \left(\int_0^1 \bar{W}^2\right)^{-1} \int_0^1 \bar{W} d\bar{W},$$

where $\bar{W} = W - \int_0^1 W$. This result also yields

$$\hat{\phi} = 1 + O_p\left(\frac{1}{n}\right).$$

Thus, the estimation error diminishes faster than the standard $\sqrt{n}$ rate, which is called super-consistency of $\hat{\phi}$. The inclusion of the constant term in the regression changes the asymptotic distribution of $\hat{\phi}$ in a significant way. Furthermore, the asymptotic distribution of $\hat{\phi}$ is not normal if $\phi = 1$. Therefore, determining if the process contains a unit root is important. We do not consider the case where $\phi > 1$.

5.2. **Unit Root Test.** Testing for the presence of a unit root for economic time series has become a routine procedure. However, the testing is rather complex due to the factors we explain below and requires careful treatment. First, the unit root test is one-sided test as the unit root hypothesis lies at the boundary of the stationarity hypothesis. Second, the asymptotic distribution of the t-test is not normal. Third, the asymptotic distribution of the statistic changes depending on the presence of a constant and/or a linear trend in the estimation. We study two most common cases here.

We distinguish the cases by the presence of the linear time trend. That is,

$$y_t = s_t + x_t,$$

where s_t is a deterministic process and x_t is a stochastic process. Our question is whether x_t contains a unit root or not. We consider either

$$s_t = \mu \quad \text{or} \quad s_t = \mu_0 + \mu_1 t.$$

And for our hypothesis testing we transform

$$x_t = \phi x_{t-1} + \varepsilon_t :\to \Delta x_t = (\phi - 1)\, x_{t-1} + \varepsilon_t,$$

where $\Delta = 1 - L$ is the difference operator. Accordingly, $\Delta s_t = 0$ or $\Delta s_t = \mu_1$.

- **Unit Root Test with no deterministic time trend:** The null and alternative hypothesis are formulated as follows.

$$\begin{aligned} \mathcal{H}_0 &: \Delta y_t = \varepsilon_t \\ vs.\ \mathcal{H}_1 &: \Delta y_t = \mu + \rho y_{t-1} + \varepsilon_t, \text{ and } \rho < 0. \end{aligned} \tag{5.1}$$

Then, one estimate the alternative model (5.1), typically, by OLS. Let the OLS estimator by $(\hat{\mu}, \hat{\rho})$. Then,

$$\begin{aligned} n\hat{\rho} &\Rightarrow \left(\int \bar{W}^2\right)^{-1} \int \bar{W} d\bar{W} \\ t_\rho &= \frac{\hat{\rho}}{\sqrt{\hat{\sigma}^2 \left(\sum_{t=1}^n \bar{y}_{t-1}^2\right)^{-1}}} \Rightarrow \left(\int \bar{W}^2\right)^{-1/2} \int \bar{W} d\bar{W}, \end{aligned}$$

where $\bar{W} = W - \int_0^1 W$, $\bar{y}_t = y_t - \frac{1}{n}\sum_{t=1}^n y_t$ and $\hat{\sigma}^2 = n^{-1}\sum_{t=1}^n \hat{\varepsilon}_t^2$.

- The consistency of $\hat{\sigma}^2$ can be derived in a straightforward way.
- As this test is one-sided, you reject the null if the sample statistic is smaller than, say, 5 percentile of the limit distribution, which is often called the Dickey-Fuller distribution. This is skewed to the left and has a negative mean.
- The estimator $\hat{\sigma}^2$ is the standard homoskedastistic variance estimator. It can be shown that even when the error is heteroskedastistic the limit distribution is valid. Thus, the unit root test is robust to the heteroskedasticity.
- Strictly speaking we need to test the joint hypothesis that $\mu = 0$ and $\rho = 0$ but typically we do as above treating $\mu = 0$ as an auxiliary assumption under the null.

- **Unit Root Test with a deterministic time trend:**

$$\begin{aligned} \mathcal{H}_0 &: \Delta y_t = \delta + \varepsilon_t \\ vs.\ \mathcal{H}_1 &: \Delta y_t = \mu_0 + \mu_1 t + \rho y_{t-1} + \varepsilon_t \text{ and } \rho < 0. \end{aligned}$$

In this case, y_t has an deterministic time trend under both hypotheses. Then, we estimate the alternative model and construct the standard t-statistic for testing the null of $\rho = 0$. It can be shown that

$$t_\rho \Rightarrow \left(\int \tilde{W}^2 \right)^{-1/2} \int \tilde{W} d\tilde{W},$$

where $\tilde{W} = W(r) - a_0 - b_0 r$ and $(a_0, b_0) = arg\min_{(a,b)} \int (W(r) - a - br)^2 dr$.

- NB. y_t is collinear to t asymptotically as $y_t = \delta t + \xi_t$ under the $\mathcal{H}_0$, where $\xi_t = y_0 + \sum_{s=1}^{t} \varepsilon_s$. But, the asymptotic distribution can still be derived for ρ after some transformation and new parametrization. See Hamilton (1994, p. 498).

In general,

- AR(p) model

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

is transformed to

$$\Delta y_t = \mu + \rho_0 y_{t-1} + \rho_1 \Delta y_{t-1} + \cdots + \rho_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

where

$$\rho_0 = -1 + \phi_1 + \cdots + \phi_p \text{ and } \rho_j = \sum_{i=j}^{p-1} \phi_i \text{ for } j = 1, ..., p-1.$$

In terms of the lag polynomial $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots$, we see that

$$\rho_0 = -\phi(1),$$

and the BN decomposition for $\tilde{\phi}(z) = \phi_1 z + \phi_2 z^2 + \cdots$,

$$\tilde{\phi}(L) = \tilde{\phi}(1) + (L-1)\tilde{\rho}(L),$$

is applied. Thus, we can test for the presence of a unit root using ρ_0, i.e.,

$$\mathcal{H}_0 : \rho_0 = 0 \quad vs. \quad \mathcal{H}_1 : \rho_0 < 0,$$

using the t-statistic.

- Implicit here is an assumption that the lag polynomial $\phi(z)$ contains at most one unit root and the others lie outside unit circle.
- The FCLT can be generalized to cover this case and the asymptotic distribution does not change according to the change in the lag length p. See Ng and Perron for the lag-order selection in unit root testing.

5.3. **Confidence Interval.** Due to the discontinuity in the asymptotic distribution of $\hat{\phi}$ at $\phi = 1$, it is difficult to construct valid confidence intervals for ϕ when it is uncertain if ϕ is away from the unit root. Hansen (1999) proposed a grid bootstrap confidence interval and Mikusheva (2007) verified that it is uniformly valid.

5.4. **Deterministic Trend.** It is often the case that a time series consists of a stochastic part η_t and a deterministic part s_t :

$$y_t = s_t + \eta_t.$$

For instance, s_t might be a linear time trend or seasonality dummies and η_t be an AR process, etc. Then, the statistical inference is made after filtering out the deterministic component. Here we consider the case with a linear time trend.

Suppose that

$$\begin{aligned} s_t &= \alpha^* + \delta^* t, \\ \phi(L)\eta_t &= \varepsilon_t, \end{aligned} \tag{5.2}$$

where $\{\varepsilon_t\}$ is an iid sequence. This can be equivalently written as

$$\phi(L) y_t = \alpha + \delta t + \varepsilon_t, \tag{5.3}$$

where

$$\phi(L)(\alpha^* + \delta^* t) = \alpha + \delta t.$$

If the series $\{\eta\}$ is stationary then the series $\{y_t\}$ is called trend stationary.

The model (5.3) can be estimated by OLS or y_t can be regressed to $(1, t)$ to collect the residuals, which is called *detrending*, and the residuals are fit to the autoregression. The two procedures are equivalent by Frisch-Waugh-Lovell theorem. It may appear that the presence of the trend complicates the asymptotic analysis. However, to some extent, our previous discussion extends to cover this case.

Refer to Hamilton Sec 16.

5.4.1. *Break.* The linear trend is often associated with breaks (so-called structural break).

$$s_t = c_0 + c_1 t + \sum_{j=1}^{J} (d_{0j} + d_{1j} t) 1 \{t > \tau_j\} .$$

The choice of the number J of breaks and the estimation of the break points τ_j are important statistical problems.

Or, it can be modelled as smoothly varying deterministic trend

$$s_t = s\left(t/n\right),$$

where $s\left(\cdot\right)$ is a smooth function on the unit interval.

