14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
September 11, 2007
revised September 9, 2013

Lecture 2

Limit Theorems, OLS, and HAC

Limit Theorems

What are limit theorems? They are laws describing behavior of sums of many random variables. The mostly used are the Law of Large Numbers and Central Limit Theorem. In fact, these are two sets of theorems, rather than just two theorems (different assumptions about moments conditions, dependence and the way of summing can lead to similar statements). The most generic form is:

If $\{x_i\}$ is a sequence of independent identically distributed (iid) random variables, with $Ex_i = \mu$, $\text{Var}(x_i) = \sigma^2$ then

1. Law of large numbers (LLN) – $\frac{1}{n}\sum_{i=1}^n x_i \to \mu$ (in $\mathcal{L}^2$, a.s., in probability)

2. Central limit theorems (CLT) – $\frac{\sqrt{n}}{\sigma}(\frac{1}{n}\sum_{i=1}^n x_i - \mu) \Rightarrow N(0,1)$

We stated these while assuming independence. In time series, we usually don't have independence. Let us explore where independence may have been used.

First, let's start at the simplest proof of LLN:

$$E\left(\frac{1}{n}\sum_{i=1}^n x_i - \mu\right)^2 = \text{Var}\left(\frac{1}{n}\sum_{i=1}^n x_i\right) \tag{1}$$

$$= \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^n x_i\right) \tag{2}$$

$$= \frac{1}{n^2}\sum_{i=1}^n \text{Var}(x_i) \tag{3}$$

$$= \frac{n\sigma^2}{n^2} \to 0 \tag{4}$$

We used independence to go from (2) to (3).

Without independence, we'd have

$$\begin{aligned}\text{Var}\left(\frac{1}{n}\sum_{i=1}^n x_i\right) =& \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \text{cov}(x_i, x_j)\\ =& \frac{1}{n^2}\left(n\gamma_0 + 2(n-1)\gamma_1 + 2(n-2)\gamma_2 + \ldots\right)\\ =& \frac{1}{n}\left[2\sum_{k=1}^n \gamma_k\left(1-\frac{k}{n}\right) + \gamma_0\right]\end{aligned}$$

If we assume absolute summability, *i.e.* $\sum_{j=-\infty}^{\infty}|\gamma_j| < \infty$, then

$$\lim_{n\to\infty}\frac{1}{n}\left[\sum_{k=1}^{n}\gamma_k\left(1-\frac{k}{n}\right)+\gamma_0\right]=0$$

Thus, we have:

Lemma 1. *If x_t is a weakly stationary time series(with mean μ) with absolutely summable auto-covariances then a law of large numbers holds (in probability and $\mathcal{L}^2$).*

Remark 2. Stationarity is not enough. Let $z \sim N(0,\sigma^2)$. Suppose $x_t = z\,\forall t$. Then $\text{cov}(x_t, x_s) = \sigma^2\,\forall t, s$, so we do not have absolute summability, and clearly we do not have a LLN for $\{x_t\}$ since the average $\frac{1}{n}\sum_{i=1}^{n} x_i$ equals to z, which is random.

Remark 3. For an MA, $x_t = c(L)e_t$, we have $\sum_{j=1}^{\infty}|c_j| < \infty$ implies $\sum_{-\infty}^{\infty}|\gamma_j| < \infty$

The proof is easy. Last time we showed that

$$\gamma_k = \sum_{j=0}^{\infty} c_j c_{j+k}$$

then

$$\begin{aligned}\sum_{k=0}^{\infty}|\gamma_k| &= \sum_{k=0}^{\infty}|\sum_{j=0}^{\infty}c_jc_{j+k}|\\ &\leq \sum_{k=0}^{\infty}\sum_{j=0}^{\infty}|c_j||c_{j+k}|\\ &\leq \sum_{l=0}^{\infty}\sum_{j=0}^{\infty}|c_j||c_l|\\ &= \left(\sum_{j=0}^{\infty}|c_j|\right)^2 < \infty\end{aligned}$$

From the new proof of LLN one can guess that the variance in a central limit theorem should change. Remember that we wish to normalize the sum in such a way that the limit variance would be 1.

$$\begin{aligned}\text{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\right) =&\gamma_0 + 2\sum_{k=1}^{n}\gamma_k\left(1-\frac{k}{n}\right)\\ \to&\gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k = \mathcal{J}\end{aligned}$$

$\mathcal{J}$ is called the *long-run variance* and is a correct scale measure.

