Modes of Convergence

Almost sure convergence

Consider a random sample Z_t, $t = 1, 2, ...$, and let ω denote the entire sequence $Z_1, Z_2, ...$, i.e. for any draw $\omega \in \Omega$, with Ω denoting the underlying probability space, we observe a different realization of the random sequence. Typically, we are interested in sample averages

$$b_T(\omega) = T^{-1} \sum_{t=1}^{T} Z_t.$$

We say that $b_T(\omega)$ converges almost surely to b, i.e. $b_T(\omega) \overset{a.s.}{\to} b$, or simply $b_T \overset{a.s.}{\to} b$ if:

$$P\left(\omega : b_T(\omega) \to b\right) = 1,$$

where P is the probability measure defined on Ω, which describes the distribution of ω and so determines the joint distribution of the entire sequence $\{Z_t\}$. Herafter, with the notation $\{Z_t\}$ we mean Z_t, with $t = 1, 2, ...$
Note that $b_T \overset{a.s.}{\to} b$, if the probability of observing a sequence $\{Z_t\}$ such that $T^{-1} \sum_{t=1}^{T} Z_t$ does not converge to b is zero.

Almost sure convergence for vectors and matrices is equivalent to almost sure convergence of each element.

Property AS1: If g is a function which is continuous at b, then $b_T \overset{a.s.}{\to} b$ implies that $g(b_T) \overset{a.s.}{\to} g(b)$.

Usefulness of this property: if $\mathbf{X'X}/T \overset{a.s.}{\to} M$ and $\det(M) > 0$, then $(\mathbf{X'X}/T)^{-1} \overset{a.s.}{\to} M^{-1}$. This is because the inversion of a positive definite matrix is a continous transformation.

Sometine in the case of heterogeneous observations, there is no a fixed value (scalar, vector, etc.) b to which b_T converges almost sure, though there exists a deterministic sequence c_T, uniformly bounded, i.e. $\sup_T |c_T| < \Delta < \infty$, such that

$$P\left(\omega : b_T(\omega) - c_T \to 0\right) = 1,$$

or $b_T(\omega) - c_T \overset{a.s.}{\to} 0$.

Uniform Continuity: If for any $\epsilon > 0$, there exists a $\delta(\epsilon)$, such that for any for any $a, b \in B$, whenever $|a - b| < \delta(\epsilon)$, $|g(a) - g(b)| < \epsilon$, with δ dependent on ϵ but not on b, then g is said to be uniformly continuous on B.

Uniform continuity implies pointwise continuity, but not the other way round. Though, if B is a compact set (i.e. a set which is bounded and closed), then pointwise continuity is equivalent to uniform continuity.

Property AS2: If g is a function which is uniform continuous in B, $\sup_T |c_T| < \Delta < \infty$, then $b_T - c_T \overset{a.s.}{\to} 0$ implies that $g(b_T) - g(c_T) \overset{a.s.}{\to} 0$.

Usefulness of this property: if $\mathbf{X}'\mathbf{X}/T - M_T \xrightarrow{a.s.} 0$ and $\inf_T \det(M_T) > 0$, then $(\mathbf{X}'\mathbf{X}/T)^{-1} - M_T^{-1} \xrightarrow{a.s.} 0$. This is because the inversion of a uniformly positive definite matrix is a uniform continous transformation.

A few definitions. (i) We say that b_T is $O_{a.s.}(1)$ (or almost surely bounded) if there exists $0 < \Delta < \infty$, and T_0, such that for all $T > T_0$, $P(\omega : |b_T(\omega)| < \Delta) = 1$.

(ii) b_T is said to be at most of almost sure order T^λ, i.e. $O_{a.s.}(T^\lambda)$, if there exists $0 < \Delta < \infty$, and T_0, such that for all $T > T_0$, $P\left(\omega : T^{-\lambda} |b_T(\omega)| < \Delta\right) = 1$

(iii) b_T is said to be almost sure order smaller than T^λ, i.e. $o_{a.s.}(T^\lambda)$, if $T^{-\lambda} b_T \xrightarrow{a.s.} 0$. Thus, if $b_T(\omega) - c_T \xrightarrow{a.s.} 0$, then $b_T(\omega) - c_T = o_{a.s.}(1)$.

(iv) A deterministic sequence c_T is said to be respectively $O(1), O(T^\lambda), o(T^\lambda)$ if there exists $0 < \Delta < \infty$, and T_0, such that for all $T > T_0$, $|c_T| < \Delta$, $T^{-\lambda} |c_T| < \Delta$, and $T^{-\lambda} c_T \to 0$

Convergence in Probability.

Let b_T be a sequence of random variables. We say that $b_T(\omega)$ converges in probablity to b, i.e. $b_T(\omega) \xrightarrow{p} b$, or simply $b_T \xrightarrow{p} b$, or $p\lim b_T = b$ if:

$$P(\omega : b_T(\omega) \to b) \to 1.$$

With almost sure convergence, the probability measure P controls the entire sequence $\{Z_t\}$, while with convergence in probability it controls only the first T elements.

Note that $b_T \xrightarrow{p} b$, if the probability of observing a sequence $\{Z_t\}$ such that $T^{-1} \sum_{t=1}^{T} Z_t$ does not converge to b becomes less and less likely as T increases.

Almost sure convergence implies convergence in probability, but not the other way round.

Property PR1: If g is a function which is continuous at b, then $b_T \xrightarrow{p} b$ implies that $g(b_T) \xrightarrow{p} g(b)$.

Usefulness of this property: if $\mathbf{X}'\mathbf{X}/T \xrightarrow{p} M$ and $\det(M) > 0$, then $(\mathbf{X}'\mathbf{X}/T)^{-1} \xrightarrow{p} M^{-1}$. This is because the inversion of a positive definite matrix is a continous transformation.

Property PR2: If g is a function which is uniform continuous in B, $\sup_T |c_T| < \Delta < \infty$, then $b_T - c_T \xrightarrow{p} 0$ implies that $g(b_T) - g(c_T) \xrightarrow{p} 0$.

Another few definitions. (i) We say that b_T is $O_p(1)$ (or bounded in probability) if there exists $0 < \Delta < \infty$, and T_0, such that for all $T > T_0$, $P(\omega : |b_T(\omega)| < \Delta) \to 1$.

(ii) b_T is said to be at most of probability order T^λ, i.e. $O_p(T^\lambda)$, if there exists $0 < \Delta < \infty$, and T_0, such that for all $T > T_0$, $P\left(\omega : T^{-\lambda} |b_T(\omega)| < \Delta\right) \to 1$

(iii) b_T is said to be of probability order smaller than T^λ, i.e. $o_p(T^\lambda)$, if $T^\lambda b_T \xrightarrow{p} 0$. Thus, if $b_T(\omega) - c_T \xrightarrow{p} 0$, then $b_T(\omega) - c_T = o_p(1)$.

Product Rule: If A_T is $o_P(1)$ and b_T is $O_P(1)$, then $A_T b_T = o_P(1)$. That is, the product of something which is bounded in probability times something which goes in probability to zero, goes in probability to zero. Analogously, if A_T is $o_{a.s.}(1)$ and b_T is $O_{a.s.}(1)$, then $A_T b_T = o_{a.s.}(1)$.

Convergence in $r-$th Mean.

Let b_T be a sequence of random variables. We say that $b_T(\omega)$ converges in $r-$th mean to b, $r > 0$, if

$$E\left(|b_T(\omega) - b|^r\right) \to 0 \text{ as } T \to \infty.$$

We also say that $b_T \overset{r.m.}{\to} b$.

Property RM1: If $b_T \overset{r.m.}{\to} b$, then for all $s < r$, $b_T \overset{s.m.}{\to} b$, i.e. convergence in r-th mean implies convergence in s-th mean for all $s < r$.

In order to prove the Property above, we need the following,

Jensen's Inequality: If g is a convex (concave) function on a set B, then for any random variable Z, such that $P(Z \in B) = 1$, it follows that $g(E(Z)) \leq E(g(Z))$ $(g(E(Z)) \geq E(g(Z)))$.

Now, for $s < r$,

$$\begin{aligned} E\left(|b_T(\omega) - b|^s\right) &= E\left(\left(|b_T(\omega) - b|^r\right)^{s/r}\right) \\ &\leq \left(E\left(\left(|b_T(\omega) - b|^r\right)\right)\right)^{s/r}, \end{aligned}$$

thus if $E\left(|b_T(\omega) - b|^r\right) \to 0$ also $E\left(|b_T(\omega) - b|^s\right) \to 0$.

Property RM1: If for $r > 0$, $b_T \overset{r.m.}{\to} b$, then $b_T \overset{p}{\to} b$, i.e. convergence in r-th mean implies convergence in probability.

In order to show the property above, we need the following,

Generalized Chebyshev Inequality (or Markov Inequality): Given a random variable Z, such that $E(|Z|^r) < \infty$, then for any $r > 0$, there exists $\varepsilon > 0$ such that,

$$P\left(|Z| > \varepsilon\right) \leq \frac{1}{\varepsilon^r} E(|Z|^r).$$

Now,

$$P\left(|b_T - b| > \varepsilon\right) \leq \frac{1}{\varepsilon^r} E(|b_T - b|^r) \to 0.$$

For later...Given a random sequence $\{Z_t\}$,

(i) if $T^{-1}\sum_{t=1}^{T}(Z_t - E(Z_t)) \overset{a.s.}{\to} 0$, we say that $\{Z_t\}$ satisfy a **Strong Law of Large Numbers**

(ii) if $T^{-1}\sum_{t=1}^{T}(Z_t - E(Z_t)) \overset{p}{\to} 0$, we say that $\{Z_t\}$ satisfy a **(Weak) Law of Large Numbers**

Convergence in Distribution

Let $b_T(\omega)$ be a sequence of random vectors, with distribution F_T. If, as $T \to \infty$, $F_T(z) \to F(z)$, for any continuity point z of F, where F is the distribution function of Z, we say that b_T converges in distribution to Z, i.e. $b_T \xrightarrow{d} Z$.

If Z is a normal random variable, then b_T is said to be *Asymptotically Normal.*

If $b_T \xrightarrow{p} b$, then b_T converges in distribution to a random variable Z which takes the value b with probability one.

.

Continuous Mapping Theorem: If $b_T \xrightarrow{d} Z$, then for any continuous function g, $g(b_T) \xrightarrow{d} g(Z)$.

Property CD1: If $b_T \xrightarrow{d} Z$, then $b_T = O_P(1)$, i.e. random sequence converging in distribution, is bounded in probability.

Proof: as $b_T \xrightarrow{d} Z$, $P(|b_T| > \Delta) \to P(|Z| > \Delta)$, and so for any Δ, there exists $\delta > 0$, such that

$$\lim P(|b_T| > \Delta) = P(|Z| > \Delta) < \delta.$$

Product Rule (again): We have see that if $A_T = o_P(1)$ and $b_T = O_P(1)$, then $A_T b_T = o_P(1)$. Thus, if $A_T \xrightarrow{p} 0$ and $b_T \xrightarrow{d} Z$, then $A_T b_T \xrightarrow{p} 0$.

Asymptotic Equivalence: If $a_T - b_T = o_P(1)$, and $b_T \xrightarrow{d} Z$, then $a_T \xrightarrow{d} Z$.

Very useful result: crucial in showing the asymptotic normality of estimators and test statistics.

Choleski decomposition: Any positive (semi) definite matrix V, can be decomposed as $V = V^{1/2}V^{1/2}$, where $V^{1/2}$ does not need to be unique.