5.4.2. *The Asymptotic Distribution of OLS (maybe skipped).* Rewriting the model

$$y_t = \alpha + \delta t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t, \tag{5.4}$$

it is assumed throughout this section that ϵ_t is i.i.d. with mean zero, variance σ^2, and finite fourth moment, and that roots of

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

lie outside the unit circle. Consider a sample of $T{+}p$ observations on y, $\left\{y_{-p+1}, y_{-p+2}, \cdots, y_T\right\}$, and let $\hat{\alpha}_T, \hat{\delta}_T, \hat{\phi}_{1,T}, \cdots \hat{\phi}_{p,T}$ denote coefficient estimates based on ordinary least squares estimation of (5.4) for $t = 1, 2, \ldots, T$.

Define

$$\begin{aligned}\delta^* &\equiv \frac{\delta}{1-\phi_1-\phi_2-\cdots-\phi_p}\\ \alpha^* &\equiv \frac{\alpha}{1-\phi_1-\phi_2-\cdots-\phi_p} - \frac{(\phi_1+2\phi_2+\cdots+p\phi_p)\,\delta^*}{1-\phi_1-\phi_2-\cdots-\phi_p}.\end{aligned}$$

Multiplying these equations by $(1-\phi_1-\phi_2-\cdots-\phi_p)$ and substituting the resulting expressions for α and δ into (5.4) produces

$$\begin{aligned} y_t &= (1-\phi_1-\phi_2-\cdots-\phi_p)\,\alpha^* + (\phi_1+2\phi_2+\cdots+p\phi_p)\,\delta^* \\ &\quad + (1-\phi_1-\phi_2-\cdots-\phi_p)\,\delta^* t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \end{aligned} \tag{5.5}$$

or

$$y_t = \alpha^* + \delta^* t + \phi_1^* y_{t-1}^* + \phi_2^* y_{t-2}^* + \cdots + \phi_p^* y_{t-p}^* + \epsilon_t, \tag{5.6}$$

where

$$\phi_j^* \equiv \phi_j \text{ for } j = 1, 2, \ldots, p$$

and

$$y^*_{t-j} \equiv y_{t-j} - \alpha^* - \delta^* \left(t - j\right) \text{ for } j = 1, 2, \ldots, p. \tag{5.7}$$

It is helpful to describe this transformation in more general notation that will also apply to more complicated models in the chapters that follow. The original regression model (5.4) can be written

$$y_t = \boldsymbol{x}_t^{'} \boldsymbol{\beta} + \epsilon_t, \tag{5.8}$$

where

$$\underset{(p+2)\times 1}{\boldsymbol{x}_t} \equiv \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \\ 1 \\ t \end{bmatrix} \qquad \underset{(p+2)\times 1}{\boldsymbol{\beta}} \equiv \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \\ \alpha \\ \delta \end{bmatrix}. \tag{5.9}$$

The algebraic transformation in arriving at (5.6) could then be described as rewriting (5.8) in the form

$$y_t = \boldsymbol{x}_t^{'} \boldsymbol{G}^{'} \left[\boldsymbol{G}^{'} \right]^{-1} \boldsymbol{\beta} + \epsilon_t = \left[\boldsymbol{x}_t^{*} \right]^{'} \boldsymbol{\beta}^{*} + \epsilon_t, \tag{5.10}$$

where

$$
\begin{aligned}
\underset{(p+2)\times(p+2)}{\boldsymbol{G}^{'}} &= \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ -\alpha^* + \delta^* & -\alpha^* + 2\delta^* & \cdots & -\alpha^* + p\delta^* & 1 & 0 \\ -\delta^* & -\delta^* & \cdots & -\delta^* & 0 & 1 \end{bmatrix} \\
\underset{(p+2)\times(p+2)}{\left[\boldsymbol{G}^{'}\right]^{-1}} &= \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ \alpha^* - \delta^* & \alpha^* - 2\delta^* & \cdots & \alpha^* - p\delta^* & 1 & 0 \\ \delta^* & \delta^* & \cdots & \delta^* & 0 & 1 \end{bmatrix}
\end{aligned}, \tag{5.11}
$$

$$\boldsymbol{x}_t^* \equiv \boldsymbol{G}\boldsymbol{x}_t = \begin{bmatrix} y_{t-1}^* \\ y_{t-2}^* \\ \vdots \\ y_{t-p}^* \\ 1 \\ t \end{bmatrix}, \tag{5.12}$$

and

$$\boldsymbol{\beta}^* \equiv \left[\boldsymbol{G}^{'}\right]^{-1}\boldsymbol{\beta} = \begin{bmatrix} \phi_1^* \\ \phi_2^* \\ \vdots \\ \phi_p^* \\ \alpha^* \\ \delta^* \end{bmatrix}. \tag{5.13}$$

Then, we can show that (e.g Hamilton's Sec 16.A)

$$\boldsymbol{\Gamma}_T\left(\boldsymbol{b}_T^* - \boldsymbol{\beta}^*\right) \xrightarrow{L} N\left(\mathbf{0}, \sigma^2\left[\boldsymbol{Q}^*\right]^{-1}\right), \tag{5.14}$$

where

$$\underset{(p+2)\times(p+2)}{\boldsymbol{\Gamma}_T} = \begin{bmatrix} \sqrt{T} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \sqrt{T} & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{T} & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sqrt{T} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & T^{3/2} \end{bmatrix}, \tag{5.15}$$

$$\underset{(p+2)\times(p+2)}{\boldsymbol{Q}^*} = \begin{bmatrix} \gamma_0^* & \gamma_1^* & \gamma_2^* & \cdots & \gamma_{p-1}^* & 0 & 0 \\ \gamma_1^* & \gamma_0^* & \gamma_1^* & \cdots & \gamma_{p-2}^* & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \gamma_{p-1}^* & \gamma_{p-2}^* & \gamma_{p-3}^* & \cdots & \gamma_0^* & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{2} & \frac{1}{3} \end{bmatrix} \tag{5.16}$$

for $\gamma_j^* \equiv E\left(y_t^* y_{t-j}^*\right)$.

What does this result imply about the asymptotic distribution of $\boldsymbol{b}$, the estimated coefficient vector for the OLS regression that is actually estimated? Writing out $\boldsymbol{b} = \boldsymbol{G}^{'}\boldsymbol{b}^{*}$ explicitly using (5.11), we have

(5.17)

$$\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \\ \hat{\alpha} \\ \hat{\delta} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ -\alpha^* + \delta^* & -\alpha^* + 2\delta^* & \cdots & -\alpha^* + p\delta^* & 1 & 0 \\ -\delta^* & -\delta^* & \cdots & -\delta^* & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\phi}_1{}^* \\ \hat{\phi}_2{}^* \\ \vdots \\ \hat{\phi}_p{}^* \\ \hat{\alpha}^* \\ \hat{\delta}^* \end{bmatrix}.$$

The OLS estimates $\hat{\phi}_j$ of the untransformed regression are identical to the corresponding coefficients of the transformed regression $\hat{\phi}_j^*$, so the asymptotic distribution of $\hat{\phi}_j$ is given immediately by (5.14). The estimate $\hat{\alpha}_T$ is a linear combination of variables that converge to a Gaussian distribution at rate $\sqrt{T}$, and so $\hat{\alpha}_T$ behaves the same way. Specifically, $\hat{\alpha}_T = \boldsymbol{g}_\alpha^{'} \boldsymbol{b}_T^*$, where

$$\boldsymbol{g}_\alpha^{'} \equiv \left[\begin{array}{cccccc} -\alpha^* + \delta^* & -\alpha^* + 2\delta^* & \cdots & -\alpha^* + p\delta^* & 1 & 0 \end{array} \right],$$

and so, from (5.14),

$$\sqrt{T}\left(\hat{\alpha}_T - \alpha\right) \xrightarrow{L} N\left(0, \sigma^2 \boldsymbol{g}_\alpha^{'} \left[\boldsymbol{Q}^*\right]^{-1} \boldsymbol{g}_\alpha\right). \tag{5.18}$$

Finally, the estimate $\hat{\delta}_T$ is a linear combination of variables converging at different rates:

$$\hat{\delta}_T = \boldsymbol{g}_\delta^{'} \boldsymbol{b}_T^* + \hat{\delta}_T^*,$$

where

$$\boldsymbol{g}_\delta^{'} \equiv \left[\begin{array}{cccccc} -\delta^* & -\delta^* & \cdots & -\delta^* & 0 & 0 \end{array} \right].$$

