14.384 Time Series Analysis, Fall 2007
Recitation by Paul Schrimpf
Supplementary to lectures given by Anna Mikusheva
October 17, 2008

Recitation 7

Empirical Process Theory

Let x_t be a real-valued random $k \times 1$ vector. Consider some $\Re^n$ valued function $g_t(x_t, \tau)$ for $\tau \in \Theta$, where Θ is a subset of some metric space.

Remark 1. In time series applications, generally, $\Theta = [0, 1]$

Let

$$\xi_T(\tau) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (g_t(x_t, \tau) - E g_t(x_t, \tau))$$

$\xi_T(\tau)$ is a random function; it maps each $\tau \in \Theta$ to an $\Re^n$ valued random variable. $\xi_T(\tau)$ is called an empirical process. Under very general conditions, standard arguments show that $\xi_T(\tau)$ converges pointwise, i.e. $\forall \tau_0 \in \Theta$, $\xi_T(\tau_0) \Rightarrow N(0, \sigma^2(\tau_0))$. Also, standard arguments imply that on a finite collection of points, $(\tau_1, ..., \tau_p)$,

$$\begin{bmatrix} \xi_T(\tau_1) \\ \vdots \\ \xi_T(\tau_p) \end{bmatrix} \Rightarrow N(0, \Sigma(\tau_1, ..., \tau_p)) \tag{1}$$

We would like to generalize this sort of result so that we talk about the convergence of $\xi_T()$.

Example 2. Suppose you want to test whether x_t has cdf $F(x)$. The cdf of x_t can be estimated by its empirical cdf,

$$\hat{F}_T(x) = \frac{1}{T} \sum \mathbf{1}(x_t \leq x)$$

Two possible statistics for testing whether $\hat{F}_n(x)$ equals $F(x)$ are the Kolmogorov-Smrinov statistic,

$$\sup_x \sqrt{n}(\hat{F}_n(x) - F(x))$$

and the Cramer-von Mises statistic

$$n \int (\hat{F}_n(x) - F(x))^2 dF(x)$$

This fits into the setup above with

$$\xi_T(\tau) = \frac{1}{\sqrt{T}} \left(\sum \mathbf{1}(x_t \leq \tau) - F(\tau) \right)$$

. For independent x_t, finite dimensional covergence is easy to verify and for any τ_1, τ_2 we have

$$\begin{bmatrix} \xi_T(\tau_1) \\ \xi_T(\tau_2) \end{bmatrix} \Rightarrow N\left(0, \begin{pmatrix} F(\tau_1)(1 - F(\tau_1)) & F(\tau_1) \wedge F(\tau_2) - F(\tau_1)F(\tau_2) \\ F(\tau_1) \wedge F(\tau_2) - F(\tau_1)F(\tau_2) & F(\tau_2)(1 - F(\tau_2)) \end{pmatrix}\right)$$

Definition 3. We define a *metric* for functions on Θ as $d(b_1, b_2) = \sup_{\tau \in \Theta} |b_1(\tau) - b_2(\tau)|$

Definition 4. $\mathcal{B}$ = bounded functions on Θ

Definition 5. $\mathcal{U}(\mathcal{B})$ = class of uniformly continuous (wrt $d()$) bounded functionals from $\mathcal{B}$ to $\Re$

Example 6. Examples of elements of $\mathcal{U}(\mathcal{B})$ include:

- Evaluation at a point: $f_{\tau_0}(\xi) = \xi(\tau_0)$
- Integration: $f(\xi) = \int_\Theta \xi(\tau) d\tau$

Definition 7. *convergence in* $\mathcal{B}$: $\xi_T \Rightarrow \xi$ iff $\forall f \in \mathcal{U}(\mathcal{B})$ we have $Ef(\xi_T) \to Ef(\xi)$

Remark 8. This definition of convergence implies pointwise convergence. If $\xi_T \Rightarrow \xi$, then by definition for each τ_0 and k, $E\xi_T(\tau_0)^k \to E\xi(\tau_0)^k$. Then, if the distribution of $\xi(\tau_0)$ is completely determined by its moments (as it is if, for example, $\xi(\tau_0)$ is normal or has bounded support), it follows that $\xi_T(\tau_0) \Rightarrow \xi(\tau_0)$.