Asymptotic Covariance Matrix: Let V_T be a $k \times k$ matrix which is positive definite for all $T > T_0$, if $\mathbf{V}_T^{-1/2} b_T \xrightarrow{d} N(0, I_K)$, then $\mathbf{V}_T$ is said to be the asymptotic covariance matrix of b_T, or $\mathbf{V}_T = avar(b_T)$. Note that if the smallest eigenvalue of $\mathbf{V}_T$ and $\mathbf{V}_T^{-1}$ are bounded away from zero, then $\mathbf{V}_T = avar(b_T) = var(b_T)$.

Quadratic Forms: If $\mathbf{V}_T^{-1/2} b_T \xrightarrow{d} N(0, I_K)$, then $b_T' \mathbf{V}_T^{-1} b_T \xrightarrow{d} \chi_k^2$ (immediate from continuous mapping theorem).

For later...Given a random sequence $\{Z_t\}$,

(i) if $\mathbf{V}_T^{-1/2} T^{-1/2} \sum_{t=1}^{T} (Z_t - E(Z_t)) \xrightarrow{d} N(0, I_K)$ we say that $\{Z_t\}$ satisfy a **Central Limit.**

Consistency and Asymptotic Normality of OLS

Linear Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{\dagger} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is $T \times 1$, $\mathbf{X}$ is $T \times k$, $\boldsymbol{\beta}^{\dagger}$ is $k \times 1$ and $\boldsymbol{\epsilon}$ is $T \times 1$, or

$$y_t = \mathbf{X}_t'\boldsymbol{\beta}^{\dagger} + \epsilon_t, \tag{2}$$

with y_t, ϵ_t scalars, $\mathbf{X}_t$ and $\boldsymbol{\beta}^{\dagger}$ $k \times 1$. Hereafter, $\mathbf{X}_t = (1, X_{2,t}, ..., X_{k,t})$.

We say that the linear model is *correctly specified for* $E(y_t|\mathbf{X}_t)$ if

$$E(y_t|\mathbf{X}_t) = \mathbf{X}_t'\boldsymbol{\beta}^{\dagger}.$$

Note that correct specification for $E(y_t|\mathbf{X}_t)$ is EQUIVALENT to $E(\epsilon_t|\mathbf{X}_t) = 0$. In the case of correct specification, $\boldsymbol{\beta}^{\dagger}$ is the parameter vector of the conditional expectation. For example, if $(y_t, \mathbf{X}_t)$ is jointly normal, then $y_t|\mathbf{X}_t \simeq N(\mathbf{X}_t'\boldsymbol{\beta}^{\dagger}, \sigma_{\epsilon}^2)$, and so the linear model is correctly specified. Otherwise, there is no particular reason why the linear model is correctly specified for $E(y_t|\mathbf{X}_t)$.

Now, $E(y_t|\mathbf{X}_t)$ implies that $E(\mathbf{X}_t\epsilon_t) = 0$, in fact, by the *Law of the Iterated Expectations,*

$$\begin{aligned} E(\mathbf{X}_t\epsilon_t) &= E\left(E(\mathbf{X}_t\epsilon_t|\mathbf{X}_t)\right) = E(\mathbf{X}_t E(\epsilon_t|\mathbf{X}_t)) \\ &= 0 \text{ if } E(\epsilon_t|\mathbf{X}_t) = 0. \end{aligned}$$

If we simply assume that $E(\mathbf{X}_t\epsilon_t) = 0$, then the linear model is not necessarily correctly specified, and $\boldsymbol{\beta}^{\dagger}$ is not necessarily the parameter of the conditional expectation, though $\boldsymbol{\beta}^{\dagger}$ can be interpreted as the **best linear predictor.**

Provided $\left(E(\mathbf{X}_t\mathbf{X}_t')\right)^{-1}$ exists, i.e. no perfect collinearities, note that by premultiplying (2) by $\mathbf{X}_t$ and taking the expectation, we have that

$$E\left(\mathbf{X}_t y_t\right) = E(\mathbf{X}_t\mathbf{X}_t')\boldsymbol{\beta}^{\dagger} + E(\mathbf{X}_t\epsilon_t),$$

so that if $E(\mathbf{X}_t\epsilon_t) = 0$,

$$\boldsymbol{\beta}^{\dagger} = \left(E(\mathbf{X}_t\mathbf{X}_t')\right)^{-1} E\left(\mathbf{X}_t y_t\right).$$

Now, it is immediate to see that $\boldsymbol{\beta}^{\dagger}$ can be also defined as:

$$\boldsymbol{\beta}^{\dagger} = \arg\min_{\beta} \frac{1}{T}\sum_{t=1}^{T} E\left(y_t - \mathbf{X}_t'\boldsymbol{\beta}\right)^2$$

and so is the best linear predictor. In fact, by the first order conditions, recalling that $E\left(\mathbf{X}_t\left(y_t - \mathbf{X}_t'\boldsymbol{\beta}^{\dagger}\right)\right) = 0$ for all $t, 2\frac{1}{T}\sum_{t=1}^{T} E\left(\mathbf{X}_t\left(y_t - \mathbf{X}_t'\boldsymbol{\beta}^{\dagger}\right)\right) = 0$, which indeed implies, $\boldsymbol{\beta}^{\dagger} = \left(E(\mathbf{X}_t\mathbf{X}_t')\right)^{-1} E\left(\mathbf{X}_t y_t\right)$.

Define the OLS (Ordinary least Square Estimator), $\widehat{\beta}_T$, where

$$\widehat{\beta}_T = \arg\min_{\beta} \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \mathbf{X}_t'\boldsymbol{\beta}\right)^2,$$

and so

$$\widehat{\boldsymbol{\beta}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t y_t\right). \tag{3}$$

Assumption OLS-1

(i) $y_t = \mathbf{X}_t'\boldsymbol{\beta}^\dagger + \epsilon_t$, with $E(\mathbf{X}_t\epsilon_t) = 0$.

(ii) $\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t \xrightarrow{a.s.} E(\mathbf{X}_t\epsilon_t) = 0$ (strong law of large numbers)

(iii) $\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t' - \mathbf{M}_T \xrightarrow{a.s.} 0$, where $\mathbf{M}_T = \frac{1}{T}\sum_{t=1}^{T}E\left(\mathbf{X}_t\mathbf{X}_t'\right) = O(1)$, with $\inf_T \det \mathbf{M}_T > 0$ (strong law of large numbers)

(iv) $\mathbf{V}_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t \xrightarrow{d} N(0, I_k)$, with $\mathbf{V}_T = var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right) = O(1)$ and $\inf_T \det \mathbf{V}_T > 0$ (central limit theorem for the score)

Theorem OLS-1:

(a) Given Assumption OLS-1(i)-(iii), then

$$\widehat{\boldsymbol{\beta}}_T \xrightarrow{a.s.} \boldsymbol{\beta}^\dagger$$

(b) Given Assumption OLS1-(i)-(iv), then

$$\mathbf{D}_T^{-1/2}T^{1/2}\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right) \xrightarrow{d} N(0, I_k),$$

where $\mathbf{D}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1} V_T \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}$.

(c) Given Assumption OLS-1(i)-(iv), if there exists $\widehat{\mathbf{V}}_T - \mathbf{V}_T \xrightarrow{pr} 0$, then

$$\widehat{\mathbf{D}}_T^{-1/2}T^{1/2}\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right) \xrightarrow{d} N(0, I_k),$$

where $\widehat{\mathbf{D}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1} \widehat{V}_T \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}$.

Proof: (a) Given (3), and (2),

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_T &= \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\left(\mathbf{X}_t'\boldsymbol{\beta}^\dagger + \epsilon_t\right)\right) \\
&= \boldsymbol{\beta}^\dagger + \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t
\end{aligned}$$

Now,

$$\begin{aligned}
& \widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger \\
= \quad & \mathbf{M}_T^{-1} \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \epsilon_t \\
& + \left(\left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' \right)^{-1} - \mathbf{M}_T^{-1} \right) \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \epsilon_t \qquad (4)
\end{aligned}$$

Given A-OLS-1(iii), $\mathbf{M}_T = O(1)$ and $\mathbf{M}_T$ uniformly positive definite (i.e. $\inf_T \det \mathbf{M}_T > 0$), it follows that $\mathbf{M}_T^{-1} = O(1)$. Now, by A-OLS-1(i)-(ii), $\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \epsilon_t = o_{a.s.}(1)$. Thus,

$$\mathbf{M}_T^{-1} \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \epsilon_t = O(1) o_{a.s.}(1) = o_{a.s.}(1)$$

by the product rule (recall that a term which is $O(1)$ is also $O_{a.s.}(1)$). As for the second term on the RHS of (4), as the inversion of a uniformly positive definite matrix is a continuous operation, given A-OLS-1(iii) and given Property AS1,

$$\left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' \right)^{-1} - \mathbf{M}_T^{-1} = o_{a.s.}(1).$$

Thus, recalling A-OLS-1(ii), the second term of the right hand side of (4) is $o_{a.s.}(1)$. This concludes the proof of part (a).

(b) Given (4), we have that

$$\begin{aligned}
& T^{1/2} \mathbf{D}_T^{-1/2} \left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger \right) \\
= \quad & \mathbf{D}_T^{-1/2} \mathbf{M}_T^{-1} \mathbf{V}_T^{1/2} \mathbf{V}_T^{-1/2} \frac{1}{T^{1/2}} \sum_{t=1}^{T} \mathbf{X}_t \epsilon_t \\
& + \mathbf{D}_T^{-1/2} \left(\left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' \right)^{-1} - \mathbf{M}_T^{-1} \right) \mathbf{V}_T^{1/2} \mathbf{V}_T^{-1/2} \frac{1}{T^{1/2}} \sum_{t=1}^{T} \mathbf{X}_t \epsilon_t \qquad (5)
\end{aligned}$$

We first show that the second term on the RHS of (5) is $o_P(1)$.

Given, A-OLS-1(iv), $\mathbf{V}_T^{-1/2} \frac{1}{T^{1/2}} \sum_{t=1}^{T} \mathbf{X}_t \epsilon_t = O_P(1)$, as it converges in distribution. Also, given the fact that $\mathbf{M}_T$ and $\mathbf{V}_T$ are $O(1)$ and uniformly positive definite, $\mathbf{D}_T^{-1/2}$ and $\mathbf{V}_T^{1/2}$ are $O(1)$.

In part (a), we have shown that $\left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' \right)^{-1} - \mathbf{M}_T^{-1} = o_{a.s.}(1)$. Thus the second term on the RHS of (5) is $o_P(1)$ by the product rule (recall that a $o_{a.s.}(1)$ term is also $o_P(1)$).

Now, consider the first term on the RHS of (5).