Its asymptotic distribution is governed by the variables with the slowest rate of convergence:

$$\begin{aligned}\sqrt{T}\left(\hat{\delta}_T-\delta\right) &= \sqrt{T}\left(\hat{\delta}_T^{*}+\boldsymbol{g}_\delta^{'}\boldsymbol{b}_T^{*}-\delta^{*}-\boldsymbol{g}_\delta^{'}\boldsymbol{\beta}^{*}\right)\\ &\overset{p}{\to} \sqrt{T}\left(\delta^{*}+\boldsymbol{g}_\delta^{'}\boldsymbol{b}_T^{*}-\delta^{*}-\boldsymbol{g}_\delta^{'}\boldsymbol{\beta}^{*}\right)\\ &= \boldsymbol{g}_\delta^{'}\sqrt{T}\left(\boldsymbol{b}_T^{*}-\boldsymbol{\beta}^{*}\right)\\ &\overset{L}{\to} N\left(0,\sigma^2\boldsymbol{g}_\delta^{'}\left[\boldsymbol{Q}^{*}\right]^{-1}\boldsymbol{g}_\delta\right).\end{aligned}$$

Thus, each of the elements of $\boldsymbol{b}_T$ individually is asymptotically Gaussian and $O_p\left(T^{-1/2}\right)$. The asymptotic distribution of the full vector $\sqrt{T}\left(\boldsymbol{b}_T-\boldsymbol{\beta}\right)$ is multivariate Gaussian, though with a singular variance-covariance matrix. Specifically, the particular linear combination of the elements of $\boldsymbol{b}_T$ that recovers $\hat{\delta}_T^{*}$, the time trend coefficient of the hypothetical regression,

$$\hat{\delta}_T^{*}=-\boldsymbol{g}_\delta^{'}\boldsymbol{b}_T^{*}+\hat{\delta}_T=\delta^{*}\hat{\phi}_{1,T}+\delta^{*}\hat{\phi}_{2,T}+\cdots+\delta^{*}\hat{\phi}_{p,T}+\hat{\delta}_T,$$

converges to a point mass around δ^* even when scaled by $\sqrt{T}$:

$$\sqrt{T}\left(\hat{\delta}_T^* - \delta^*\right) \xrightarrow{p} 0.$$

However, (5.14) establishes that

$$T^{3/2}\left(\hat{\delta}_T^* - \delta^*\right) \xrightarrow{L} N\left(0, \sigma^2\left(q^*\right)^{p+2,p+2}\right)$$

for $(q^*)^{p+2,p+2}$ the bottom right element of $\left[\boldsymbol{Q}^*\right]^{-1}$.

6. VAR

We consider an r-dimensional *vector autoregressive (VAR)* process $\{Y_t\}$ generated by

$$Y_t = \Pi_1 Y_{t-1} + \cdots + \Pi_p Y_{t-p} + u_t \tag{6.1}$$

where $\{u_t\}$ is MDS with $E\varepsilon_t\varepsilon_t' = \Omega$. The model is commonly written more compactly as

$$\Pi(L)\, Y_t = u_t$$

where L is the lag operator and

$$\Pi(x) = I - \Pi_1 x - \cdots - \Pi_p x^p.$$

Therefore, $\Pi(L)$ becomes a matrix consisting of polynomials in lag operator. The model (6.1) includes p lags, and for this reason, is called a p-th *order* VAR and denoted by VAR(p).

When and only when

$$\det \Pi(x) \neq 0 \quad \text{for} \quad |x| \leq 1,$$

which is called invertibility, $\{Y_t\}$, defined by (6.1), is stationary and has an MA representation

$$Y_t = \Phi(L)\, u_t$$

where

$$\Phi(x) = \sum_{i=0}^{\infty} \Phi_i x^i$$

with Φ_i's satisfying

$$\sum_{i=1}^{\infty} |\Phi_i|_{\infty} < \infty$$

where $|M|_{\infty} = \max |m_{pq}|$ for a matrix $M = (m_{pq})$. The MA coefficients Φ_i's can be obtained from the power series expansion of $\Pi(x)^{-1}$. This MA representation will be used in the impulse response analysis.

If $\det \Pi\,(1) = 0$, then some of the eigenvalues of $\Pi\,(1)$ are zero, which implies that some elements of Y_t is $I\,(1)$.

6.1. **Structural VAR (SVAR).** For r-dimensional time series $\{Y_t\}$, consider the model given by

$$B_0 Y_t = B_1 Y_{t-1} + \cdots + B_p Y_{t-p} + \varepsilon_t \tag{6.2}$$

with

$$var\,(\varepsilon_t) = \Lambda$$

a *diagonal* matrix.

Compare the model with the usual VAR

$$Y_t = \Pi_1 Y_{t-1} + \cdots + \Pi_p Y_{t-p} + u_t$$

with $var\,(u) = \Omega$. Model (6.2) is referred to as VAR in structural form (SF) or *structural* VAR (SVAR) and, in contrast, model (6.1) as VAR in reduced form (RF). The $\{u_t\}$ in RF VAR and the $\{\varepsilon_t\}$ in SF VAR are called, respectively, *reduced form errors* and *structural innovations*, which are related by

$$B_0 u_t = \varepsilon_t \tag{6.3}$$

Note the differences between SF and RF VAR's: First, VAR in SF allows for contemporaneous relationships in the components of Y_t, contrary to VAR in RF. Second, Ω for reduced form errors is unrestricted, while Λ for structural innovations is restricted to be diagonal.

SVAR in (6.2) can be extended to a more general form

$$C_0 Y_t = C_1 Y_{t-1} + \cdots + C_p Y_{t-p} + D\varepsilon_t \tag{6.4}$$

where contemporaneous relationships in the components of structural errors ε_t, as well as of Y_t, are permitted. The model may simply be regarded as SVAR (6.2) with

$$B_0 = D^{-1} C_0$$

and treated as such in our subsequent exposition.

6.1.1. *Sims (1980)'s Recursive model.*

m	money	$m = \varepsilon_m$
y/p	real GNP	$y/p = \beta_{21}m + \varepsilon_{y/p}$
u	unemployment	$u = \beta_{31}m + \beta_{32}y/p + \varepsilon_u$
w	wage level	$w = \beta_{41}m + \beta_{42}y/p + \beta_{43}u + \varepsilon_w$
p	price level	$p = \beta_{51}m + \beta_{52}y/p + \beta_{53}u + \beta_{54}w + \varepsilon_p$
pm	import price	$pm = \beta_{61}m + \beta_{62}y/p + \beta_{63}u + \beta_{64}w + \beta_{65}p + \varepsilon_{pm}$

This is also known as a triangular System as B_0 is a lower triangular matrix with unit diagonals. This system is *identified* in the sense that different parameter values of B and Λ are associated with different RF's so that different distribution of the data.

- Suppose there are more than one (B, Λ) which yields the same RF (Π, Ω). Then, they must belong to $\{PB, P\Sigma P' : |P| \neq 0\}$. Why?
- Suppose that there is a p.d. P and

$$\left(\tilde{B}_0, \tilde{B}_1, \cdots, \tilde{B}_p, \tilde{\Lambda}\right) = (PB_0, PB_1, \cdots, PB_p, P\Lambda P')$$

has the same reduced form as (B, Λ). Then, we get identification by showing that if $\tilde{B}_0$ is lower triangular with unit diagonals and $\tilde{\Lambda}$ is diagonal matrix then $P = I$. How?

- Since B_0 and $\tilde{B}_0 = PB_0$ should be lower triangular with unit diagonal, P should also be lower triangular with unit diagonal. Similarly, argue that since Λ and $\tilde{\Lambda} = P\Lambda P'$ should be diagonal and P is lower triangular with unit diagonal, we must have $P = I$.