There are many Central Limit Theorems for serially correlated observations. The simplest is for MA(∞).

Theorem 4. *Let $y_t = \mu + \sum_{j=0}^{\infty} c_j e_{t-j}$, where e_t is independent white noise and $\sum_{j=0}^{\infty}|c_j| < \infty$, then*

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}y_t - \mu\right) \Rightarrow N(0,\mathcal{J})$$

For another version we have to introduce the following notations.

- Let I_t be information available at time t, *i.e.* I_t is the sigma-algebra generated by $\{y_j\}_{j=-\infty}^{t}$
- Let $\tau_{t,k} = E[y_t|I_{t-k}] - E[y_t|I_{t-k-1}]$ is the revision of forecast about y_t as the new information arrives at time $t-k$.

Definition 5. A strictly stationary process $\{y_t\}$ is *ergodic* if for any t, k, l and any bounded functions, g and h,

$$\lim_{n\to\infty} \text{cov}(g(y_t, ..., y_{t+k}), h(y_{t+k+n}, ...y_{t+k+n+l})) = 0$$

Theorem 6 (Gordon's CLT)**.** *Assume that we have a strictly stationary and ergodic series* $\{y_t\}$ *with* $Ey_t^2 < \infty$ *satisfying:*

1. $\sum_j (E\tau_{t,j}^2)^{1/2} < \infty$
2. $E[y_t|I_{t-j}] \to 0$ *in* $\mathcal{L}^2$ *as* $j \to \infty$

then

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_t \Rightarrow N(0, \mathcal{J}),$$

where $\mathcal{J} = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k$ *is a long-run variance.*

Remark 7. Notice, that $y_t = \sum_{j=0}^{\infty}\tau_{t,j}$. The condition 1 is intended to make the dependence between distant observations to decrease to 0. Condition 1 can be checked (see an example below). I'm not sure how the ergodicity can be easily checked. Condition 2 is aimed at the correct centering, in particular, it implies that $E[y_t] = 0$

Example 8. *AR(1)* $y_t = \rho y_{t-1} + e_t$
We can check condition 2. We have $E[y_t|I_{t-k}] - E[y_t|I_{t-k-1}] = \rho^k e_{t-k}$ and $E\tau_{t,j}^2 = \rho^{2k}\sigma^2$, so condition 2 is satisfied. More generally, if the MA has absolutely summable coefficients, then condition 2 will hold. One can notice that $E[y_t|I_{t-k}] = \rho^k y_{t-k}$, so condition 3 holds. Now let's calculate the long-run variance:

$$\gamma_k = \frac{\sigma^2\rho^k}{1-\rho^2}$$

$$\mathcal{J} = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k = \frac{\sigma^2}{1-\rho^2}\left(1 + 2\sum_{k=1}^{\infty}\rho^k\right) = \frac{\sigma^2}{(1-\rho)^2}$$

Remark 9.

$$\mathcal{J} = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k = \sum_{k=-\infty}^{\infty}\gamma_k = \gamma(1)$$

where $\gamma(1)$ is the covariance function from last lecture evaluated at 1. Recall:

$$\gamma(\xi) = \sum_{i=-\infty}^{\infty}\gamma_i\xi^i$$

and if $a(L)y_t = b(L)e_t$, then

$$\gamma(\xi) = \sigma^2\frac{b(\xi)b(\xi^{-1})}{a(\xi)a(\xi^{-1})}$$

so

$$\mathcal{J} = \left(\frac{b(1)}{a(1)}\right)^2\sigma^2$$

Remark 10. If $\{y_t\}$ is a vector, then let $\Gamma_k = \text{cov}(y_t, y_{t+k})$ and $\mathcal{J} = \sum_{-\infty}^{\infty} \Gamma_k$. The only thing that's different from the scalar case is that $\Gamma_k \neq \Gamma_{-k}$. Instead, $\Gamma_k = \Gamma'_{-k}$. All the formulas above also hold, except in matrix notation. For example, for a VARMA,

$$\mathcal{J} = A^{-1} B \Sigma B' A^{-1'}$$

Remark 11. If y_t is a martingale difference: $E[y_t|I_{t-1}] = 0$, then there is no serial correlation and $\mathcal{J} = \sigma^2$.