First, note that, as $\mathbf{D}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1} V_T \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}$,

$$\mathbf{D}_T^{1/2} = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}\mathbf{V}_T^{1/2}$$

and so

$$\mathbf{D}_T^{-1/2} = \mathbf{V}_T^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right).$$

So, given A-OLS-1(iii), and recalling the product rule,

$$\begin{aligned}
& \mathbf{D}_T^{-1/2}\mathbf{M}_T^{-1}\mathbf{V}_T^{1/2} \\
= \;& \mathbf{V}_T^{-1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)\mathbf{M}_T^{-1}\mathbf{V}_T^{1/2} \\
= \;& \mathbf{V}_T^{-1/2}\left(\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)-\mathbf{M}_T\right)\mathbf{M}_T^{-1}\mathbf{V}_T^{1/2} \\
& +\mathbf{V}_T^{-1/2}\mathbf{M}_T\mathbf{M}_T^{-1}\mathbf{V}_T^{1/2} \\
= \;& o_{a.s.}(1) + I_k
\end{aligned}$$

Thus, the

$$\begin{aligned}
& \mathbf{D}_T^{-1/2}\mathbf{M}_T^{-1}\mathbf{V}_T^{1/2}\mathbf{V}_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t \\
= \;& \mathbf{V}_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t + o_p(1)
\end{aligned}$$

Given A-OLS-1(iv), $\mathbf{V}_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t \xrightarrow{d} N(0, I_k)$, and by the asymptotic equivalence lemma, the LHS (which is the first term on the RHS of (5)) converges in distribution to a $N(0, I_k)$. Recalling that the second term on the RHS of (5) is $o_p(1)$, part (b) follows by applying again the asymptotic equivalence lemma.

(c)

$$\begin{aligned}
& T^{1/2}\widehat{\mathbf{D}}_T^{-1/2}\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right) \\
= \;& T^{1/2}\mathbf{D}_T^{-1/2}\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right) \\
& + \left(\widehat{\mathbf{D}}_T^{-1/2} - \mathbf{D}_T^{-1/2}\right)T^{1/2}\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right) \qquad (6)
\end{aligned}$$

By part (ii), we know that the first term on the RHS of (6) converges in distribution to a $N(0, I_k)$. Now, given that $\widehat{\mathbf{V}}_T - \mathbf{V}_T \xrightarrow{pr} 0$, it follows that $\widehat{\mathbf{D}}_T^{-1/2} - \mathbf{D}_T^{-1/2} = o_P(1)$; $T^{1/2}\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right)$ is $O_P(1)$ as it converges in distribution. Thus, the second term on the RHS of (6) is $o_p(1)$ by the product rule.

Part (c) then follows by the asymptotic equivalence lemma.

Remark 1: If in Assumption OLS-1(ii)-(iii), we replace almost sure convergence with convergence in probability, the statement in part (a) would be $\widehat{\boldsymbol{\beta}}_T \xrightarrow{p} \boldsymbol{\beta}^\dagger$, while the statements in part (b) and (c) would not change.
Remark 2: The proof of Theorem OLS-1 is based on four elements: law of large numbers, central limit, product rule and asymptotic equivalence. So far, we have assumed that law of large numbers and central limit hold (Assumptions OLS-1(ii),(iii),(iv)). Though, these are NOT primitive assumptions. In the sequel, we shall provide different set of assumptions on the degree of memory and dependence of the data, ensuring that law of large numbers and central limit hold.
Remark 3: In statement (c), we have assumed that there exists a consistent estimator of the (asymptotic) variance of $\frac{1}{T^{1/2}} \sum_{t=1}^{T} \mathbf{X}_t \epsilon_t$. In the sequel, we shall provide sufficient conditions for that. We anticipate that, in the case of dynamically misspecified models, this is somewhat complex, i.e. cannot rely just on law of large numbers.

Before addressing the issues in Remarks 2 and 3, we first outline the asymptotic behavior of the three classical tests, Wald, Lagrange Multiplier and Likelihood Ratio, still assuming that Assumptions OLS-1(ii)-(iv) hold.

Hypothesis Testing: Wald, Lagrange Multiplier and Likelihood Ratio Tests

Often we are interested in testing linear restriction about parameters, that is we want to test the null hypothesis

$$H_0 : \mathbf{R}\boldsymbol{\beta}^\dagger = \mathbf{r} \tag{7}$$

versus the alternative,

$$H_A : \mathbf{R}\boldsymbol{\beta}^\dagger \neq \mathbf{r} \tag{8}$$

where $\mathbf{R}$ is a $q \times k$ selection matrix and $\mathbf{r}$ is a $q \times 1$ vector, q denotes the number of linear restrictions.

Examples:Suppose we want to test $H_0 : \beta_2^\dagger = 0$ vs $H_1 : \beta_2^\dagger \neq 0$.

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{r} = 0$$

1. Suppose we want to test $H_0 : \beta_2^\dagger = \beta_3^\dagger = ... = \beta_k^\dagger = 0$ vs $H_1 : \beta_i^\dagger \neq 0$ for at least one $i = 2, ..., k$. In this case we test

$$\mathbf{R}\boldsymbol{\beta}^\dagger = \mathbf{r} \text{ vs } \mathbf{R}\boldsymbol{\beta}^\dagger \neq \mathbf{r}$$

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & & 1 & & \\ \vdots & & & \ddots & \\ 0 & 0 & & & 1 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

   where $\mathbf{R}$ is $(k-1) \times k$ and $\mathbf{r}$ is $(k-1) \times 1$**.**

2. Suppose we want to test $H_0 : \beta_2^\dagger + \beta_3^\dagger = 0$ vs $H_1 : \beta_2^\dagger + \beta_3^\dagger \neq 0$. In this case we test

$$\mathbf{R}\boldsymbol{\beta}^\dagger = \mathbf{r} \text{ vs } \mathbf{R}\boldsymbol{\beta}^\dagger \neq \mathbf{r}$$

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 1 & 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{r} = 0$$

3. Suppose we want to test $H_0 : \beta_3^\dagger - \beta_4^\dagger = 0$ and $\beta_2^\dagger = 1$ vs $H_1 : \beta_3^\dagger - \beta_4^\dagger \neq 0$ and/or $\beta_2^\dagger \neq 1$. In this case we test

$$\mathbf{R}\boldsymbol{\beta}^\dagger = \mathbf{r} \text{ vs } \mathbf{R}\boldsymbol{\beta}^\dagger \neq \mathbf{r}$$

   $\mathbf{R}$ has two rows, one for $\beta_2^\dagger = 1$, and one for $\beta_3^\dagger - \beta_4^\dagger = 0$:

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & \dots & 0 \end{bmatrix}, \qquad \mathbf{r} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Wald Test

When performing Wald test, we estimate only the unrestricted model. Hereafter, let $\mathbf{V}_T = var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right)$, and let $\widehat{\mathbf{V}}_T$ be a consistent estimator of $\mathbf{V}_T$ (under H_0), and let $\widehat{\boldsymbol{\beta}}_T$ be the OLS estimator of the unrestricted model. Define the Wald statistic as

$$\mathcal{W}_T = T\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{r}\right)'\left(\mathbf{R}\widehat{\mathbf{D}}_T\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{r}\right)$$

where

$$\widehat{\mathbf{D}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}\widehat{V}_T\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1} \tag{9}$$

.

Theorem Wald-1: Let the assumptions of Theorem OLS-1(c) hold. Then:
(i) under H_0, $\mathcal{W}_T \xrightarrow{d} \chi^2_q$
(ii) under H_A, $\mathcal{W}_T$ diverges to infinity.
Proof: (i) Under H_0, $\mathbf{r} = \mathbf{R}\boldsymbol{\beta}^\dagger$, thus

$$\mathcal{W}_T = T\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{R}\boldsymbol{\beta}^\dagger\right)'\left(\mathbf{R}\widehat{\mathbf{D}}_T\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{R}\boldsymbol{\beta}^\dagger\right)$$

By Theorem OLS-1(c), $T^{1/2}\left(\mathbf{R}\widehat{\mathbf{D}}_T\mathbf{R}'\right)^{-1/2}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{R}\boldsymbol{\beta}^\dagger\right) \xrightarrow{d} N(0, I_q)$ and so $\mathcal{W}_T \xrightarrow{d} \chi^2_q$ because of the continuous mapping theorem.

(ii) Under H_A, $\mathbf{r} \neq \mathbf{R}\boldsymbol{\beta}^\dagger$, thus

$$\begin{aligned}
\mathcal{W}_T &= T\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{R}\boldsymbol{\beta}^\dagger\right)'\left(\mathbf{R}\widehat{\mathbf{D}}_T\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{R}\boldsymbol{\beta}^\dagger\right) \\
&\quad + T\left(\mathbf{R}\boldsymbol{\beta}^\dagger - \mathbf{r}\right)'\left(\mathbf{R}\widehat{\mathbf{D}}_T\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{R}\boldsymbol{\beta}^\dagger\right) \\
&\quad + T\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{R}\boldsymbol{\beta}^\dagger\right)'\left(\mathbf{R}\widehat{\mathbf{D}}_T\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{r}\right) \\
&\quad + T\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{r}\right)'\left(\mathbf{R}\widehat{\mathbf{D}}_T\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{r}\right)
\end{aligned} \tag{10}$$

The first term on the RHS of (10) is $O_P(1)$ as it converges in distribution under both hypotheses. The second and third term are of order $O_P(T^{1/2})$ as $\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{R}\boldsymbol{\beta}^\dagger\right) = O_P(T^{-1/2})$ and $\mathbf{R}\boldsymbol{\beta}^\dagger - \mathbf{r} \neq \mathbf{0}$. Finally, the last term diverges to (plus) infinity at rate T, as $\mathbf{R}\boldsymbol{\beta}^\dagger - \mathbf{r} \neq \mathbf{0}$.

Remarks on Wald Test

(i) The Wald test is not an invariant test, in the sense that is not invariant to parameters reparametrization (one to one transformation of the parameter space), see Dagenais and Dufour (Econometrica 1991). Broadly speaking, if instead of test $H_0 : \beta = 0$, we test $H_0' : \ln(\beta) = 1$, we may get quite different answers.

(ii) If $var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right) = T^{-1}\sum_{t=1}^{T}E\left(\mathbf{X}_t\mathbf{X}_t'\right)\sigma_\epsilon^2$, i.e. in the case of conditional homoskedasticity, the Wald statistics writes as:

$$\mathcal{W}_T = T\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{r}\right)'\left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}/n\right)^{-1}\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T - \mathbf{r}\right)/\widehat{\sigma}_\epsilon^2$$

and is numerically identical to qF, where$F$ is the standard F-statistic for the same null.

Lagrange Multiplier Test

An alternative approach to the Wald approach above is to impose the restrictions, i.e. estimate only the restricted model. Let $\widetilde{\boldsymbol{\beta}}_T$ be defined as:

$$\widetilde{\boldsymbol{\beta}}_T = \arg\min_{\beta}\frac{1}{T}\sum_{t=1}^{T}\left(y_t - \mathbf{X}_t'\boldsymbol{\beta}\right)^2 \text{ subject to } \mathbf{R}\boldsymbol{\beta} = \mathbf{r}, \tag{11}$$

i.e.Now let $\widetilde{\epsilon}_t$ be the restricted residuals, i.e. $\widetilde{\epsilon}_t = y_t - \mathbf{X}_t'\widetilde{\boldsymbol{\beta}}_T$. Note that is the restricted least square estimators, which is DIFFERENT from the unrestricted OLS estimator of the restricted model. In fact, $\widetilde{\boldsymbol{\beta}}_T$ is $k \times 1$ even if there are zero restrictions.

The Lagrange Multiplier statistic is defined as:

$$\begin{aligned}
& \mathcal{LM}_T \\
= \; & T\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\widetilde{\epsilon}_t\right)'\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}\mathbf{R}'\left(\mathbf{R}\widetilde{\mathbf{D}}_T\mathbf{R}'\right)^{-1} \\
& \times\mathbf{R}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\widetilde{\epsilon}_t\right),
\end{aligned} \tag{12}$$

where

$$\widetilde{\mathbf{D}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}\widetilde{\mathbf{V}}_T\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1} \tag{13}$$

and $\widetilde{\mathbf{V}}_T$ is an estimator of $\mathbf{V}_T$ based on the restricted residuals, recall that $\mathbf{V}_T = var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right)$.