Estimation of the recursive system is easy. The Cholesky decomposition of a given Ω produces the corresponding B_0 directly, that is, apply the decomposition to write

$$\Omega = LL',$$

where L is known to be uniqe. Then we have

$$B_0 = \Lambda^{\frac{1}{2}} L^{-1}.$$

As B_0 has unit diagonals, $\Lambda^{1/2}$ should be the diagonal matrix whose diagonals are the inverses of those of L^{-1}. Thus, let

$$\hat{\Omega} = \frac{1}{n}\hat{u}'\hat{u} = \hat{L}\hat{L}', \tag{6.5}$$

where $\hat{u}$ is the matrix stacking the OLS residuals $\hat{u}_t's$. Then, $\hat{\Lambda}$ is the diagonal matrix of the inverse of the diagonals of $\hat{L}^{-1}$ and

$$\hat{B}_0 = \hat{\Lambda}^{\frac{1}{2}}\hat{L}^{-1}.$$

6.2. **Impulse Response Analysis and Forecast Error Variance Decomposition.** Write Y_t in the MA representation

$$Y_t = \sum_{i=0}^{\infty}\Phi_i u_{t-i} = \sum_{i=0}^{\infty}\Phi_i^* \varepsilon_{t-i}^*,$$

where $\{\varepsilon_t^*\}$ are normalized $\{\varepsilon_t\}$ to have unit variances, and

$$\Phi_i^* = \Phi_i B_0^{-1}\Lambda^{\frac{1}{2}}. \tag{6.6}$$

Recall that u_t is the error from reduced form and ε_t the one from structural form. Then, the response of the p-th variable in the i-period ahead to an impulse in the q-th structural innovation is given by ${}_i\pi_{pq}$, where $\Pi_i = ({}_i\pi_{pq})$. Note that the unit shock in a component of $\{\varepsilon_t^*\}$ is identical to one standard deviation shock in the corresponding component of $\{\varepsilon_t\}$. This is the impulse response analysis (IRA).

Another popular approach to IRA is via the so-called "local projection".

6.3. **Identification.** In the parametric inference, the identification of the unknown parameters $\theta_{p\times 1}$ is defined through an *objective function,*

$$Q\left(x;\theta\right)=Q\left(\theta\right).$$

The function $Q(\theta)$ measures the risk or loss incurred by the decision to adopt the value θ.

Example 6.1. Consider the linear regression model

$$y_i=\beta' z_i+v_i,\ i=1,...,n.$$

If our objective function $Q\left(\beta\right)$ is

$$Q\left(\beta\right)=\left(Y-Z\beta\right)'\left(Y-Z\beta\right),$$

then

$$\widehat{\beta}=\arg\min_{\beta\in\mathbb{R}^p}Q\left(\beta\right),$$

which is the *least squares estimator* (LSE) of β. On the other hand, if we choose as our objective function

$$Q(\beta) = (Y - Z\beta)' \Sigma^{-1} (Y - Z\beta),$$

we then obtain the *generalized least squares estimator*

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} Q(\beta)$$

when $\Sigma = E(vv')$. If we knew the probability density function of v_i, then our objective function would become

$$Q(\theta) = -\log p\left(Y - \beta Z; \sigma^2\right),$$

where $\theta = \left(\beta', \sigma^2\right)' \in \mathbb{R}^p \times (0, \infty)$, and then you would be able to compute the *Maximum Likelihood Estimator* (MLE). The *Instrumental variable estimator*

(IVE) is obtained as

$$\begin{aligned} Q(\beta) &= (Y - Z\beta)' (WW') (Y - Z\beta) \\ \tilde{\beta}^{IV} &= \arg\min_{\beta \in \mathbb{R}^p} Q(\beta) \\ &= (W'Z)^{-1} W'Y. \end{aligned}$$

The pair (Q, Θ) is thought of as containing all the information that will be used to choose a θ-value. For a given data set X, two or more values give rise to the same Q, in which case one cannot choose between them by means of Q. Such an X may occur with small or zero probability. But if it exists for all possible X, there is some structural problem.

Definition 6.2. We say that two values θ_1 and $\theta_2 \in \Theta$ are observationally equivalent $(O.E.)$ with respect to Q, if

$$Q(x; \theta_1) = Q(x; \theta_2) \;\; \forall x \in \mathcal{X}.$$

Example 6.3. Suppose that

$$Q(\beta) = Y'Y + (\beta' Z' - 2Y') Z\beta,$$

and rank$(Z) < K$, that is, there is multicolinearity. Then,

$$Q(\beta) = Y'Y + (\beta' Z' - 2Y') Z\beta = Y'Y + \left((\beta + c)' Z' - 2Y'\right) Z(\beta + c) = Q(\beta + c),$$

and thus β and $\beta + c$ are O.E., yielding identification failure of LSE.

Definition 6.4. $\theta_1 \in \Theta$ is identifiable with respect to (Q, Θ) if there is not any other $\theta_2 \in \Theta$ such that $\theta_2 \neq \theta_1$ which is O.E. to θ_1 with respect to Q. Otherwise, θ_1 is unidentifiable.

Theorem 6.5. *(Sufficient condition for identifiability.) Let us assume that there exists a function $\phi(x)$ such that $\forall\ \theta \in \Theta$, we have that*

$$\theta = \int \phi(x) f(Q(x;\theta))\, dx, \tag{6.7}$$

where $f(\cdot)$ is some given function. Then, θ is identifiable with respect to $(Q;\theta)$.

Proof. Suppose the contrary. Then, we have that $\forall x \in \mathcal{X}$, $Q(x;\theta_1) = Q(x;\theta_2)$. The latter implies that

$$\phi(x) f(Q(x;\theta_1)) = \phi(x) f(Q(x;\theta_2)),$$

and thus by (6.7) we obtain that

$$\begin{aligned} \theta_1 &= \int \phi(x) f(Q(x;\theta_1))\, dx \\ &= \int \phi(x) f(Q(x;\theta_2))\, dx \\ &= \theta_2, \end{aligned}$$

which concludes the proof. □

Consider the linear Simultaneous Equations System (SEM):

$$By_t = Cz_t + u_t.$$

Let

$$\underset{n\times G}{Y = (y_1, ..., y_n)'}\,;\quad \underset{n\times K}{Z = (z_1, ..., z_n)'};\quad \underset{n\times(G+K)}{X = (x_1, ..., x_n)'} = (Y, Z)\,,$$

denote the matrix of observations, U that of the errors, and

$$A = (B, -C)\,.$$

Then,

$$XA' = YB' - ZC' = U.$$

Here, the structural form parameters are (A, Σ), whereas the reduced form parameters are (Π, Ω), where

$$\Pi = -B^{-1}C \quad \text{and} \quad \Omega = B^{-1}\Sigma(B^{-1})'.$$

Assume that

$$u_t \sim \text{iid } N\,(0, \Sigma)\,.$$

Then, the (negative of) log-likelihood becomes[1]

$$Q(\theta) = Q(A, \Sigma) = \frac{nG}{2} \log 2\pi + \frac{n}{2} \log \det(\Omega) + \frac{1}{2} \operatorname{tr} \left\{ (Y + Z\Pi') \Omega^{-1} (Y + Z\Pi')' \right\}.$$

It depends on A and Σ via Π and Ω.

Let

$$\Theta = \{\Pi, \Omega, \Omega > 0\} \quad \theta = vec[\Pi, \Omega],$$

and assume that there is no multicolinearity. Then, θ is identified. To see this, choose as our $\phi(x)$ the function

$$\phi(x) = vec\left(\widehat{\Pi}, \widehat{\Omega}\right),$$

where

$$\widehat{\Pi} = -Y'Z(Z'Z)^{-1} \quad \widehat{\Omega} = \frac{1}{n-K} Y' \left(I_n - Z(Z'Z)^{-1} Z'\right) Y.$$

[1] $\operatorname{tr}(ABCD) = \operatorname{vec}(C)'(D \otimes B')\operatorname{vec}(A')$, where the operator vec stacks columns of a matrix.

Now, choose as the objective function the log-likelihood, and as f

$$f = e^{-Q}.$$

Then, we obtain that

$$\int \phi\left(x\right) f\left(Q\left(x;\theta\right)\right) dx = E_{\theta}\left(vec(\widehat{\Pi},\widehat{\Omega})\right) = vec\left[E_{\theta}(\widehat{\Pi},\widehat{\Omega})\right] = \theta.$$

So, the parameters of the reduced form are identifiable if the model is not multicolinear.