Definition 9. ξ is *stochastically equicontinuous* if $\forall \epsilon > 0$, $\forall \eta > 0$, there exists $\delta > 0$ s.t.

$$\lim_{T\to\infty} P(\sup_{|\tau_1-\tau_2|<\delta} |\xi_T(\tau_1) - \xi_T(\tau_2)| > \eta) < \epsilon$$

Theorem 10. Functional Central Limit Theorem*: If*

1. *Θ is bounded*
2. *there exists a finite-dimensional distribution convergence of ξ_T to ξ (as in (1))*
3. *$\{\xi_T\}$ are stochastically equicontinuous*

then $\xi_T \Rightarrow \xi$

Remark 11. Condition 1 can be removed. Without it, condition 3 must be strengthened to: $\forall\ \epsilon, \eta > 0$ there exists a partition of Θ into finitely many sets, $\Theta_1, ...\Theta_k$ such that

$$\lim_{T\to\infty} \sup P(\max_i \sup_{\tau_1,\tau_2\in\Theta_i} |\xi_T(\tau_1) - \xi_T(\tau_2)| > \eta) < \epsilon$$

Proving the theorem involves constructing a metric on Θ such that Θ is bounded with respect to that metric, so condition 1 is really a consequence of this stronger version of condition 3.

Remark 12. Condition 2 can be checked. Condition 3 is difficult to check, but lots of work has been done to derive simpler sufficient conditions. See Andrews (1994 HoE) for some sufficient conditions. Necessary and sufficient conditions for stochastic equicontinuity are not known. However, very general sufficient conditions are known. Classes of functions for which the functional CLT holds are called P-Donsker.

Sufficient Conditions for Stochastic Equicontinuity

This is largely tangential to what we will do in class.

Definition 13. A class of functions, $\mathcal{G}$, is *P-Donsker* if for every $g \in \mathcal{G}$,

$$\frac{1}{\sqrt{T}} \left(\sum g(x_t, \cdot) - E[g(x_t, \cdot)]\right) \Rightarrow \xi$$

where $\xi \in \ell^\infty(\mathcal{G})$

In order for a class of functions to be P-Donsker, stochastic equicontinuity requires that the function class not be too complex. One way of measuring the complexity of a function class is by bracketing numbers. An ϵ bracket in L^2, $[l, u]$ is the set of all functions, f, such that $l \leq f \leq u$ pointwise with $E[|l-u|^2]^{1/2} < \epsilon$. The ϵ bracketing number written as $N_{[]}(\epsilon, \mathcal{G})$ is the minimal number of ϵ brackets needed to cover $\mathcal{G}$. An important sufficient condition for a class to be P-Donsker is the following:

Theorem 14. *Every class $\mathcal{G}$ of measurable functions with*

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{G})} d\epsilon < \infty$$

is P-Donsker. [1]

Although this condition looks strange and difficult, it can be verified in a number of interesting situations.

Example 15. Classes that are P-Donsker include

- *Distribution functions*: using brackets of the form $[\mathbf{1}(x < x_i), \mathbf{1}(x < x_{i+1})]$ with $F(x_{i+1}) - F(x_i) < \epsilon$ we can cover $\mathcal{G}$ with C/ϵ^2 brackets, so

  $$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{G})} d\epsilon \le \int_0^1 \sqrt{\log(c/\epsilon^2)} d\epsilon = \log(c) + 1$$

  is finite.

- *Parametric Classes*: if $\mathcal{G} = \{g_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ with Θ bounded, and a Lipschitz condition holds:

  $$|g_{\theta_1}(x) - g_{\theta_2}(x)| \le m(x) ||\theta_1 - \theta_2||$$

  with $E[m(x)^2] < \infty$.

- *Smooth functions* from $\mathbb{R}^d \to \mathbb{R}$ with uniformly bounded derivates of order up to $\alpha > d/2$

Another way of characterizing complexity is through uniform covering numbers and uniform entropy integrals, but I am not going to say anything about it here.