Theorem LM-1: Let the assumptions of Theorem OLS-1(b) hold. Then:
(i) Assume that $\widetilde{\mathbf{V}}_T - \mathbf{V}_T = o_p(1)$, then under H_0, $\mathcal{LM}_T \xrightarrow{d} \chi_q^2$
(ii) under H_A, $\mathcal{LM}_T$ diverges to infinity.

Proof: (i) Consider

$$\mathcal{LM}_T^{1/2} = T^{1/2}\left(\mathbf{R}\widetilde{\mathbf{D}}_T\mathbf{R}'\right)^{-1/2}\mathbf{R}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\widetilde{\epsilon}_t\right)$$

and, recalling that $\widetilde{\epsilon}_t - \epsilon_t = -\mathbf{X}_t' \left(\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger \right)$, note that

$$\begin{aligned} \mathcal{LM}_T^{1/2} &= T^{1/2} \left(\mathbf{R} \widetilde{\mathbf{D}}_T \mathbf{R}' \right)^{-1/2} \mathbf{R} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{X}_t \mathbf{X}_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{X}_t \epsilon_t \right) \\ &\quad - T^{1/2} \left(\mathbf{R} \widetilde{\mathbf{D}}_T \mathbf{R}' \right)^{-1/2} \mathbf{R} \left(\widetilde{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger \right). \end{aligned} \tag{14}$$

Now, under H_0, $\mathbf{R}\boldsymbol{\beta}^\dagger = \mathbf{r}$, and by construction $\mathbf{R}\widetilde{\boldsymbol{\beta}}_\mathbf{T} = \mathbf{r}$. Thus, under the null, the second term on the RHS of (14) is equal to zero. Now, from the proof of Theorem OLS-1(b), we have seen that

$$T^{1/2} \left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger \right) = \left(\frac{1}{T} \sum_{t=1}^T \mathbf{X}_t \mathbf{X}_t' \right)^{-1} \frac{1}{T^{1/2}} \sum_{t=1}^T \mathbf{X}_t \epsilon_t.$$

Thus, the first term on the RHS of (14) writes as:

$$T^{1/2} \left(\mathbf{R} \widetilde{\mathbf{D}}_T \mathbf{R}' \right)^{-1/2} \mathbf{R} \left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger \right).$$

As we have assumed that $\widetilde{\mathbf{V}}_T - \mathbf{V}_T = o_p(1)$, $T^{1/2} \left(\mathbf{R} \widetilde{\mathbf{D}}_T \mathbf{R}' \right)^{-1/2} \mathbf{R} \left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger \right) \xrightarrow{d} N(0,1)$ by the same argument used in the proof of Theorem OLS-1(c). Part (i) then follows from the continuous mapping theorem.

(ii) under H_A, $\mathbf{R}\boldsymbol{\beta}^\dagger \neq \mathbf{r}$, and by construction $\mathbf{R}\widetilde{\boldsymbol{\beta}}_\mathbf{T} = \mathbf{r}$, thus by adding and subtracting $\mathbf{r}$ in the second term on the RHS of (14), we have that

$$\begin{aligned} \mathcal{LM}_T^{1/2} &= T^{1/2} \left(\mathbf{R} \widetilde{\mathbf{D}}_T \mathbf{R}' \right)^{-1/2} \mathbf{R} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{X}_t \mathbf{X}_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{X}_t \epsilon_t \right) \\ &\quad + T^{1/2} \left(\mathbf{R} \widetilde{\mathbf{D}}_T \mathbf{R}' \right)^{-1/2} \left(\mathbf{R}\boldsymbol{\beta}^\dagger - \mathbf{r} \right) \end{aligned} \tag{15}$$

The first term on the RHS of (15) is $O_P(1)$ as it converges in distribution by the same argument as in part (i), but the second term is of $O_P(1)O(T^{1/2})$ and thus when premultiplied by its transpose it diverges to infinity at rate T.

Remarks on Lagrange Multiplier Test

(i) Also, the Lagrange Multiplier test is (in general) not an invariant test.

(ii) If $var \left(\frac{1}{T^{1/2}} \sum_{t=1}^T \mathbf{X}_t \epsilon_t \right) = T^{-1} \sum_{t=1}^T E\left(\mathbf{X}_t \mathbf{X}_t' \right) \sigma_\epsilon^2$, i.e. in the case of conditional homoskedasticity, then the LM test can be performed in a TR^2 format, at least for the case in which we are testing zero restrictions. That is, we estimate the restricted model, and then regress the residuals on a constant and on the omitted variables, and compute the R^2 from the latter regression. Under the null, $TR^2 \xrightarrow{d} \chi_q^2$, under the alternative, it diverges to infinity at rate T.

Likelihood Ratio Test

If $\epsilon_t|\mathbf{X}_t$ is iid $N(0,\sigma_\epsilon^2)$, then the OLS $\widehat{\boldsymbol{\beta}}_T$ is also the Maximum Likelihood estimator (MLE); on the other hand if $\epsilon_t|\mathbf{X}_t$ is NOT iid $N(0,\sigma_\epsilon^2)$ then $\widehat{\boldsymbol{\beta}}_T$ is a Quasi Maximum Likelihood estimator (QMLE), i.e. an estimator constructed under the wrong assumption that the marginal of the error is normal. Let $\widetilde{\boldsymbol{\beta}}_T$ be defined as in (11), then define the likelihood ratio statistics as $\mathcal{LR}_T$,

$$\mathcal{LR}_T = \frac{T}{2}\ln\left(\widehat{\sigma}_\epsilon^2/\widetilde{\sigma}_\epsilon^2\right), \tag{16}$$

where $\widehat{\sigma}_\epsilon^2 = \frac{1}{T}\sum_{t=1}^T\left(y_t - \mathbf{X}_t'\widehat{\boldsymbol{\beta}}_T\right)^2$ and $\widetilde{\sigma}_\epsilon^2 = \frac{1}{T}\sum_{t=1}^T\left(y_t - \mathbf{X}_t'\widetilde{\boldsymbol{\beta}}_T\right)^2$.

Theorem LRT-1: Let the assumptions of Theorem OLS-1(b) hold. Also, assume conditional homoskedasticity, i.e. $var\left(\frac{1}{T^{1/2}}\sum_{t=1}^T\mathbf{X}_t\epsilon_t\right) = T^{-1}\sum_{t=1}^T E\left(\mathbf{X}_t\mathbf{X}_t'\right)\sigma_\epsilon^2$. Then:
(i) under H_0, $-2\mathcal{LR}_T \xrightarrow{d} \chi_q^2$
(ii) under H_A, $-2\mathcal{LR}_T$ diverges to infinity.

The proof of Theorem LRT-1 requires the following result:
Intermediate value theorem. Let $s: R^k \to R$ be defined on an open set $\Theta \subset R^k$ and it's differentiable on Θ, with gradient ∇s. Then, for any θ and θ_0 in Θ, $s(\theta) = s(\theta_0) + \nabla s(\overline{\theta})'\left(\theta - \theta_0\right)$, with $\overline{\theta} \in (\theta, \theta_0)$.

Proof: (i) First, recalling that $\mathbf{X}'\widehat{\epsilon} = 0$ by construction, note that

$$\widetilde{\sigma}_\epsilon^2 = \widehat{\sigma}_\epsilon^2 + \left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right)'\left(\mathbf{X}'\mathbf{X}/\mathbf{T}\right)\left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right),$$

and so

$$\mathcal{LR}_T = -\frac{T}{2}\ln\left(1 + \left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right)'\left(\mathbf{X}'\mathbf{X}/T\right)\left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right)/\widehat{\sigma}_\epsilon^2\right).$$

We now make use of the intermediate value theorem, setting $\theta_0 = 0$ and

$$\theta = \left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right)'\left(\mathbf{X}'\mathbf{X}/T\right)\left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right)/\widehat{\sigma}_\epsilon^2$$

Then,

$$\mathcal{LR}_T = -\frac{T}{2}\left(1+\overline{\theta}\right)^{-1}\left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right)'\left(\mathbf{X}'\mathbf{X}/T\right)\left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right)/\widehat{\sigma}_\epsilon^2,$$

where $\overline{\theta} \in (\theta, 0)$. As $\overline{\theta} \to 1$, $\left(1+\overline{\theta}\right)^{-1}$, and so

$$-2\mathcal{LR}_T - T\left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right)'\left(\mathbf{X}'\mathbf{X}/T\right)\left(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_T\right)/\widehat{\sigma}_\epsilon^2 \xrightarrow{p} 0.$$

By the asymptotic equivalence Lemma, it suffice to consider the limiting distribution of

$$T\left(\widehat{\boldsymbol{\beta}}_T-\widetilde{\boldsymbol{\beta}}_T\right)'\left(\mathbf{X}'\mathbf{X}/T\right)\left(\widehat{\boldsymbol{\beta}}_T-\widetilde{\boldsymbol{\beta}}_T\right)/\widehat{\sigma}_\epsilon^2.$$

Now (i leave the proof for homework),

$$\left(\widehat{\boldsymbol{\beta}}_T-\widetilde{\boldsymbol{\beta}}_T\right)=\left(\mathbf{X}'\mathbf{X}/T\right)^{-1}\mathbf{R}'\left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}/T\right)^{-1}\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T-\mathbf{r}\right).$$

Thus, $-2\mathcal{LR}_T$ is asymptotically equivalent to $T\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T-\mathbf{r}\right)'\left(\mathbf{R}\left(\mathbf{X}'\mathbf{X}/n\right)^{-1}\mathbf{R}'\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_T-\mathbf{r}\right)/\widehat{\sigma}_\epsilon^2$, which is the Wald test for the case of conditional homoskedasticity.

(ii) Do it! easy.

Remarks on LR test.

(i) Advantage: It is an invariant test

(ii) Disadvantage. If conditional homoskedasticity does not hold, it does no longer have a chi-squared limiting distribution (though it has a distribution).

Other Remarks on Wald, LM and LR.

(iii) We have seen that under the null, and in the presence of conditional homoskedasticity, $\mathcal{W}_T$, $\mathcal{LM}_T$ and $-2\mathcal{LR}_T$ are asymptotic equivalent. Though, in finte sample, the following hold:

$$\mathcal{W}_T\geq-2\mathcal{LR}_T\geq\mathcal{LM}_T.$$

(iv) Suppose we are interested in testing nonlinear restrictions, i.e. we have to test $H_0:h(\boldsymbol{\beta}^\dagger)=0$ vs $H_A:h(\boldsymbol{\beta}^\dagger)\neq0$, where $h:\mathcal{R}^k\to\mathcal{R}^q$, $q\leq k$. Then

$$\mathcal{W}_T=Th(\widehat{\boldsymbol{\beta}}_T)'\left(\nabla_\beta h(\widehat{\boldsymbol{\beta}}_T)\widehat{\mathbf{D}}_T\nabla_\beta h(\widehat{\boldsymbol{\beta}}_T)'\right)^{-1}h(\widehat{\boldsymbol{\beta}}_T),$$

and the statement in theorem Wald-1 still applies. As for the Lagrange Multiplier and Likelihood Ratio, they are constructed as in (12) and (16), but with $\widetilde{\boldsymbol{\beta}}_T$ compute as $\widetilde{\boldsymbol{\beta}}_T=\arg\min_\beta\frac{1}{T}\sum_{t=1}^T\left(y_t-\mathbf{X}_t'\boldsymbol{\beta}\right)^2$ subject to $h(\boldsymbol{\beta})=\mathbf{0}$. The statement in Theorems LM-1 and LR-1 still apply.