The next issue to examine is when the parameters of the structural form are identifiable. Here, we focuse on the linear restrictions on A only. We have that

$$\Pi = -B^{-1}C, \quad \text{that is,} \quad B\Pi + \underset{G\times K}{C} = 0,$$

or equivalently

$$\underset{K\times G}{\Pi'} B' + C' = 0. \tag{2.2}$$

Thus, we have GK equations, but we have $G^2 + GK$ unknowns. Then,

$$\begin{aligned} vec\,(\Pi' B' + C') &= vec\,(\Pi' B' I_G) + vec\,(C') \\ &= (I_G \otimes \Pi')\, vec\,(B') + vec\,(C') \qquad (2.3) \\ &= ((I_G \otimes \Pi')\,; I_{GK}) \underset{G(G+K)\times 1}{\alpha} \end{aligned}$$

where

$$\alpha = (\beta', \gamma')', \; \beta = vec\,(B'), \text{ and } \gamma = vec\,(C').$$

From here, we see clearly that we have

$$G\,(G+K) \text{ unknowns}$$

$$GK \text{ equations.}$$

The latter implies that we need at least $G\,(G+K) - GK$ extra equations to be able to identify A. Let $W_{r\times G(G+K)}$ be a matrix (known) and $W\alpha = w$ vector. Then, a necessary condition to identify A is that $r \geq G^2$. With this new set of extra constraints on the parameters α, the system of equations becomes

$$\begin{aligned} \left[(I_G \otimes \Pi')\, ; I_{GK}\right] \alpha &= 0 \\ W\alpha &= w, \end{aligned}$$

or in matrix notation

$$\begin{pmatrix} V \\ W \end{pmatrix} \alpha = \Psi\alpha = \begin{pmatrix} 0 \\ \omega \end{pmatrix} \neq 0.$$

Theorem 6.6. *α is identified <u>iff</u> rank$(\Psi) = G\,(G+K)$ (rank condition).*

The condition $r \geq G^2$ is known as the <u>order condition</u>.

Let D denote $(I_G \otimes B; I_G \otimes C)$.

Theorem 6.7. *α identifiable iff rank$(WD') = G^2$.*

Proof. Note that

$$\begin{aligned}\Psi &= \begin{bmatrix} I_G \otimes \Pi' & I_{GK} \\ \multicolumn{2}{c}{W} \end{bmatrix} = \begin{bmatrix} I_G \otimes \Pi' & I_{GK} \\ W_1 & W_2 \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} 0_{G^2} & I_{GK} \\ WD' & W_2 \end{bmatrix}}_{\Psi_1} \underbrace{\begin{bmatrix} I_G \otimes (B')^{-1} & 0 \\ I_G \otimes \Pi' & I_{GK} \end{bmatrix}}_{\Psi_2}.\end{aligned}$$

Because $rank\,(B) = G$, it implies that $\text{rank}(\Psi_2) = G\,(G + K)$, and hence $rank\,(\Psi) = rank\,(\Psi_1)$.

However, I_{GK} is an invertible matrix, so

$$\begin{aligned} rank\,(\Psi_1) &= GK + rank\,(WD') \\ &= G\,(G + K) \end{aligned}$$

if and only if

$$G^2 = rank\,(WD')\,.$$

This concludes the proof. □

When parameters are subject only to linear constraints (not only in linear SEM) we could equivalently reparametrize in terms of a set of unrestricted parameters.

Normally, the type of constraints that we have among the parameters are *homogeneous*, that is, some linear combination of the parameters is zero. For example, the exclusion restriction is homogeneous since the corresponding row of W and w are given, respectively, by

$$(0,0,...,0,1,0,...,0)\,; \quad \text{and} \quad 0.$$

Another example is

$$(0,...,0,-1,0,...,0,1,0,...,0)\,; \quad \text{and} \ \ 0.$$

We always need at least one element of w to be different than zero, such as the normalization restriction where the corresponding element in w is 1.

Typically we use the normalization

$$diag\,(B) = (1,1,...,1)\,.$$

6.3.1. ***Identification of Subset.*** Denote by θ the underlying parameters of the system, that is

$$\theta = \left(\underset{1\times p_1}{\theta_1'}, \underset{1\times p_2}{\theta_2'} \right)'; \quad \theta_0 = (\theta_{01}'; \theta_{02}').$$

Definition 6.8. θ_{01} is identifiable w.r.t. (Q, Θ) iff all $\theta \in \Theta$ that are O.E. to θ_0 have the same value of θ_1.

Let $\alpha_1' = (\beta_1'; \gamma_1')$ be the parameters of the 1^{st} equation, e.g. the 1^{st} row of A. Then, in this case, what we have is that

$$\beta_1'\Pi + \gamma_1' = 0.$$

In this case, we have K equations and $G + K$ unknowns, so that we need at least G extra equations to be able to identify α_1.

Let r_1 be additional constraints on α_1,

$$W_1\alpha_1 = w_1.$$

Theorem 6.9. *(Rank condition) The parameters*

$$\alpha_1 = \begin{pmatrix} \beta_1 \\ \gamma_1 \end{pmatrix}$$

are identified iff

$$rank\,(W_1 A') = G.$$

The order condition is $r_1 \geq G$.

Proof. Write

$$\left.\begin{array}{c} \Pi'\beta_1 + \gamma_1 = 0 \\ W_1\alpha_1 = w_1 \end{array}\right\} = \begin{bmatrix} \Pi' & I_K \\ W_{11} & W_{12} \end{bmatrix} \alpha_1 = \begin{pmatrix} 0 \\ w_1 \end{pmatrix}.$$

Now,

$$\begin{pmatrix} \Pi' & I_K \\ \underset{r_1 \times G}{W_{11}} & \underset{r_1 \times K}{W_{12}} \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & I_K \\ W_1 A' & W_{12} \end{pmatrix}}_{\Psi_1} \times \underbrace{\begin{pmatrix} (B')^{-1} & 0 \\ \Pi' & I_K \end{pmatrix}}_{\Psi_2}.$$

Because $|B| \neq 0$, we have that Ψ_2 is a full rank matrix, and then

$$rank \begin{bmatrix} \Pi' & I_K \\ W_{11} & W_{12} \end{bmatrix} = rank \begin{pmatrix} 0 & I_K \\ W_1 A' & W_{12} \end{pmatrix}.$$

From here we conclude that the the left side of the last displayed equality is $G+K$ if and only if

$$rank(\Psi_1) = G + K = K + rank(W_1 A'),$$

which concludes the proof. □

Theorem 6.10. *Assume that in the 1^{st} equation all the constraints are zero, except the first one, $\alpha_{11} = 1$. Then α_1 is identifiable iff $r_1 \geq G$ and rank $(A^*) = G - 1$, where A^* is the matrix formed from the second $G-1$ rows corresponding to zeroes in α_1.*

Proof. Rearrange x_i as $\tilde{x}_i$, and A as $\widetilde{A} = \left(\widetilde{A}_1, \widetilde{A}_2\right)$, where all the restrictions are imposed in $\widetilde{A}_1 : G \times r_1$ and the normalization restriction is imposed on the $(1,1)$

element in $\widetilde{A}_1$. Then,

$$W_1 = \left(I_{r_1} \vdots 0 \right); \quad \widetilde{A}_1' = \left. \begin{matrix} 1 & \cdots \\ 0 & A^{*\prime} \end{matrix} \right\} r_1.$$

So,

$$W_1 \widetilde{A}' = \widetilde{A}_1'$$

and $rank\left(W_1 \widetilde{A}' \right) = G$ if and only if $rank\left(A^* \right) = G - 1$. □

Typically, this means validity of IVs.

Sometimes we have more constraints than we need for identification. θ_0 is said to be <u>overidentified</u> if ∃ two (or more) sets of prior constraints, each of which is capable of identifying θ_0, and the union of the sets is linearly independent. And θ_0 is <u>just-identified</u> if it is identified but not overidentified. Important as far as the efficiency of estimators of θ_0 is concerned.

6.3.2. ***Non-Linearity and Local Identification.*** In a linear SEM, there could be constraints in the variance-covariance matrix of the structural form error u_i, i.e.such as

$$\Sigma = \sigma^2 I \quad \text{or} \quad \underset{\text{(variance component models)}}{\Sigma = \lambda 1 1' + \sigma^2 I.}$$

If we have restrictions

$$w(A, \Sigma) = 0,$$

then we can use

$$B \Omega B' = \Sigma,$$

as well as

$$B \Pi + C = 0.$$

But $B \Omega B' = \Sigma$ is nonlinear in B.

Another Example is to put restrictions on the impulse response function (dynamic causal effect) of a SVAR model. See for example Blanchard and Quah (1989) and Gali (1992) . They typically concern the *long-run cumulative effects of a structural shock j on another variable i*. For instance, one may impose the hypothesis that the aggregate demand shocks do not have long-run effects on real GDP, or the long-run money neutrality. Given a SVAR and its Moving average representation (or impulse responce function),

$$A(L)\,y_t = u_t, \qquad \Rightarrow \qquad y_t = A(L)^{-1} u_t = D(L)\,u_t,$$

the restriction takes the form

$$D(1)_{ij} = 0.$$

Note that

$$D(1) = D_0 + D_1 + \cdots ,$$

which is the reason why it is called "long-run restriction." In the BQ's original work,

$$y_t = \begin{pmatrix} \Delta dgp_t \\ unempl_t \end{pmatrix} \text{ and } A(L) = B,$$

and the long-run effect is imposed by setting B as a lower-triangular matrix. With added assumption of the diagonality of Σ, this model becomes the recursive model. With lagged term in $A(L)$, the identification is more complex.