Continuous Mapping Theorem

The following theorem is important for making the functional central limit theorem useful.

Theorem 16. Continuous Mapping Theorem*: if $\xi_T \Rightarrow \xi$, then $\forall$ continuous functionals, f, $f(\xi_T) \Rightarrow f(\xi)$*

Example 17. We can use the continuous mapping theorem to get the distribution of the Kolmogorov-Smirnov and Cramer-von Mises statistics. Both: $\sup_\tau \xi(\tau)$ and $\int \xi(\tau)^2 dF(\tau)$ are continuous functionals, so

$$\sup_x \sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} \sup_x \xi(x)$$

and

$$n \int (\hat{F}_n(x) - F(x))^2 dF(x) \xrightarrow{d} \int \xi(x)^2 dF(x)$$

where $\xi(x)$ is a Brownian bridge, i.e. Gaussian with covariance function as above. We can simulate $\sup_x \xi(x)$ and $\int \xi(x)^2 dF(x)$ to find critical values for hypothesis tests.

[1] This theorem definitely holds for iid data. It might need to be modified for dependent data (e.g. the form of the integral depends on mixing coefficients), but I'm not certain.

Random Walk Asymptotics

In lecture 12, we saw that if y_t is a random walk and we estimate an $AR(1)$, then

$$T(\hat{\rho}-1) \Rightarrow \frac{\frac{1}{2}(W^2(1)-1)}{\int_0^1 W(s)^2 ds} = \frac{\int_0^1 W(s)dW(s)}{\int_0^1 W^2(s)ds}$$

It is important to understand how we derived these expressions because, unlike in the stationary case, small changes to the estimated model can greatly alter the asymptotic distribution. For example, suppose we estimate an $AR(1)$ with a constant, so we estimate,

$$y_t = \alpha + \rho y_{t-1} + u_t$$

Let $\beta = [\alpha\, \rho]' = [0\, 1]'$ and $\hat{\beta}$ be the OLS estimate. We know that:

$$\begin{bmatrix} \hat{\alpha}-\alpha \\ \hat{\rho}-\rho \end{bmatrix} = \begin{bmatrix} T & \sum y_{t-1} \\ \sum y_{t-1} & \sum y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum u_t \\ \sum y_{t-1}u_t \end{bmatrix} \tag{2}$$

To find the asymptotic distribution, we need to examine each of the sums, determine appropriate scaling factors, and write down what they converge to. We've already seen each of these sums in lecture 12, so I won't rewrite the steps here, but recall that

$$\begin{aligned}
\frac{1}{T^{3/2}}\sum_{t=1}^T y_t \Rightarrow & \sigma \int_0^1 W(t)dt \\
T^{-2}\sum_{t=1}^T y_t^2 \Rightarrow & \sigma^2 \int_0^1 W^2(s)ds \\
\frac{1}{\sqrt{T}}\sum u_t \Rightarrow & \sigma W(1) \\
\frac{1}{T}\sum y_{t-1}u_t \Rightarrow & \frac{\sigma}{2}(W^2(1)-\sigma^2)
\end{aligned}$$

These results suggest scalling $\hat{\beta}$ by $\begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}$ to arrive at a nondegenerate asymptotic distribution, *i.e.*

$$\begin{aligned}
\begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}\begin{bmatrix} \hat{\alpha}-\alpha \\ \hat{\rho}-\rho \end{bmatrix} =& \left(\begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}\begin{bmatrix} T & \sum y_{t-1} \\ \sum y_{t-1} & \sum y_{t-1}^2 \end{bmatrix}\begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}\right)^{-1}\begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}\begin{bmatrix} \sum u_t \\ \sum y_{t-1}u_t \end{bmatrix} \\
=& \begin{bmatrix} 1 & T^{-3/2}\sum y_{t-1} \\ T^{-3/2}\sum y_{t-1} & T^{-2}\sum y_{t-1}^2 \end{bmatrix}^{-1}\begin{bmatrix} T^{-1/2}\sum u_t \\ T^{-1}\sum y_{t-1}u_t \end{bmatrix} \\
\begin{bmatrix} T^{-1/2}(\hat{\alpha}-\alpha) \\ T^{-1}(\hat{\rho}-\rho) \end{bmatrix} \Rightarrow & \begin{bmatrix} \sigma & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & \int W(s)ds \\ \int W(s)ds & \int W(s)^2 ds \end{bmatrix}^{-1}\begin{bmatrix} W(1) \\ \frac{1}{2}(W(1)^2-1) \end{bmatrix}
\end{aligned}$$