OLS

Suppose $y_t = x_t\beta + u_t$. In cross-section x_t is always independent from u_s if $s \neq t$ due to iid assumption, so the exclusion restriction is formulated as $E(u_t|x_t) = 0$. In time series, however, we have to describe the dependence between error terms and all regressors.

Definition 12. x_t is *weakly exogenous* if $E(u_t|x_t, x_{t-1}, ...) = 0$

Definition 13. x_t is *strictly exogenous* if $E(u_t|\{x_t\}_{t=-\infty}^{\infty}) = 0$

Usually, strict exogeneity is too strong an assumption, it is difficult to find a good empirical example for it. The weak exogeneity is much more functional (and we will mainly assume it).

OLS estimator: $\hat{\beta} = (X'X)^{-1}(X'y)$
What is the asymptotic distribution?

$$\begin{aligned}\sqrt{T}(\hat{\beta} - \beta) =& (\frac{1}{T}X'X)^{-1}(\frac{1}{\sqrt{T}}X'u) \\ =& (\frac{1}{T}\sum_t x_t x_t')^{-1}(\frac{1}{\sqrt{T}}\sum_t x_t u_t)\end{aligned}$$

Appropriate assumptions will gives us a LLN for $(\frac{1}{T}\sum_t x_t x_t') \to M$. Assume also Gordon's condition for $z_t = x_t u_t$. If u_t is weakly exogenous, then centering is OK. Gordon's CLT gives $(\frac{1}{\sqrt{T}}\sum_t x_t u_t) \Rightarrow N(0, \mathcal{J})$, which means that

$$\sqrt{T}(\hat{\beta} - \beta) \Rightarrow N(0, M^{-1}\mathcal{J}M^{-1})$$

The only thing that is different from usual is the $\mathcal{J}$. $\mathcal{J} = \sum_{-\infty}^{\infty} \gamma_j$ (where γ_j are the autocovariances of $z_t = x_t u_t$) is called the long-run variance. The long-run variance usually arise from potentially auto-dependent error terms u_t. The errors usually contains everything that is not in the regression, which is arguably auto-correlated. It also may arise from x_t being autocorrelated and from conditional heteroskedasticity of the error terms. We need to figure out how to estimate $\mathcal{J}$. This is called HAC (heteroskedasticity autocorrelation consistent) standard errors.

Remark 14. A side note on GLS. If one believes in strict exogeneity, then the estimation can be done more efficiently by using GLS. However, GLS is generally invalid if only weak exogeneity holds.

The logic here is the following. In many settings error terms u_t are arguably auto-correlated, one may think that estimation is not fully efficient (as Gauss-Markov theorem assumes that observations are uncrrelated) and could be improved. Assume for a moment that

$$y_t = \beta x_t + u_t; \quad \text{and} \quad u_t = \rho u_{t-1} + e_t.$$

Assume also for a moment that ρ is known and e_t are serially uncorrelated(white noise). You may think of transforming the system of observations and replace t' s equation with the quasi-differenced one:

$$y_t - \rho y_{t-1} = \beta(x_t - \rho x_{t-1}) + e_t;$$

or $\widetilde{y}_t = \beta\widetilde{x}_t + e_t$, where $\widetilde{y}_t = y_t - \rho y_{t-1}$ and $\widetilde{x}_t = x_t - \rho x_{t-1}$. The new system seems to be better since the errors are not autocorrelated and have the same variance (with the exception of the first one). If we have strong exogeneity then the OLS for the new system (the first equation should be corrected to have the same variance) is the efficient(BLUE). What we described is efficient GLS in this case. The problem thought is that

$$E[e_t|x_t, x_{t-1}, ...] = E[u_t|x_t, x_{t-1}, ...] - \rho E[u_{t-1}|x_t, x_{t-1}, ...]$$

However, if u_t satisfied only weak exogenuity but not strong exogenuity assumption, then the new error may not satisfy the exogenuity condition, and the OLS in the transformed system will be biased. So, unless you believe in strong exogeneity (which is extremely rare), you should not use GLS.