Suppose identifiability happens when there is a unique solution to a set of (possibly) nonlinear equations

$$\underset{q\times 1}{\psi}\,(\underset{p\times 1}{\theta}) = 0, \tag{6.8}$$

where ψ is a vector of given functions. Notice that in (6.8) we have suppressed any reference to known quantities.

We already know that if $\psi(.)$ is linear then there is either $\underline{\text{one}}$ unique solution to (6.8) or there are uncountable ones, (in fact, a continuous set of them). If $\psi(.)$ is nonlinear, then things are a bit different. We may have $\underline{\text{one}}$, $\underline{\text{uncountable}}$ as before, solutions but in addition to these we may have also countable or finite number of solutions.

Definition 6.11. θ_0 is locally identifiable (L.I.) w.r.t. (Q, θ) if $\exists$ an open neighbourhood of θ_0 containing no other θ which is O.E. to θ_0 w.r.t. Q.

Definition 6.12. θ_0 is globally identifiable (G.I.) if and only if it is identifiable for every neighbourhood of θ_0.

In the linear case, we have that $G.I. \equiv L.I.$. Indeed, if

$$W\alpha_1 = w \quad and \;\; W\alpha_2 = w,$$

we then have that

$$W\left(\lambda\alpha_1 + (1-\lambda)\,\alpha_2\right) = w$$

which it implies that

$$\lambda\alpha_1 + (1-\lambda)\,\alpha_2 = \alpha_3$$

also satisfies the constraints. So, now let λ vary to conclude.

If we can reduce Θ, then $L.I.$ becomes $G.I.$, e.g. using inequality restrictions.

Let

$$\Psi(\theta) = \frac{\partial}{\partial\theta'}\psi(\theta)\,, \quad \Psi_0 = \Psi(\theta_0)\,.$$

Definition 6.13. θ_0 is a regular point of $\Psi(\theta)$ if $\Psi(\theta)$ does not change its rank in a neighbourhood of θ_0.

For an example, consider a restriction

$$\theta = (\theta_1, \theta_2)'\,, \quad \psi(\theta) = \theta_1^2 + \theta_2^2 \quad (p = 2,\; q = 1)\,.$$

Thus $\theta = 0$ is identifiable, but

$$rank\left(\Psi\left(\theta\right)\right) = rank\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{cases} 0 & \text{if } \theta = 0 \\ 1 & \text{if } \theta \neq 0 \end{cases},$$

Then $\theta = 0$ is not a regular point of $\Psi\left(\theta\right)$.

Theorem 6.14. *Let θ_0 be a solution of* (6.8) *($\psi(\theta) = 0$). Let $\psi(\theta)$ be a continuous and differentiable function in a neighbourhood of θ_0. Then,*

(a) If $rank\left(\Psi_0\right) = p$, then θ_0 is L.I.

(b) If θ_0 is a regular point of $\Psi\left(\theta\right)$ and θ_0 is L.I., then

$$rank\left(\Psi_0\right) = p.$$

We need obviously $p \leq q$ for (a) and (b).

Proof. This proof is heuristic.

First, note that

$$\psi\left(\theta\right) = \psi\left(\theta_0\right) + \Psi_0\left(\theta - \theta_0\right) + O\left(\left|\theta - \theta_0\right|^2\right).$$

Thus, if θ_0 is not L.I., there exists a sequence θ_n such that $\psi(\theta_n) = \psi(\theta_0) = 0$. However,

$$\Psi_0 \frac{(\theta_n - \theta_0)}{|\theta_n - \theta_0|} = O(|\theta - \theta_0|) \to 0.$$

Furthermore, since $x'\Psi_0'\Psi_0 x$ is continuous in x and $\{x : |x| = 1\}$ is compact, $\min_{x:|x|=1} x'\Psi_0'\Psi_0 x$ exists. If Ψ_0 is of full column rank, there is no x such that $|x| = 1$ and $x'\Psi_0'\Psi_0 x = 0$. This yields contradiction and proves (a) □

Next we will study the case where there are constraints on Σ, e.g.

(6.9) $$B\Pi + C = 0$$

(6.10) $$B\Omega B' = \Sigma$$

(6.11) $$w(\theta) = 0,$$

where $\theta = vec(B', C', \Sigma)$. Then we have the following theorem.

Theorem 6.15. *Let θ_0 be a solution to $(6.9)-(6.11)$. Let $w(\theta)$ be a continuously differentiable function in a neighbourhood of θ_0. Denote*

$$W(\theta) = \frac{\partial w(\theta)}{\partial \theta'}$$

$$H(\theta) = W(\theta)\begin{pmatrix} I_G \otimes B' \\ I_G \otimes C' \\ I_G \otimes 2\Sigma \end{pmatrix}.$$

Then,

(a) *if $rank(H(\theta_0)) = G^2$, then θ_0 is L.I.*

(b) *if θ_0 is a regular point of $H(.)$ and θ_0 is L.I., then $rank(H(\theta_0)) = G^2$.*

Proof. Let

$$\theta = (\beta', \gamma', \sigma')' = [vec'(B'); vec'(C'); vec(\Sigma)],$$

then

$$\Psi(\theta) = \begin{bmatrix} I_G \otimes \Pi' & I_{GK} & 0 \\ \Delta & 0 & -I_{GG} \\ W_\beta & W_\gamma & W_\sigma \end{bmatrix},$$

where

$$W_\beta = \left[\frac{\partial w}{\partial \beta}\right];\ W_\gamma = \left[\frac{\partial w}{\partial \gamma}\right];\ W_\sigma = \left[\frac{\partial w}{\partial \sigma}\right].$$

But also $\Psi(\theta)$ can be written as

$$\Psi(\theta_0) = \begin{pmatrix} 0 & I_{GK} & 0 \\ 0 & 0 & -I_{GG} \\ H^* & W_\gamma & W_\sigma \end{pmatrix} \begin{pmatrix} I_G \otimes (B')^{-1} & 0 & 0 \\ I_G \otimes \Pi' & I_{GK} & 0 \\ -\Delta & 0 & I_{GG} \end{pmatrix}, \tag{2.5}$$

where

$$H^* = W_\beta (I_G \otimes B') + W_\gamma (I_G \otimes C') + W_\sigma (I_G \otimes 2\Sigma).$$

cf.[2]

$$\Delta = \frac{\partial \operatorname{vec}(B\Omega B')}{\partial \beta'} = 2\frac{\partial}{\partial \beta'}(I_G \otimes B\Omega)\beta = 2(I_G \otimes B\Omega) = 2\left(I_G \otimes \Sigma (B')^{-1}\right).$$

[2] $2\frac{\partial \operatorname{vec}(B\Omega B')}{\partial \beta'} = \frac{\partial \operatorname{vec}(B(\beta)\Omega B')}{\partial \beta'} + \frac{\partial \operatorname{vec}(B\Omega B(\beta)')}{\partial \beta'} = 2\frac{\partial \operatorname{vec}(B\Omega B(\beta)')}{\partial \beta'}$ due to the symmetry of $B\Omega B'$.

So, because the rank of the second matrix on the right of (2.5) is full rank, then

$$rank\left(\Psi\left(\theta_0\right)\right) = rank(1^{st} \text{ matrix})$$

which is

$$G\left(G+K\right) + rank\left(H^*\right).$$

So, we apply Theorem 6.14 here to the matrix

$$H^* = W\left(\theta\right)\begin{pmatrix} I_G \otimes B' \\ I_G \otimes C' \\ I_G \otimes 2\Sigma \end{pmatrix}$$

to conclude the proof. □

Global identification is difficult to establish in general. However, some restrictions may be helpful such as linearity, monotonicity. Revisit the recursive system and see Rubio-Ramírez et al (2010) as well.

6.4. **Cointegration.**

Definition 6.16. The $r \times 1$ series Y_t is cointegrated if Y_t is $I(1)$ yet there exists $r \times h$ matrix β whose rank is h and $z_t = \beta' Y_t$ is $I(0)$. The h vectors in β are called the cointegrating vectors.