We have derived the asymptotic normality of OLS estimators, as well as as the limiting distribution of Wald, Lagrage Multiplier and Likelihood Ratio test, assuming that:

(a) $\mathbf{X}'\mathbf{X}/T$ and $\mathbf{X}'\boldsymbol{\epsilon}$ satisfy a *Strong Law of Large Numbers*

(b) $\mathbf{X}'\boldsymbol{\epsilon}/T^{1/2}$ satisfy a *Central Limit Theorem*

(c) There exist a *consistent estimator for var* $\left(\mathbf{X}'\boldsymbol{\epsilon}/T^{1/2}\right)$.

In the sequel we shall provide various sets of primitive sufficient conditions on the data $(y_t, \mathbf{X}_t)$ ensuring that the law of large number is statisfied, central limit theorem applies and we can consistently estimate *var* $\left(\mathbf{X}'\boldsymbol{\epsilon}/T^{1/2}\right)$.

Strong Law of Large Numbers

We shall proceed from the easier case of independent and identically distributed data to the more complex in which we allow for both dependence and heterogeneity.

In the sequel we need the following inequality:

Holder Inequality: If $p > 1$, $p^{-1} + q^{-1} = 1$, and if $E(|Y|^p) < \infty$ and $E(|Z|^q) < \infty$,

$$E\left(|YZ|\right) \leq \left(E(|Y|^p)\right)^{1/p} \left(E(|Z|^q)\right)^{1/q}$$

For $p = q = 2$, we have the well known *Cauchy-Schwarz Inequality, i.e.*

$$E\left(|YZ|\right) \leq \left(E(|Y|^2)\right)^{1/2} \left(E(|Z|^2)\right)^{1/2}.$$

Proposition IID-1: If $\{Z_t\}$ is an iid sequence, then for any continuous functiuon g, $\{g(Z_t)\}$ is also iid (this is true more a more general class than continuous function, in fact it holds for all measurable function).

Identically and Independently Distributed Observations (Kolmogorov SLLN)

Let $\{Z_t\}$ be a sequence of *iid* observation. Then $T^{-1}\sum_{t=1}^{T} Z_t \overset{a.s.}{\to} \mu$ if and only if $E(|Z_t|) < \infty$ and $E(Z_t) = \mu$.

Proposition SLLN-1: Let $\{y_t, X_t\}$ be independently and identically distributed random sequence. If $E({X_{i,t}}^2) < \infty$ for $i = 1, ...k$ and if $E(\epsilon_t^2) < \infty$, then:

$$\text{(i) } \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t\epsilon_t \overset{a.s.}{\to} E(\mathbf{X}_1\epsilon_1)$$

$$\text{(ii) } \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t\mathbf{X}_t' \overset{a.s.}{\to} \mathbf{M}, \text{ where } \mathbf{M} = E(\mathbf{X}_1'\mathbf{X}_1)$$

If $E(\mathbf{X}_1\epsilon_1) = 0$, then A-OLS-1(ii) is satisfied. Also, if $\mathbf{M}$ is positive definite A-OLS-1(iii) is also satisfied.

Proof of (i): By Chauchy-Schwartz inequality, $E\left|\mathbf{X}_1\epsilon_1\right| \leq \left(E\left(\mathbf{X}_1\mathbf{X}_1'\right)\right)^{1/2}\left(E(\epsilon_i^2)\right)^{1/2} < \infty$. Also, by Proposition IID-1, $\mathbf{X}_t\epsilon_t$ is iid (the product is a continuous function); thus statement (i) follows from Kolmogorov Law of Large Numbers

(ii) By CS inequality, $E(|X_{i,t}X_{j,t}|) \leq \left(E({X_{i,t}}^2)\right)^{1/2}\left(E({X_{j,t}}^2)\right)^{1/2}$. Thus, for all $i,j = 1,...,\ \frac{1}{T}\sum_{t=1}^{T} X_{i,t}X_{j,t} \overset{a.s.}{\to} E(X_{i,1}X_{j,1})$. Almost sure convergence of each element ensures almost sure convergence of the matrix.

Independent and Heterogeneous Observations (SLNN)

Let $\{Z_t\}$ be a a sequence of independent observations, with $E(Z_t) = \mu_t$. If for some $\delta > 0$, $E(|Z_t|^{1+\delta}) < \Delta < \infty$ then $T^{-1}\sum_{t=1}^{T}(Z_t - \mu_t) \overset{a.s.}{\to} 0$.

Proposition SLLN-2: Let $\{y_t, X_t\}$ be independently and heterogeneously distributed random sequence. If $E({X_{i,t}}^{2(1+\delta)}) < \infty$ for $i = 1,...k$ and if $E(\epsilon_t^{2(1+\delta)}) < \infty$, then

$$\text{(i) } \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{X}_t\epsilon_t - E(\mathbf{X}_t\epsilon_t)\right) \overset{a.s.}{\to} 0$$

$$\text{(ii) } \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t' - \mathbf{M}_T \overset{a.s.}{\to} 0, \text{ where } \mathbf{M}_T = \frac{1}{T}\sum_{t=1}^{T}E(\mathbf{X}_t'\mathbf{X}_t)$$

If $E(\mathbf{X}_t\epsilon_t) = 0$, then A-OLS-1(ii) is satisfied. Also, if $\mathbf{M}_T$ is uniformly positive definite A-OLS-1(iii) is also satisfied.

Dependent and Homogeneously Distributed Observations

While the independent assumption often holds for cross section data, it cannot hold for time series data. The issue is to see how much dependence we can allow for and still have (S)LLN holding (we'll see that the restriction on dependence necessary for CLT are stronger than those necessary for (S)LNN).

We need some preliminary and then investigate some memory conditions often used in the literature.

$\sigma-$Algebra: A family (collection) $\mathcal{F}$ of subsets of Ω is a $\sigma-$field ($\sigma-$algebra), if: (a) the empty set ($\varnothing$) and Ω belong to $\mathcal{F}$, (ii) if $F \in \mathcal{F}$, then $F^c \in \mathcal{F}$ (F^c is the complement of F) (iii) If $\{F_i\}$ is a sequence of sets in $\mathcal{F}$, then $\cup_{i=1}^{\infty}F_i \in \mathcal{F}$.

The pair $(\Omega, \mathcal{F})$ is a measurable space whenever $\mathcal{F}$ is a $\sigma-$field. Any $F \in \mathcal{F}$, can be interpreted as an event.

Probability Measure. Let $(\Omega, \mathcal{F})$ be a measurable space. A mapping $P : \mathcal{F} \to [0,1]$ is a probability measure on $(\Omega, \mathcal{F})$, provided: (i) $P(\varnothing) = 0$ and $P(\Omega) = 1$, (ii)

for any $F \in \mathcal{F}$, $P(F^c) = 1 - P(F)$, (iii) For any disjoint sequence $F_i \in \mathcal{F}$, i.e. $F_i \cap F_j = 0$ for all $i \neq j$, $P\left(\cup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i)$.

The triplet $(\Omega, \mathcal{F}, P)$ is said to be a probaility space.

Measurable Function: A function $g : \Omega \to \mathcal{R}$ is said to be $\mathcal{F}-$measurable if for any real number a, the set $[\omega : g(\omega) \leq a] \in \mathcal{F}$.

Any continuous g is $\mathcal{F}-$measurable.

Measure Preserving Transformation. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A mapping $\Upsilon : \Omega \to \Omega$ is a measure preserving transformation if for any $F \in \mathcal{F}$, $P(\Upsilon^{-1}F) = P(F)$.

Stationarity: Given a random sequence $\{Z_t\}$, we say that $\{Z_t\}$ is stationary if for any $\tau = ... - 1, 0, 1, ...$ the joint distribution of $(...Z_{-1}, Z_0, Z_1, ...)$ is identical to the joint distribution of $(...Z_{-1+\tau}, Z_\tau, Z_{1+\tau}, ...)$.

Stationarity is sometime termed STRICT STATIONARITY to distinguish it from *covariance stationarity*. Recall that Z_t, is covariance stationary if (i) $E(Z_t) = \mu$ (ii) $Var(Z_t) = \sigma^2 < \infty$ (iii) $Cov(Z_t, Z_{t-k}) = \gamma_k$. A (strictly) stationary sequence with finite variance is covariance stationary.

Proposition SS-1: $\{Z_t\}$ is stationary if and and only if there exists a measure preserving transformation Υ such that, $Z_1(\omega) = Z_1(\omega)$, $Z_2(\omega) = Z_1(\Upsilon\omega)$, ..., $Z_T(\omega) = Z_1(\Upsilon^{T-1}\omega)$.

Note that iid implies stationarity. Though, if we relax the independence assumptions, the fact that $\{Z_t\}$ is identically distributed, does NOT imply strict stationarity. In fact, identically distributed means that the marginal distribution of Z_t is the same for all t, while strict stationarity means that the JOINT distribution of $\{Z_t\}$ is the same as the JOINT distribution of $\{Z_{t+\tau}\}$, for all τ.

Ergodicity: Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\{Z_t\}$ be a stationary sequence, and let Υ be a measure preserving transformation. Then, $\{Z_t\}$ is ergodic if:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P\left(F \cap \Upsilon^t G\right) = P(F)P\left(G\right),$$

for all $F, G \in \mathcal{F}$.

In the independent case, $P\left(F \cap G\right) = P(F)P\left(G\right)$. Can think at $\Upsilon^t G$ as at the event G shifted t periods ahead. Since $P(G) = P\left(\Upsilon^t G\right)$, ($\Upsilon$ is measure preserving) ergodicity means that F and $\Upsilon^t G$ are independent on average in the limit. Thus ergodicity is a form of asymptotic independence on average.

Property SS-2: If $\{Z_t\}$ is stationary ergodic, then for any $\mathcal{F}$-measurable function g, $g(...Z_{t-1}, Z_t, Z_{t+1}, ...)$ is also stationary ergodic. Note that g may be function of the infinite history of Z_t.

SLNN for Stationary Ergodic Sequence (Ergodic Theorem)

Let $\{Z_t\}$ be a stationary ergodic sequence, with $E|Z_t| < \infty$. Then $\frac{1}{T}\sum_{t=1}^{T} Z_t \overset{a.s.}{\to} \mu = E(Z_1)$.

Proposition SLLN-3: Let $\{y_t, \mathbf{X}_t\}$ be stationary ergodic random sequences. If $E({X_{i,t}}^2) < \infty$ for $i = 1, ...k$ and if $E(\epsilon_t^2) < \infty$, then:

$$\text{(i) } \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t\epsilon_t \overset{a.s.}{\to} E(\mathbf{X}_1\epsilon_1)$$

$$\text{(ii) } \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t\mathbf{X}_t' \overset{a.s.}{\to} \mathbf{M}, \text{ where } \mathbf{M} = E(\mathbf{X}_1'\mathbf{X}_1)$$

If $E(\mathbf{X}_1\epsilon_1) = 0$, then A-OLS-1(ii) is satisfied. Also, if $\mathbf{M}$ is positive definite A-OLS-1(iii) is also satisfied.
Proof: By the same argument as in the proof of Proposition SLLN1, recalling that if $\{y_t, \mathbf{X}_t\}$ are stationary ergodic, then also $\mathbf{X}_t\mathbf{X}_t'$ and $\mathbf{X}_t\epsilon_t$ are stationary ergodic.