HAC

Assume we have a series $\{z_t\}$ satisfying Assumptions of CLT, and we want to estimate $\mathcal{J} = \sum_{-\infty}^{\infty} \gamma_k$. There are two main ways: parametric and non-parametric.

Parametric

Assume z_t is AR(p):

$$z_t = a_1 z_{t-1} + ... + a_p z_{t-p} + e_t$$

then $\mathcal{J} = \frac{\sigma^2}{a(1)^2}$, where $a(L) = 1 - a_1 L - ...a_p L^p$. We can proceed in the following way: run OLS regression of z_t on $z_{t-1}, ..., z_{t-p}$, get $\hat{a}_1, ..., \hat{a}_p$ and $\hat{\sigma}^2$, then use $\hat{a}(L) = 1 - \hat{a}_1 L - .. - \hat{a}_p L^p$ to construct $\hat{\mathcal{J}}$,

$$\hat{\mathcal{J}} = \frac{\hat{\sigma}^2}{\hat{a}(1)^2}.$$

Two important practical questions:

- What p should we use? – model selection criteria, BIC (Bayesian informaiton criteria)
- What if z_t is not AR(p)?

The second question is still an open question. Den Haan and Levin (1997) showed that if z_t is AR(p), then the convergence of the parameteric estimator is faster than the kernel estimator described below.

Non-parametric

A naïve approach

$\mathcal{J}$ is the sum of all auto-covariance. We can estimate $T-1$ of these, but not all. What if we just use the ones we can estimate, *i.e.*

$$\tilde{\mathcal{J}} = \sum_{k=T-1}^{T-1} \hat{\gamma}_k \, , \, \hat{\gamma}_k = \frac{1}{T}\sum_{j=1}^{T-k} z_j z_{j+k}$$

It turns out that this is a very bad idea.

$$\begin{aligned}\tilde{\mathcal{J}} =& \sum_{k=-(T-1)}^{T-1} \hat{\gamma}_k \\ =& \frac{1}{T}\sum_{k=-(T-1)}^{T-1}\sum_{j=1}^{T-k} z_j z_{j+k} \\ =& \frac{1}{T}(\sum_{t=1}^{T} z_t)^2 \\ =& (\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t)^2 \\ \Rightarrow & N(0,\mathcal{J})^2\end{aligned}$$

so $\tilde{\mathcal{J}}$ is not consistent; it converges to a distribution instead of a point. The problem is that we're summing too many imprecisely estimated covariances. So, the noise does not die out. For example, to estimate γ_{T-1} we use only one observation, how good can it be?

Truncated sum of sample covariances

What if we don't use all the covariances?

$$\tilde{\mathcal{J}}_2 = \sum_{k=-S_T}^{S_T} \hat{\gamma}_k$$

where $S_T < T$ and $S_T \to \infty$ as $T \to \infty$, but more slowly.

First, we have to notice that due to truncation there will be a finite sample bias. As S_T will increase the bias due to truncation should be smaller and smaller. But we don't want to increase S_T too fast for the reason stated above (we don't want to sum up noises). Assume that we can choose S_T in such a way that this estimator is consistent. Then we might face another bad small sample property: the estimate of long run variance may be negative: $\tilde{\mathcal{J}}_2 < 0$ (or in vector case, $\tilde{\mathcal{J}}_2$ not positive definite)

Example 15. Take $S_T = 1$, then $\tilde{\mathcal{J}}_2 = \hat{\gamma}_0 + 2\hat{\gamma}_1$. In small samples, we may find $\hat{\gamma}_1 < -1/2\hat{\gamma}_0$, then $\tilde{\mathcal{J}}_2$ will be negative.

Weighted, truncated sum of sample covariances

The renewed suggestion is to create a weighted sum of sample auto-covariances with weights guaranteeing positive-definiteness:

$$\hat{J} = \sum_{j=-S_T}^{S_T} k_T(j)\hat{\gamma}_j$$

Remark 16. $k_T()$ is called a kernel.

We need conditions on S_T and $k_T()$ to give us consistency and positive-definiteness. S_T should increase $S_T \to \infty$ as $T \to \infty$, but but not too fast.
$k_T()$ needs to be such that it guarantees positive-definiteness by down-weighting high lag covariances. Also need $k_T() \to 1$ for consistency.