If the series Y_t is not cointegrated, then $h = 0$. If $h = r$, then $Y_t = I(0)$. For $0 \leq h < r$, Y_t is $I(1)$ and cointegrated, and shares $r - h$ common stochastic trends. Due to Phillips (1991), it can be represented by a triangular system after reordering,

$$Y_{1t} = \Gamma Y_{2t} + z_t$$

$$Y_{2t} = Y_{2,t-1} + u_{2t},$$

where z_t is a h-dimensional vector of cointegrating relations, i.e. stationary, and u_{2t} is a $(r-h)$-dimensional stationary errors.

Let $\{Y_t\}$ be an r-dimensional VAR(p) process. More explicitly, we write

$$A(L)Y_t = \varepsilon_t,$$

where $A(L)$ is a matrix of lag polynomials given by

$$A(z) = I - A_1 z - \cdots - A_p z^p.$$

Theorem 6.17. ***(Granger-Johansen Representation)*** *Suppose the polynomial equation*

$$\det\ A(z) = 0$$

has m-roots at $z = 1$, where $m \leq r$, and all the other roots be outside the unit circle. Moreover, in the representation

$$A(z) = -z\,A(1) + (1 - z)\Gamma(z),$$

we assume that $\operatorname{rank} A(1) = h$*, where $h = r - m$, and $\Gamma(z)$ is nonsingular for all $|z| \leq 1$. Then, we may represent the long run impact matrices $A(1)$ as*

$$A(1) = \alpha\beta'$$

where β and α are $h \times r$ matrices of full column ranks.

If the rank of $A(1)$ is r, then $\det(A(1)) \neq 0$ which may imply that Y_t is $I(0)$. On the other hand, the rank is 0, then there is no cointegration. The model in this representation is also called the vector error correction model (VECM). The representation theorem was presented in Engle and Granger (1987) and updated in Johansen (2008).

6.4.1. ***Johansen's Cointegration Test.*** Its objective is testing general cointegrating relations using LR test. That is, testing

$$H_0 : h = h_0$$

against

$$H_1 : h = h_0 + h_1,$$

where h stands for the number of the cointegrating vector.

Write the model as

$$\Delta Y_t = A(1) Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \cdots + \Gamma_{p-1} \Delta Y_{t-p+1} + \varepsilon_t$$

where ε_t is iid $\mathrm{N}_r(0_r, \Sigma)$. Here $A(1)$ is the one in Granger representation above so that it includes all the informations regarding the stationarity of the series. Under the null hypothesis, we can write using the Granger Representation Theorem

$$A(1) = \alpha\beta'$$

where α and β are $r \times h_0$ matrices. Note that this representation is not unique so that we need a kind of normalization which will be mentioned below. Since ε_t follows iid Normal, the log-likelihood is

$$\mathcal{L}(\Sigma, \Gamma) = -\frac{Tr}{2}\log 2\pi - \frac{T}{2}\log|\Sigma| - \frac{1}{2}\sum_{t=1}^{T}\varepsilon_t'\Sigma^{-1}\varepsilon_t$$

First note that for a fixed β MLE is OLS under Normality and $\hat{\Sigma} = T^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t'$ where $\hat{\varepsilon}_t$ is the OLS residual. Since

$$\sum_{t=1}^{T}\hat{\varepsilon}_t'\hat{\Sigma}^{-1}\hat{\varepsilon}_t = tr\left(\sum_{t=1}^{T}\hat{\varepsilon}_t'\left(T^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t'\right)^{-1}\hat{\varepsilon}_t\right) = rT,$$

we write the concentrated likelihood as $\mathcal{L}(\beta)$,

$$\arg\max_{\beta} \mathcal{L}(\beta) \;=\; \arg\min_{\beta} \left| \sum_{t=1}^{T} \hat{\varepsilon}_t(\beta)\, \hat{\varepsilon}_t(\beta)' \right|$$

where $\hat{\varepsilon}_t(\beta)$ is the OLS residual of ΔY_t on $(1, \Delta Y_{t-1}, \cdots, \Delta Y_{t-p+1}, \beta' Y_{t-1})$. Let $\Delta\tilde{Y}_t$ and $\tilde{Y}_{t-1}$ be the regression residuals of

$$\Delta Y_t \text{ and } Y_{t-1} \text{ on } 1, \Delta Y_{t-1}, \cdots, \Delta Y_{t-p+1}, \tag{6.12}$$

respectively. Let

$$S_{00} = \sum_{t}^{n} \Delta\tilde{Y}_t \Delta\tilde{Y}_t', S_{01} = \sum_{t}^{n} \Delta\tilde{Y}_t \tilde{Y}_{t-1}', S_{11} = \sum_{t}^{n} \tilde{Y}_{t-1} \tilde{Y}_{t-1}'.$$

By FWL Theorem,

$$\begin{aligned} (6.13) \qquad & \min_{\beta}\left|\sum_{t=1}^{T}\hat{\varepsilon}_t(\beta)\,\hat{\varepsilon}_t(\beta)'\right| \\ = \; & \min_{\beta}\left|S_{00}-S_{01}\beta\left(\beta'S_{11}\beta\right)^{-1}\beta'S_{10}\right| \\ = \; & \left|S_{00}\right|\min_{\beta}\left|I-S_{00}^{-1}S_{01}\beta\left(\beta'S_{11}\beta\right)^{-1}\beta'S_{10}\right|, \end{aligned}$$

where the last equality follows from the fact that $|AB|=|A|\,|B|$.

Theorem 6.18. *We have*

$$\begin{aligned} & \min_{\beta}\left|I-\left(S_{00}\right)^{-1}\left(S_{01}\beta\right)\left(\beta'S_{11}\beta\right)^{-1}\left(\beta'S_{10}\right)\right| \\ = \; & \prod_{i=1}^{h_0}\left(1-\lambda_i\right), \end{aligned}$$

where $\lambda_1,\cdots,\lambda_{h_0}$ are the largest h_0 eigenvalues of the sample canonical correlations between ΔY_t and Y_{t-1} after controlling the constant and $\left(\Delta Y_{t-1},\ldots,\Delta Y_{t-p+1}\right)$,that *is, those of* $(S_{00})^{-1}(S_{01})(S_{11})^{-1}(S_{10})$. *And the MLE of* β*, which is normalized by* $\hat{\beta}'S_{11}\hat{\beta} = I$*, consists of the corresponding eigenvectors.*

NB As $\sum_{t=1}^{T}\hat{\varepsilon}_t(\beta)\hat{\varepsilon}_t(\beta)'$ is p.s.d, $(1-\lambda_i) \geq 0$ for all i.

Proof. If we impose the normalization assumption of β, then, since $|I_k - X'X| = |I_n - XX'|$ and the determinant of a matrix equals to the product of its eigenvalues,

$$\begin{aligned} & \left|I - (S_{00})^{-1}(S_{01}\beta)(\beta'S_{11}\beta)^{-1}(\beta'S_{10})\right| \\ = & \left|I_r - (S_{00})^{-1/2}(S_{01}\beta)(\beta'S_{10})(S_{00})^{-1/2}\right| \\ = & \left|I_{h_0} - \beta'S_{10}(S_{00})^{-1}S_{01}\beta\right| \\ = & \prod_{i=1}^{h_0}(I-\gamma_i), \end{aligned} \tag{6.14}$$

where $\gamma_i' s$ are eigenvalues of $\beta' S_{10} \left(S_{00}\right)^{-1} S_{01}\beta$. That is, γ_is solve the following eigenvalue problem

$$\left|\beta' S_{10} S_{00}^{-1} S_{01}\beta - \gamma_i I\right| = 0. \tag{6.15}$$

But, using the normalization assumption that $I = \beta' S_{11}\beta$,

$$\beta' \left(S_{10}S_{00}^{-1}S_{01} - \gamma_i S_{11}\right)\beta = \beta' S_{11}\left[\left(S_{11}^{-1}S_{10}S_{00}^{-1}S_{01} - \gamma_i I_r\right)\beta\right],$$

whose rank is smaller than h_0 by (6.15). As β and S_{11} are full column rank, $\left(S_{11}^{-1}S_{10}S_{00}^{-1}S_{01} - \gamma_i I_r\right)$ should have a deficient rank, i.e., $\left|S_{11}^{-1}S_{10}S_{00}^{-1}S_{01} - \gamma_i I_r\right| = 0$, which in turn implies that γ_i is one of eigenvalues of the canonical correlation. Then, γ_i should be one of the h_0 largest eigenvalues of the canonical correlation to minimize (6.14), which can be obtained if β are the corresponding eigenvectors. If one of the columns of β is the corresponding eigenvector of γ_i, then a column of the matrix in the braces [] is zero, i.e., we can say its determinant is zero. ■

Therefore the achieved maximum likelihood is

$$\mathcal{L}\left(\hat{\Sigma},\hat{\Gamma}\right)=-\frac{T}{2}\log 2\pi\left|S_{00}\right|-\frac{1}{2}rT-\frac{T}{2}\log\left|\prod_{i=1}^{h_0}\left(1-\lambda_i\right)\right|$$

and the MLE of β is the matrix of the corresponding eigenvectors.