Dependent and Heterogeneously Distributed Observations

We have seen that that ergodicity applies only to stationary sequences. Also, ergodicity is a too weak memory requirement in several circumstances. For heterogeneously distributed sequences, we need some stronger conditions, known as *mixing conditions.*
$\phi-$Mixing and $\alpha-$Mixing: Let $\mathcal{B}_{-\infty}^{n} = \sigma(..., Z_{n-1}, Z_n)$ and $\mathcal{B}_{n+m}^{\infty} = \sigma(Z_{n+m}, Z_{n+m+1}, ...)$ and define the ϕ and α mixing coefficient as:

$$\phi\left(\mathcal{B}_{-\infty}^{n}, \mathcal{B}_{n+m}^{\infty}\right) = \sup_{\{G\in\mathcal{B}_{-\infty}^{n}, F\in\mathcal{B}_{n+m}^{\infty}: P(G>0)\}} |P(H|G) - P(H)|$$

$$\alpha\left(\mathcal{B}_{-\infty}^{n}, \mathcal{B}_{n+m}^{\infty}\right) = \sup_{\{G\in\mathcal{B}_{-\infty}^{n}, F\in\mathcal{B}_{n+m}^{\infty}:\}} |P(H\cap G) - P(H)P(G)|$$

Note that $\phi\left(\mathcal{B}_{-\infty}^{n}, \mathcal{B}_{n+m}^{\infty}\right)$ and $\alpha\left(\mathcal{B}_{-\infty}^{n}, \mathcal{B}_{n+m}^{\infty}\right)$ measure the degree of dependence among two events which are $m-$periods apart. Note that, as $P(H\cap G) = P(H|G)P(G)$, $\phi\left(\mathcal{B}_{-\infty}^{n}, \mathcal{B}_{n+m}^{\infty}\right) \geq \alpha\left(\mathcal{B}_{-\infty}^{n}, \mathcal{B}_{n+m}^{\infty}\right)$. Now, let

$$\phi(m) = \sup_n \phi\left(\mathcal{B}_{-\infty}^{n}, \mathcal{B}_{n+m}^{\infty}\right) \text{ and } \alpha(m) = \sup_n \alpha\left(\mathcal{B}_{-\infty}^{n}, \mathcal{B}_{n+m}^{\infty}\right)$$

(i) If as $m \to \infty$ $\phi(m) \to 0$, $\{Z_t\}$ is $\phi-$mixing (or uniform mixing)
(ii) If as $m \to \infty$ $\alpha(m) \to 0$, $\{Z_t\}$ is $\alpha-$mixing (or strong mixing)
Note that $\phi-$mixing implies $\alpha-$mixing, as for all m $\phi(m) \geq \alpha(m)$.

Basically a mixing random sequence is a sequence which is asymptotically independent, i.e. two events which are m periods apart, become independent as $m \to \infty$.

For (S)LLN and CLT hold, we need that the mixing coefficient goes to zero fast enough.

Mixing Size

For $a \in \mathcal{R}$, if $\phi(m) = O(m^{-a-\varepsilon})$, for some $\varepsilon > 0$, then ϕ is of size $-a$; similarly if $\alpha(m) = O(m^{-a-\varepsilon})$, for some $\varepsilon > 0$, then α is of size $-a$.

The smaller is α the higher the dependence.

Property Mix1: If $\{Z_t\}$ is a $\phi-$mixing process of size $-a$, then for any measure function g, $g(Z_t, Z_{t+1}, ..., Z_{t+\tau})$, with $\tau < \infty$, is $\phi-$mixing of size $-a$; similarly If $\{Z_t\}$ is a $\alpha-$mixing process of size $-a$, then for any measure function g, $g(Z_t, Z_{t+1}, ..., Z_{t+\tau})$, with $\tau < \infty$, is $\alpha-$mixing of size $-a$.

Thus, any measurable function of a FINITE history of a mixing process, is mixing of the same size.

Covariance stationary ARMA processes are $\alpha-$mixing with mixing coefficient decaying at an exponential rate (therefore are said exponentially mixing), but they are not necessarily $\phi-$mixing.

SLNN Dependent Heterogeneous Sequences (McLeish)

Let $\{Z_t\}$ be a $\phi-$mixing sequence with size $-r/(2r-1)$, $r \geq 1$ or a $\alpha-$mixing sequence with size $-r/(r-1)$, $r > 1$, and with $E(|Z_t|^{r+\delta}) < \Delta < \infty$, for $\delta > 0$. Then, $\frac{1}{T}\sum_{t=1}^{T}(Z_t - E(Z_t)) \overset{a.s.}{\to} 0$.

Note trade-off between moments and memory conditions: the higher is r the more dependence we allow, but at the cost of stronger moment conditions (and less heterogeneity).

Proposition SLLN-4: Let $\{y_t, X_t\}$ be ϕ (α) random sequence of size $-r/(2r-1)$, $r \geq 1$ ($-r/(r-1)$, $r > 1$). If $E({X_{i,t}}^{2(1+\delta)}) < \infty$ for $i = 1, ...k$ and if $E(\epsilon_t^{2(1+\delta)}) < \infty$, then

$$\text{(i) } \frac{1}{T}\sum_{t=1}^{T}(\mathbf{X}_t\epsilon_t - E(\mathbf{X}_t\epsilon_t)) \overset{a.s.}{\to} 0$$

$$\text{(ii) } \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t' - \mathbf{M}_T \overset{a.s.}{\to} 0, \text{ where } \mathbf{M}_T = \frac{1}{T}\sum_{t=1}^{T}E(\mathbf{X}_t'\mathbf{X}_t)$$

If $E(\mathbf{X}_t\epsilon_t) = 0$, then A-OLS-1(ii) is satisfied. Also, if $\mathbf{M}_T$ is uniformly positive definite A-OLS-1(iii) is also satisfied.

Proof: By the same argument as in the proof of Proposition SLLN1, recalling that if $\{y_t, \mathbf{X}_t\}$ are be ϕ (α) random sequence of size $-a$, then also $\mathbf{X}_t\mathbf{X}_t'$ and $\mathbf{X}_t\epsilon_t$ are also ϕ (α) random sequence of size $-a$.

Central Limit Theorems

We now need to find moments and memory conditions under which $\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t$ satisfies a CLT, and so Assumption A-OLS-1(iv) is satisfied. As we see such conditions are slightly in stronger than those needed for the SLLN to hold. In particular, as for the law of large numbers, in the stationary case, ergodicity suffices, but this is no longer the case for the CLT.

In the sequel, we'll need the following result:

Cramer-Wold device

Let b_T a $k\times 1$ random vector, and let λ be a $k\times 1$ vector, such that $\lambda'\lambda=1$ and $\lambda' b_T \xrightarrow{d} \lambda' Z$. Then $b_T \xrightarrow{d} Z$.

CLT are stated for scalars, then the Cramer Wold device is used to obtain the CLT for vectors.

Identically and Independent Observations (Linderberg-Levy CLT)

Let $\{Z_t\}$ be a *iid* sequence, with $E(Z_t)=\mu$, $Var(Z_t)=\sigma^2$, with $0<\sigma^2<\infty$, then $\frac{1}{T^{1/2}}\sum_{t=1}^{T}\left(\frac{Z_t-\mu}{\sigma}\right)\xrightarrow{d}N(0,1)$.

Proposition CLT-1: Let $\{y_t, X_t\}$ be an independently and identically distributed random sequence. If $E({X_{i,t}}^4)<\infty$ for $i=1,...k$ and if $E(\epsilon_t^4)<\infty$, then if $E(\mathbf{X}_t\epsilon_t)=0$:

$$V^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\xrightarrow{d}N(0,I_k),$$

where $V=Var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right)$. Thus, A-OLS-1(iv) is satisfied.

Proof: By Cauchy-Schwarz inequality, for $i=1,...,k$ $E\left(X_{i,t}^2\epsilon_t^2\right)<\left(E\left(X_{i,t}^4\right)\right)^{1/2}\left(E(\epsilon_t^4)\right)<\infty$. By Proposition IID-1, $X_t\epsilon_t$ is iid. For λ $k\times 1$ vector, such that $\lambda'\lambda=1$, $Var\left(\lambda' V_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right)=1$, and Linderberg-Levy CLT, $\lambda' V_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\xrightarrow{d}\lambda' N(0,I_k)$. Thus, by the Cramer-Wold device, $V_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\xrightarrow{d}N(0,I_k)$.

Heterogeneous and Independent Observations (Liapunov CLT)

Let $\{Z_t\}$ be an independent sequence, with $E(Z_t)=\mu_t$, $Var(Z_t)=\sigma_t^2$, $E(|Z_t-\mu_t|^{2+\delta})<\Delta<\infty$, with $\delta>0$, and $\overline{\sigma}_T^2=\frac{1}{T}\sum_{t=1}^{T}\sigma_t^2>\delta'>0$. Then $\frac{1}{T^{1/2}}\sum_{t=1}^{T}\left(\frac{Z_t-\mu_t}{\overline{\sigma}_T}\right)\xrightarrow{d}N(0,1)$.

In the heterogenous case, we need to have a CLT for $V_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t$, where the summands $V_T^{-1/2}\mathbf{X}_t\epsilon_t$ may depend on T. We need the following generalization of the theorem above,

Identically and Heterogeneous Observations (Loeve CLT)

Let $\{Z_{tT}\}$ be an independent sequence, with $E(Z_{tT})=\mu_{tT}$, $Var(Z_{tT})=\sigma_{tT}^2$, $E(|Z_{tT}-\mu_{tT}|^{2+\delta})<\Delta<\infty$, with $\delta>0$, and $\overline{\sigma}_T^2=\frac{1}{T}\sum_{t=1}^{T}\sigma_{tT}^2>\delta'>0$. Then $\frac{1}{T^{1/2}}\sum_{t=1}^{T}\left(\frac{Z_{tT}-\mu_{tT}}{\overline{\sigma}_T^2}\right)\xrightarrow{d}N(0,1)$.

Proposition CLT-2: Let $\{y_t, X_t\}$ be an independently distributed random sequence. If $E(X_{i,t}^{2(2+\delta)}) < \Delta < \infty$, $E(\epsilon_{i,t}^{2(2+\delta)}) < \Delta < \infty$, for $i = 1, ...k$ and some $\delta > 0, V_T$ is uniformly positive definite. Then if $E(\mathbf{X}_t\epsilon_t) = 0$:

$$V_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t \xrightarrow{d} N(0, I_k),$$

where $V_T = Var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right)$. Thus, A-OLS-1(iv) is satisfied.

Proof: By Cauchy-Schwartz inequality, $E(|X_{i,t}\epsilon_{i,t}|^{2+\delta}) \leq \left(E(X_{i,t}^{2(2+\delta)})\right)^{1/2}\left(E(\epsilon_{i,t}^{2(2+\delta)})\right)^{1/2} < \Delta$. By construction, $V_T = \frac{1}{T}\sum_{t=1}^{T} var\left(\mathbf{X}_t\epsilon_t\right)$. Recalling that continuous function of independent rv is independent, the desired result follows by Loeve CLT and Cramer Wold device.

Dependent Observations.

Our objective is to provide primitive conditions on the observations, so that a CLT applies and A-OLS-1(iv) hold.
In the case of independent conditions, CLT hold by just strengthen the moment conditions required for the SLLN.
In the case of dependent observations, we need also to impose stronger conditions on the allowed degree of memory. In particular, stationary-ergodicity is not an enough strong memory requirement for ensuring a CLT to hold.