From which, we see that neither $\hat{\alpha}$ nor $\hat{\rho}$ are asymptotically normal. Also, $\hat{\alpha}$ converges at the usual $1/\sqrt{T}$ rate, but $\hat{\rho}$ converges at rate $1/T$.

with Drift

Now, let's consider another modification of the model. Suppose that y_t is a random walk with drift, $y_t = y_{t-1} + \alpha + e_t$. As above, let's assume we estimate by OLS an $AR(1)$ with a constant. As above we need to analyze each of the sums in the matrices in (2). We cannot just use the results from lecture 12 because now the process for y_t is different.

- $\sum y_{t-1}$:

$$\begin{aligned}\sum y_{t-1} =& \sum_t (\alpha(t-1) + y_0 + \sum_{s<t} e_s) \\ =& (\sum_t \alpha(t-1)) + Ty_0 + \sum_t \xi_T((t-1)/T)\end{aligned}$$

  For $\sum_t \alpha t$ to have a finite limit, we must normalize it by T^{-2}. We know that $T^{-2}Ty_0 \to 0$, and $T^{-2}\sum \xi_T((t-1)/T) \to 0$ (since $T^{-3/2}\sum \xi_T((t-1)/T) = \frac{1}{T}\sum T^{-1/2}\frac{\xi_T((t-1)/T)}{\sqrt{T}} \Rightarrow \int W(s)ds$). Therefore, $T^{-2}\sum y_{t-1} \to \lim T^{-2}\sum \alpha(t-1) = \alpha/2$.

- $\sum y_{t-1}^2$: identical reasoning shows that we must normalize by T^{-3} and $\sum y_{t-1}^2 \to \alpha^2/3$
- $\sum e_t$: is unchanged, $\to \sigma W(1)$
- $\sum y_{t-1}e_t$:

$$\begin{aligned}\sum y_{t-1}e_t =& \sum((t-1)\alpha + y_0 + \sum_{s<t} e_s)e_t \\ =& \sum e_t(t-1)\alpha + \sum e_t y_0 + \frac{1}{2}(y_T^2 - \sum e_t^2)\end{aligned}$$

  The first term, $\sum e_t(t-1)\alpha$ is $O_p(T^{3/2})$, so we must normalize by at least $T^{-3/2}$. $\sum e_t y_0$ is $O_p(T^{1/2})$ and $y_T^2 - \sum e_t^2$ is $O_p(T)$, so they vanish. This leaves,

$$T^{-3/2}\sum y_{t-1}e_t \Rightarrow T^{-3/2}\sum e_t(t-1)\alpha \Rightarrow N(0, \alpha^2/3)$$

  Furthermore, jointly we have:

$$\begin{bmatrix} T^{-1/2}\sum e_t \\ T^{-3/2}\sum y_{t-1}e_t \end{bmatrix} \Rightarrow N(0, \sigma^2 \begin{bmatrix} 1 & \alpha/2 \\ \alpha/2 & \alpha^2/3 \end{bmatrix})$$

Combining these results, we see that

$$\begin{bmatrix} T^{1/2}(\hat{\alpha} - \alpha) \\ T^{3/2}(\hat{\rho} - \rho) \end{bmatrix} \Rightarrow N(0, \sigma^2 \begin{bmatrix} 1 & \alpha/2 \\ \alpha/2 & \alpha^2/3 \end{bmatrix}^{-1})$$

Thus, we obtain asymptotic normality when we estimate a random walk with drift. Also, the asymptotic variance matrix is the same as standard OLS. However, $\hat{\rho}$ converges at a faster rate than usual.