If we use only h_0 informations of the sample canonical correlations, it is natural to pick out its largest eigen values to minimize the determinant in (6.13). Or to minimize the prediction error we use the linear combinations which show the highest correlations to $\Delta\bar{Y}$.

If the null hypothesis is correct, the rank of the sample canonical correlations should be equal to h_0 in large enough sample. Hence

$$\mathcal{H}_0:\lambda_{h_0+1}=\cdots=\lambda_{h_0+h_1}=0.$$

We can directly test this finding using LR test statistic.

From the above we can write

$$
\begin{aligned}
LR &= 2\left(\mathcal{L}_1^* - \mathcal{L}_0^*\right) \\
&= -T \sum_{j=h_0+1}^{h_0+h_1} \log\left(1-\lambda_j\right) \\
&= T \sum_{j=h_0+1}^{h_0+h_1} \lambda_j + o_p(1)
\end{aligned} \tag{6.16}
$$

where $\mathcal{L}_1^*$ and $\mathcal{L}_0^*$ are the attained maximum likelihood under H_1 and H_0 respectively. The last equality holds because the $\lambda_j\,'s, j = h_0+1, \ldots, h_0+h_1$ should be close to 0 under H_0. Equation(6.16) shows why the LR statistic is sometimes called the ”Maximal eigenvalue statistic” when $h_0 = 0$ and $h_1 = 1$ and the ”Trace statistic” when $h_0 = 0$ and $h_1 = r$.

The critical values differ depending on whether we include a constant in the regression (6.12) as in the unit root testing. The tables can be found in e.g., Hamilton (1994) .

It was extended to high-dimensional VARs by Onatski and Wang (2018).

6.5. **Factor-Augmented VAR.** The factor-augmented VAR (FAVAR), Bernanke et al. (2005), is often employed to overcome the dimensionality-issue of the VAR:

$$\Pi(L)\, Y_t = u_t,$$

where some of the elements in Y_t are the latent factors f_t such that

$$x_{it} = \lambda_i' F_t + e_{it}, \tag{6.17}$$

for $i = 1, ..., N, \quad t = 1, ..., T$, where λ_i is a r-dimensional column vector, referred to as the factor loading, F_t is the *unobserved* common factor, and e_{it} is the idiosyncratic error.

More recently, more general transformations of F_t and its lagged ones are being used to augment the macro regressions emphasizing the role of the uncertainty, e.g. Jurado et al. (2015). While oftentime x_{it} are macro and financial time series, more and more micro-level data is being utilized, e.g. Bloom (2018).

In this model, a few factors drive comovements of a high-dimensional time series variables x_t. The idiosyncratic shocks account for measurement error and special features that are specific to an individual series. For instance, in finance, the return x_{it} on the asset i at the time t is determined by the exposure λ_i to the systemetic risk F_t and the idiosyncratic returns. The term $C_{it} = \lambda_i' F_t$ is called the common component of the model.

In vector and matrix form,

$$X_t = \Lambda F_t + e_t \quad : \quad N \times 1$$

and

$$X = F\Lambda' + e \quad : \quad T \times N.$$

Although the factor F_t is subject to certain dynamic system such as

$$A(L) F_t = u_t,$$

the above model is static because only the current F_t appears in the model. On the contrary, the dynamic model is given by

$$x_{it} = \lambda_i (L)' f_t + e_{it}.$$

The factor model exhibits the following covariance structure

$$\Sigma := EX_tX_t' = \Lambda\Lambda' + \Omega,$$

where $\Omega = Ee_te_t'$, under the identifying assumption of

$$EF_tF_t' = I_r. \tag{6.18}$$

Furthermore, assume that

$$\frac{1}{N}\Lambda'\Lambda \to V_\lambda > 0 \quad \text{and} \quad \lambda_{max}\left(\Omega\right) \leq c < \infty.$$

If Ω is diagonal, the model is known as the *strict* factor model and as the *approximate* factor model otherwise.

A key feature of Σ when $N \to \infty$ is that its largest r eigenvalues increase with N while other eigenvalues and those of Ω stay bounded.

Another way to look at it is that the scaled sum of common components across i does not vanish with N while that of e_{it} averages out by the weak law of large numbers as they are idiosyncratic shocks.

The restriction (6.18) is a normalization restriction and λ_i and F_t cannot be separately identified without it. That is, for any p.d. Q, $\lambda_i' F_t = \tilde{\lambda}_i' \tilde{F}_t$ where $\tilde{\lambda}_i = Q' \lambda_i$ and $\tilde{F}_t = Q^{-1} F_t$. So, we need additional $r\,(r-1)\,/2$ restrictions for identification. Let

$$\Lambda' \Lambda \quad \text{be diagonal.}$$

Alternatively, we can set $EF_t' F_t$ as diagonal and $\Lambda \Lambda' / N = I_r$.

The method of the principal component yields an estimator for the common factors and the factor loadings for a prespecified r, which is the least squares estimation under the normalization restriction:

$$\min_{F,\Lambda} \sum_{i,t} \left(x_{it} - \lambda_i' F_t\right)^2$$

$$\text{s.t.} \quad F' F / T = I_r \text{ and } \Lambda' \Lambda \text{ is diagonal.}$$

The solution can be given by the eigenvalue decomposition of the matrix

$$\underbrace{XX'}_{T\times T} = BDB',$$

where D is the diagonal matrix of eigenvalues in descending order and B the matrix of orthonormal eigenvectors such that $B'B = I_T$. The $\sqrt{T}$ times of the first r columns of B are the estimated factors $\hat{F}_t$. For the given estimated factors $\hat{F}_t$, the loading Λ can be estimated by the OLS, that is,

$$\hat{\Lambda}' = \hat{F}'X/T$$

given the normalization $\hat{F}'\hat{F}/T = I_r$.

Note that this estimation method needs a balanced panel. See the MLE in Stock and Watson (2016). The estimated factors are used for forecasting, in the factor augmented regressions such as FAVAR, and instrumental variable regressions.

A key issue in practice is how to determine the number of factors r. Several methods were proposed such as the information criteria by Bai and Ng (2002) and the eigenvalue ratio test by Ahn and Horenstein (2009).

An important avenue for future research is to accomodate unstabilities in the factor loading over time such as breaks (Cheng and Schorfheide 2016 REstud), multiple regimes.

With economic time series data, the presence of unit root and/or cointegration in the series is another important issue. There are methods to take care of this directly such as Bai and Ng (2004) but it is also often the case that the data are differenced and standardized in advance.

References

Andrews, Donald WK (1991). "Heteroskedasticity and autocorrelation consistent covariance matrix estimation". In: *Econometrica: Journal of the Econometric Society*, pp. 817–858.

Berk, Kenneth N (1974). "Consistent autoregressive spectral estimates". In: *The Annals of Statistics*, pp. 489–502.

Bertsimas, Dimitris, Angela King, and Rahul Mazumder (2016). "Best subset selection via a modern optimization lens". In.

Davidson, James (1994). *Stochastic limit theory: An introduction for econometricians*. OUP Oxford.

Jansson, Michael (2002). "Consistent covariance matrix estimation for linear processes". In: *Econometric Theory* 18.6, pp. 1449–1459.

Kiefer, Nicholas M and Timothy J Vogelsang (2002). "Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation". In: *Econometrica* 70.5, pp. 2093–2095.

Kiefer, Nicholas M and Timothy J Vogelsang (2005). “A new asymptotic theory for heteroskedasticity-autocorrelation robust tests”. In: *Econometric Theory* 21.6, pp. 1130–1164.

Lazarus, Eben et al. (2018). “HAR inference: Recommendations for practice”. In: *Journal of Business & Economic Statistics* 36.4, pp. 541–559.

Sun, Yixiao, Peter CB Phillips, and Sainan Jin (2008). “Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing”. In: *Econometrica* 76.1, pp. 175–194.

Tibshirani, Robert (1996). “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288.

Wang, Hansheng, Guodong Li, and Chih-Ling Tsai (2007). “Regression coefficient and autoregressive order shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69.1, pp. 63–78.

Email address: `myunghseo@snu.ac.kr`