From Proposition SS1 and MIX1, we have see that $\mathbf{X}_t\epsilon_t$ cannot have more memory than $\{y_t, \mathbf{X}_t\}$. In fact, if $\{y_t, \mathbf{X}_t\}$ is stationary ergodic, then $\mathbf{X}_t\epsilon_t$ is also stationary ergodic, and if $\{y_t, \mathbf{X}_t\}$ is α-mixing of size $-a$, then $\mathbf{X}_t\epsilon_t$ is also α-mixing of size $-a$.

Nevertheless, $\mathbf{X}_t\epsilon_t$ can display much less memory than $\{y_t, \mathbf{X}\}$. This depends on whether the linear model is *dynamically correctly specified.*

Hereafter, let $\mathcal{F}_t = \sigma\left(y_1, ..., y_{t-1}, X_1, X_{2,}...X_t\right)$.[1]
(*Dynamic) Correct Specification*

The linear model $y_t = \mathbf{X}_t'\boldsymbol{\beta}^\dagger + \epsilon_t$ is dynamically correctly specified if $E(y_t|\mathcal{F}_t) = E(y_t|\mathbf{X}_t) = \mathbf{X}_t'\boldsymbol{\beta}^\dagger$.

Note that dynamic correct specification implies correct specification, but the reverse does not hold. For example, suppose that data are generated by a AR(2) process, i.e. $y_t = \alpha_{0,1}y_{t-1} + \alpha_{0,2}y_{t-2} + \epsilon_t$. However, i estimate a AR(1), say $y_t = \alpha_1^\dagger y_{t-1} + \epsilon_t$ ($\alpha_{0,1} \neq \alpha_1^\dagger$). Then, the AR1 model is correctly specified, in the sense that $E(y_t|y_{t-1}) = \alpha_1^\dagger y_{t-1}$, but dynamically misspecified, as $E(y_t|\mathcal{F}_t) = E(y_t|y_{t-1}, y_{t-2}) \neq E(y_t|y_{t-1})$.

[1] Recall that $\mathbf{X}_t$ is $k \times 1$, if one of its the component is y_{t-1}, then

$$\begin{aligned}\mathcal{F}_t &= \sigma\left(y_1, ..., y_{t-1}, X_1, X_{2,}...X_t\right)\\ &= \sigma\left(X_1, X_{2,}...X_t\right)\end{aligned}$$

Below we shall show that if the model is dynamically correctly specified, then $\mathbf{X}_t\epsilon_t$ is a martingale difference sequence.

Martingale Difference Sequence (mds)

Let Z_t be $\mathcal{F}_t-$measurable random sequence, with $E(Z_t) = 0$. Then $\{Z_t, \mathcal{F}_t\}$ is called a martingale difference sequence, if

$$E(Z_t|\mathcal{F}_t) = 0$$

Note that a continuous function of a martingale difference is NOT a martingale difference, i.e. if Z_t is a mds, then Z_t^2 is NOT mds.

Proposition DCS (dynamic correct specification)*:* If the linear model is dynamically correctly specified, then $\mathbf{X}_t\epsilon_t$ is a martingale difference sequence.
Proof:

$$\begin{aligned} E(\epsilon_t|\mathcal{F}_t) &= E(y_t|\mathcal{F}_t) - E(\mathbf{X}_t'\boldsymbol{\beta}^\dagger|\mathcal{F}_t) \\ &= E(y_t|\mathcal{F}_t) - \mathbf{X}_t'\boldsymbol{\beta}^\dagger = 0 \end{aligned}$$

Now, by the law of the iterated expectations,

$$E(\mathbf{X}_t\epsilon_t|\mathcal{F}_t) = E\left(\mathbf{X}_t E(\epsilon_t|\mathcal{F}_t)\right) = 0.$$

Below, we provide conditions under which A-OLS-1(iv) (CLT) hold, for dynamically correctly specified models, distinguishing between the stationary and the heterogeneous case.

Proposition CLT3 (CLT for stationary martingale difference sequences)

Let (y_t, X_t) be a stationary ergodic sequence, with $E(X_{i,t}^4) < \infty$, $E(\epsilon_{i,t}^4) < \infty$, and let $\{\mathbf{X}_t\epsilon_t, \mathcal{F}_t\}$ be a martingale difference sequence. Then, if $V = var\left(\frac{1}{T^{1/2}}\sum_{t=1}^T \mathbf{X}_t\epsilon_t\right)$ positive definite, then

$$V_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^T \mathbf{X}_t\epsilon_t \xrightarrow{d} N(0, I_k),$$

where $V_T = Var\left(\frac{1}{T^{1/2}}\sum_{t=1}^T \mathbf{X}_t\epsilon_t\right)$. Thus, A-OLS-1(iv) is satisfied.

Proof: By a similar argument as in the proof of Proposition CLT1, as the CLT theorem for stationary mds follows under the same conditions as the CLT for iid.

Proposition CLT4 (CLT for heterogeneous martingale difference sequences)

Let $\{\mathbf{X}_t\epsilon_t, \mathcal{F}_t\}$ be a martingale difference sequence, with $E(X_{i,t}^{2(2+\delta)}) < \Delta < \infty$, $E(\epsilon_t^{2(2+\delta)}) < \Delta < \infty$. Then, if $V_T = var\left(\frac{1}{T}\sum_{t=1}^T \mathbf{X}_t\epsilon_t\right)$ is uniformly positive definite, then

$$V_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^T \mathbf{X}_t\epsilon_t \xrightarrow{d} N(0, I_k),$$

Thus, A-OLS-1(iv) is satisfied.

Proof: By a similar argument as in the proof of Proposition CLT3, as the CLT theorem for heterogeneous mds follows under the same conditions as the CLT for heterogeneous independent sequence.

In the case of dynamic misspecification. $\mathbf{X}_t\epsilon_t$ is no longer a martingale difference sequence, and simply inherits the same degree of dependence of the observations $\{y_t, \mathbf{X}_t\}$.
For the case of homogeneous series, stationarity and ergodicity (regardless the moment conditions we are willing to impose) do not suffice for a CLT. Given that, we restrict out attention to mixing sequences. We provide a general CLT for mixing, heterogeneous observations.

CLT for heteregeneous mixing sequences (Wooldridge and White)

Let $\{Z_{tT}\}$ be $\phi-$mixing sequence of size $-r/2(r-1)$ $r>1$, or $\alpha-$mixing with size $-r/(r-2)$ $r>2$, with $E(Z_{t,T})=0$, $E\left(|Z_{tT}|^r\right)<\Delta<\infty$, $r>2$, $var(\frac{1}{T^{1/2}}\sum_{t=1}^{T}Z_t)=\overline{\sigma}_T^2>\delta>0$. Then $\frac{1}{T^{1/2}}\sum_{t=1}^{T}\left(\frac{Z_t}{\overline{\sigma}_T}\right)\xrightarrow{d}N(0,1)$.

Proposition CLT5 (CLT for mixing sequences)

Let $\{y_t, \mathbf{X}_t\}$ be a $\phi-$mixing sequence of size $-r/2(r-1)$ $r>1$, or $\alpha-$mixing with size $-r/(r-2)$ $r>2$, with $E(|X_{i,t}|^{2r})<\Delta<\infty$ and $E(|\epsilon_t|^{2r})<\Delta<\infty$, $r>2$, and let $V_T=var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right)$ be uniformly positive definite. Then:

$$V_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\xrightarrow{d}N(0,I_k),$$

Thus, A-OLS-1(iv) is satisfied.
Proof: First, by CS inequality, $E\left(|X_{i,t}\epsilon_t|^r\right)<\Delta<\infty$, also

$$var\left(\frac{1}{T^{1/2}}\lambda'V_T^{-1/2}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right)=\lambda'V_T^{-1/2}V_TV_T^{-1/2}\lambda=1,$$

For $\lambda\in\mathcal{R}^k$, $\lambda'\lambda=1$. Thus, recalling that by Proposition MIX1, $\mathbf{X}_t\epsilon_t$ are mixing of the same size as $(y_t,\mathbf{X}_t)$, by the Wooldrige-White CLT, $\lambda'V_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\xrightarrow{d}$ $\lambda'N(0,I_k)$, and the statement then follows by the Cramer-Wold device.

Consistent Estimation of Asymptotic Covariance Matrices

We have provided primitive conditions, in terms of memory and heterogeneity of the observations, under which $\mathbf{V}_T^{-1/2}\frac{1}{T^{1/2}}\sum_{t=1}^T \mathbf{X}_t\epsilon_t$ is asymptotically standard normal. In practice, we do not know V_T, and we need a consistent estimator $\widehat{V}_T$, i.e. such that $\widehat{V}_T - V_T = o_P(1)$. We have seen (Theorem OLS-1(d), Wald-1, LM-1, LRT-1) that test statistics based on V_T and $\widehat{V}_T$ are asymptotic equivalent.

Now, recalling that $E(\mathbf{X}_t\epsilon_t) = 0$,

$$Var\left(\frac{1}{T^{1/2}}\sum_{t=1}^T \mathbf{X}_t\epsilon_t\right) = \frac{1}{T}\sum_{t=1}^T E\left(\mathbf{X}_t\mathbf{X}_t'\epsilon_t^2\right)$$

$$+\frac{1}{T}\sum_{t=1}^T\sum_{s\neq t} E\left(\mathbf{X}_t\mathbf{X}_s'\epsilon_t\epsilon_s\right) + \frac{1}{T}\sum_{t=1}^T\sum_{s\neq t} E\left(\mathbf{X}_s\mathbf{X}_t'\epsilon_t\epsilon_s\right). \tag{17}$$

Now, in the case of independent observations, $\mathbf{X}_t\epsilon_t$ is an independent sequence, and so $E\left(\mathbf{X}_s\mathbf{X}_t'\epsilon_t\epsilon_s\right) = 0$ for all $t \neq s$.
Also, we have seen that in the case of dynamic correct specification, $\mathbf{X}_t\epsilon_t$ is a martingale sequence, and so for $s > t$, by the law of the iterated expectations,

$$E\left(\mathbf{X}_t\mathbf{X}_s'\epsilon_t\epsilon_s\right) = E\left(E\left(\mathbf{X}_t\mathbf{X}_s'\epsilon_t\epsilon_s\right)|\mathcal{F}_t\right) = E\left(\mathbf{X}_t\epsilon_t E\left(\mathbf{X}_s'\epsilon_s\right)|\mathcal{F}_t\right) = 0.$$

Therefore, in the case of either independent observations or dynamic correct specification, we can ignore all the covariance term. Therefore, we just need to provide a consistent estimator for $\frac{1}{T}\sum_{t=1}^T E\left(\mathbf{X}_t\mathbf{X}_t'\epsilon_t^2\right)$. Though, we need to distinguish two possible cases, conditional homoskedasticity and conditional heteroskedasticity.
Conditional Homoskedasticity: In this case, we know that $E(\epsilon_t^2|\mathbf{X}_t) = \sigma_\epsilon^2$.[2] Now,

$$E\left(\mathbf{X}_t\mathbf{X}_t'\epsilon_t^2\right) = E\left(E\left(\mathbf{X}_t\mathbf{X}_t'\epsilon_t^2\right)|\mathbf{X}_t\right)$$

$$= E\left(\mathbf{X}_t\mathbf{X}_t'E\left(\epsilon_t^2\right)|\mathbf{X}_t\right) = E\left(\mathbf{X}_t\mathbf{X}_t'\right)\sigma_\epsilon^2$$

and so

$$\frac{1}{T}\sum_{t=1}^T E\left(\mathbf{X}_t\mathbf{X}_t'\epsilon_t^2\right) = \frac{1}{T}\sum_{t=1}^T E\left(\mathbf{X}_t\mathbf{X}_t'\right)\sigma_\epsilon^2$$

We state the theorem for the case of mixing observations, thus implicitely assuming that we have dynamic correct specification, otherwise we could not ignore the cross terms.

Proposition Var1: Let $\{y_t, \mathbf{X}_t\}$ be a $\phi-$mixing sequence of size $-r/2(r-1)$ $r > 1$, or $\alpha-$mixing with size $-r/(r-2)$ $r > 2$, with $E(|X_{i,t}|^{2r}) < \Delta < \infty$ and

[2] Note that we are not ruling out the possibility that ϵ_t be unconditionally heteroskedastic.

$E(|\epsilon_t|^{2r}) < \Delta < \infty$, $r > 2$. Also, $E(\mathbf{X}_t\epsilon_t) = 0$ and $E\left(\mathbf{X}_t\mathbf{X}_t'\epsilon_t^2\right) = E\left(\mathbf{X}_t\mathbf{X}_t'\right)\sigma_\epsilon^2$. Then,

$$\widehat{V}_T - V_T = o_P(1)$$

where,

$$\widehat{V}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\right)\widehat{\sigma}_{\epsilon T}^2$$

$$V_T = \left(\frac{1}{T}\sum_{t=1}^{T}E(\mathbf{X}_t\mathbf{X}_t')\right)\sigma_\epsilon^2$$

and[3] $\widehat{\sigma}_{\epsilon T}^2 = \frac{1}{T}\sum_{t=1}^{T}\widehat{\epsilon}_t^2$.

Proof: Let $\widehat{M}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'$ and $M_T = \frac{1}{T}\sum_{t=1}^{T}E(\mathbf{X}_t\mathbf{X}_t')$.

$$\begin{aligned}\widehat{V}_T - V_T &= \left(\widehat{M}_T - M_T\right)\sigma_\epsilon^2 + M_T\left(\widehat{\sigma}_{\epsilon T}^2 - \sigma_\epsilon^2\right) \\ &\quad + \left(\widehat{M}_T - M_T\right)\left(\widehat{\sigma}_{\epsilon T}^2 - \sigma_\epsilon^2\right) \qquad (18)\end{aligned}$$

It suffices to show that the first two terms on the RHS of (18) are $o_P(1)$. The third term will then be $o_P(1)$ as a product of $o_P(1)$ is $o_P(1)$. The first term is $o_{a.s.}(1)$, by the SLLN, given the moments and memory conditions assumed. As for the second term,

$$\begin{aligned}\widehat{\sigma}_{\epsilon T}^2 &= \frac{1}{T}\sum_{t=1}^{T}\widehat{\epsilon}_t^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_t - \mathbf{X}_t'\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right)\right)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2 + \left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right)'\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right) \\ &\quad -\frac{2}{T}\sum_{t=1}^{T}\epsilon_t\mathbf{X}_t'\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right) \qquad (19)\end{aligned}$$

Now, $\frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2 - \sigma_\epsilon^2 = o_{a.s.}(1)$ by the SLLN, while the last two terms on the last equality (19) are $o_p(1)$, as $\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right) = O_P(T^{-1/2})$, because of the CLT.

Conditional Heteroskedasticity

We still assume that $E\left(\mathbf{X}_s\mathbf{X}_t'\epsilon_t\epsilon_s\right) = 0$ for all $t \neq s$, but we relax the conditional homoskedasticity assumption.

Proposition Var2: Let $\{y_t, \mathbf{X}_t\}$ be a $\phi-$mixing sequence of size $-r/2(r-1)$ $r > 1$, or $\alpha-$mixing with size $-r/(r-2)$ $r > 2$, with $E(|X_{i,t}|^{2(r+\delta)}) < \Delta < \infty$

[3] Note that $\widehat{V}_T$ is the default estimator used by computer packages.

and $E(|\epsilon_t|^{2(r+\delta)}) < \Delta < \infty, r > 2$. Also, $E(\mathbf{X}_t\epsilon_t) = 0$ and $Var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right) = \frac{1}{T}\sum_{t=1}^{T}E\left(\mathbf{X}_t\mathbf{X}_t'\epsilon_t^2\right)$. Then,

$$\widehat{V}_T - V_T = o_P(1) \tag{20}$$

where,

$$\widehat{V}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\widehat{\epsilon}_t^2$$

$$V_T = \frac{1}{T}\sum_{t=1}^{T}E\left(\mathbf{X}_t\mathbf{X}_t'\epsilon_t^2\right).$$

Note that the variance estimator defined in (20) is known as White Covariance estimator.

Proof: As ϵ_t is a scalar, $\widehat{\epsilon}_t' = \widehat{\epsilon}_t$.

$$\begin{aligned}
\widehat{V}_T &= \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\epsilon_t'\mathbf{X}_t' \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\left(\epsilon_t - \mathbf{X}_t'\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right)\right)\left(\epsilon_t' - \left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right)'\mathbf{X}_t\right)\mathbf{X}_t' \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\epsilon_t'\mathbf{X}_t' + \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right)\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right)'\mathbf{X}_t\mathbf{X}_t' \\
&\quad - \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right)\mathbf{X}_t'\epsilon_t' - \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\left(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^\dagger\right)'\mathbf{X}_t\mathbf{X}_t'. \qquad (21)
\end{aligned}$$

Given the moment conditions above, by the SLLN,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\epsilon_t'\mathbf{X}_t' - V_T = o_{a.s.}(1).$$

Thus, it remains to show that the last three terms on the RHS of (21) are $o_P(1)$.[4] We begin by considering the second term on the last equality on the RHS of

[4] Given a $n \times m$ matrix $A = [a_{ij}]$, $vec(A) = (a_{11}, a_{12}, ..., a_{nm})$.
For B $p \times q$, $A \otimes B$ is $(np \times mq)$ matrix

$$A \otimes B = \begin{pmatrix} a_{11}B & . & a_{1m}B \\ . & . & . \\ a_{1n}B & . & a_{nm}B \end{pmatrix}$$

Given ABC,

$$vec\,(ABC) = \left(C' \otimes A\right) vecB$$

(21).

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^{T} vec\left(\mathbf{X}_t\mathbf{X}_t'\left(\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}^{\dagger}\right)\mathbf{X}_t'\epsilon_t\right) \\ = & \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{X}_t\epsilon_t\otimes\mathbf{X}_t\mathbf{X}_t'\right) vec\left(\left(\widehat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}^{\dagger}\right)\right) \\ = & O_p(1)o_p(1)=o_p(1). \end{aligned}$$

The third and fourth terms on the last equality on the RHS of (21) can be treated in an analogous manner.

Estimation of Asymptotic Covariance Matrices in the Dynamically Misspecified Case

We now consider the case in which we have dependent observations and our model is dynamically misspecified. In this case, $\mathbf{X}_t\epsilon_t$ is no longer a martingale difference sequence, and thus we do no loger have that $E\left(\mathbf{X}_t\mathbf{X}_s'\epsilon_t\epsilon_s\right)=0$ for $t\neq s$. In this case, we need to take into account the cross terms too. Recall that, in the general case

$$Var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right)=\frac{1}{T}\sum_{t=1}^{T}E\left(\mathbf{X}_t\mathbf{X}_t'\epsilon_t^2\right)$$

$$+\frac{1}{T}\sum_{\tau=1}^{T}\sum_{t=\tau+1}^{T}\left(E\left(\mathbf{X}_t\mathbf{X}_{t-\tau}'\epsilon_t\epsilon_{t-\tau}\right)+E\left(\mathbf{X}_{t-\tau}\mathbf{X}_t'\epsilon_t\epsilon_{t-\tau}\right)\right).$$

Broadly speaking as we have a sum of T^2 covariance terms divided by T, in order to ensure that $Var\left(\frac{1}{T^{1/2}}\sum_{t=1}^{T}\mathbf{X}_t\epsilon_t\right)$ is finite, we need conditions under which $E\left(\mathbf{X}_t\mathbf{X}_{t-\tau}'\epsilon_t\epsilon_{t-\tau}\right)\to 0$ fast enough as $\tau\to\infty$. The speed at which the covariance terms are approaching zero, depends on the speed at which the mixing coefficient approach zero. Hereafter, with the notation $\|Z\|_p$ we mean $\left(E\left(|Z|^p\right)\right)^{1/p}$.

Lemma MIX (Covariance Inequality)

If for all $\tau\geq 0$, $E(Z_{t+\tau})=0$, $var(Z_t)<\infty$, and $E(|Z_{t+\tau}|^q)<\infty$, $q\geq 2$, for all $\tau\geq 1$, then

$$\left|E\left(Z_tZ_{t+\tau}\right)\right|\leq 2\phi(\tau)^{1-1/q}\left(var(Z_t)\right)^{1/2}\|Z_{t+\tau}\|_q$$

and

$$\left|E\left(Z_tZ_{t+\tau}\right)\right|\leq 2(2^{1/2}+1)\alpha(\tau)^{1/2-1/q}\left(var(Z_t)\right)^{1/2}\|Z_{t+\tau}\|_q$$

Thus, the faster the mixing coeffient are going to zero, the faster the covariance terms are going to zero.

Intuitively, if the covariance terms approach zero fast enough, then it will be enough to estimate say m_T covariance terms, where $m_T \to \infty$ as $T \to \infty$ but m_T goes to infinity slow enough.

Define:

$$\begin{aligned}\widehat{V}_T &= \frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t'\widehat{\epsilon}_t^2 \\ &\quad + \frac{1}{T}\sum_{\tau=1}^{m_T}\sum_{t=\tau+1}^{T} w_{\tau T}\left(\mathbf{X}_t\mathbf{X}_{t-\tau}'\widehat{\epsilon}_t\widehat{\epsilon}_{t-\tau} + \mathbf{X}_{t-\tau}\mathbf{X}_t'\widehat{\epsilon}_t\widehat{\epsilon}_{t-\tau}\right), \qquad (22)\end{aligned}$$

where as $T \to \infty$, $m_T \to \infty$, $m_T/T^{1/4} \to 0$, and $w_{\tau T} \to 1$. A commonly used weight is

$$w_{\tau T} = 1 - \frac{\tau}{m_T - 1}.$$

Note that the role of the weight $w_{\tau T}$ is to ensure that the estimator is positive definite (Newey-West 1987). $\widehat{V}_T$ is known as HAC (heteroskedasticity and autocorrelation rrobut) covariance estimator.

The following theorem (adapted from ATE Thm 6.21) provide sufficient conditions for the consistency of the HAC estimators.

Proposition Var3 (Consistency of HAC covariance estimators).
Let $\{y_t, \mathbf{X}_t\}$ be a $\phi-$mixing sequence of size $-r/2(r-1)$ $r > 1$, or $\alpha-$mixing with size $-r/(r-2)$ $r > 2$, with $E(|X_{i,t}|^{4(r+\delta)}) < \Delta < \infty$ and $E(|\epsilon_t|^{4(r+\delta)}) < \Delta < \infty$, $\delta > 0$, $r > 2$. If as $T \to \infty$, $m_T \to \infty$, $m_T/T^{1/4} \to 0$, and $w_{\tau T} \to 1$, then

$$\widehat{V}_T - V_T = o_P(1),$$

where $\widehat{V}_T$ is defined as in (22).

In practice one has to choose m_T, this is delicate...There are data driven way (e.g. Andrews Econometrica 1991). Typically, one tries a few values...