Convergence, Posterior simulation

October 6, 2013

Types of stochastic convergence

- Almost sure convergence; convergence with probability one.
- Convergence in probability.
- Convergence in mean square, or in quadratic mean, or q.m.
- Convergence in distribution.

Almost sure

$$X_n \xrightarrow[n\to\infty]{a.s.} X_\infty \Leftrightarrow P[X_n \to X] = 1$$

$$\Leftrightarrow \forall(\varepsilon > 0) P[\forall(n > N)\, |X_n - X| < \varepsilon] \xrightarrow[N\to\infty]{} 1$$

This asserts that the realized values of the X_j sequence converge in the ordinary calculus sense to X_∞, with probability one.

In probability

$$X_n \xrightarrow[n\to\infty]{P} X_\infty \Leftrightarrow \forall(\varepsilon > 0) P[|X_n - X_\infty| < \varepsilon] \xrightarrow[n\to\infty]{} 1$$

This asserts that the probability that X_n is near X_∞ gets closer to 1 as n increases. Clearly at least no stronger than a.s. convergence.

In mean square

$$X_n \xrightarrow{q.m.} X_\infty \Leftrightarrow E[(X_n - X_\infty)^2] \xrightarrow[n\to\infty]{} 0$$

X_n and X_∞ must all have finite second moments, which is not necessary for convergence in probability.

In distribution

$$X_n \xrightarrow[n\to\infty]{D} X_\infty \Leftrightarrow \forall(f \text{continuous, bounded}) E[f(X_n)] \xrightarrow[n\to\infty]{} E[f(X_\infty)]$$

There is an equivalent way to state this. Let $F_n(c) = P[X_n \leq c]$ be the cdf (cumulative distribution function) of X_n.

$$X_n \xrightarrow[n\to\infty]{D} X_\infty \Leftrightarrow \forall(c : F_\infty \text{ continuous at } c) F_n(c) \xrightarrow[n\to\infty]{} F_\infty(c)$$

Examples

- You and I each flip a coin at each time t. Your coin is fair, with probability of heads $p = .5$. I use an unfair coin with probably of heads $p_t < .5$. But every so often I change coins, bringing them closer and closer to fair, so $p_t \to .5$ as $t \to \infty$.

Examples

- You and I each flip a coin at each time t. Your coin is fair, with probability of heads $p = .5$. I use an unfair coin with probably of heads $p_t < .5$. But every so often I change coins, bringing them closer and closer to fair, so $p_t \to .5$ as $t \to \infty$.

  Convergence in Distribution

Examples

- You and I each flip a coin at each time t. Your coin is fair, with probability of heads $p = .5$. I use an unfair coin with probably of heads $p_t < .5$. But every so often I change coins, bringing them closer and closer to fair, so $p_t \to .5$ as $t \to \infty$.

  Convergence in Distribution

- We generate a record of two sequences of coin tosses. You again use a fair coin and flip it once for each t. I'm lazy, so usually I just record your flip as if it were mine. The probability that I flip my own coin gets smaller and smaller, converging to zero as $t \to \infty$.

Examples

- You and I each flip a coin at each time t. Your coin is fair, with probability of heads $p = .5$. I use an unfair coin with probably of heads $p_t < .5$. But every so often I change coins, bringing them closer and closer to fair, so $p_t \to .5$ as $t \to \infty$.

  Convergence in Distribution

- We generate a record of two sequences of coin tosses. You again use a fair coin and flip it once for each t. I'm lazy, so usually I just record your flip as if it were mine. The probability that I flip my own coin gets smaller and smaller, converging to zero as $t \to \infty$.

  Convergence in Probability

Examples

- You and I each flip a coin at each time t. Your coin is fair, with probability of heads $p = .5$. I use an unfair coin with probably of heads $p_t < .5$. But every so often I change coins, bringing them closer and closer to fair, so $p_t \to .5$ as $t \to \infty$.

  Convergence in Distribution

- We generate a record of two sequences of coin tosses. You again use a fair coin and flip it once for each t. I'm lazy, so usually I just record your flip as if it were mine. The probability that I flip my own coin gets smaller and smaller, converging to zero as $t \to \infty$.

  Convergence in Probability

- Same as above, but eventually I just stop flipping my coin at all, so from then on my sequence of coin flips is the same as yours.

- Same as above, but eventually I just stop flipping my coin at all, so from then on my sequence of coin flips is the same as yours.

  Almost sure convergence

- Same as above, but eventually I just stop flipping my coin at all, so from then on my sequence of coin flips is the same as yours.

  Almost sure convergence

- Two rules for changing the probability that I will flip my own coin.

  1. Every period there is a probability q_t that I will flip my own coin. $q_t \to 0$, but I keep q_t constant until I actually flip a coin myself, and only then switch to a lower value of q_t.
  2. I adjust q_t every period, with $q_t = q_0^t$ and $q_0 < .5$.

  The first rule generates convergence in probability, but not a.s. convergence. The second generates a.s. convergence.

Properties of convergence measures

$$X_n \xrightarrow{a.s.} X_\infty \Rightarrow X_n \xrightarrow{P} X_\infty$$

$$X_n \xrightarrow{q.m.} X_\infty \Rightarrow X_n \xrightarrow{P} X_\infty$$

$$X_n \xrightarrow{P} X_\infty \Rightarrow X_n \xrightarrow{D} X_\infty$$

$$X_n \xrightarrow{P} X_\infty \text{ and } f \text{ continuous} \Rightarrow f(X_n) \xrightarrow{P} f(X_\infty)$$

$$C \text{ a constant, } X_n \xrightarrow{P} C,\ Y_n \xrightarrow{D} Y_\infty,\ f \text{ continuous } \Rightarrow f(X_n, Y_n) \xrightarrow{D} f(C, Y_\infty)$$

Application to OLS estimator

Suppose $\{X_t, Y_t\}$ i.i.d., $E[Y_t \mid X_t] = X_t\beta$, $\mathrm{Var}(Y_t \mid X_t) = \sigma^2$, $E[X_t'X_t] = \Sigma_X$, with Σ_X finite and non-singular ($|\Sigma_X| > 0$). Then

i. $\hat{\beta}_{OLS} \xrightarrow{P} \beta$. (OLS is consistent.)

ii. $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, \sigma^2\Sigma_X^{-1})$.

Handling the unknown σ^2 and Σ_X

If we use the notation $\hat{\varepsilon} = Y - X\hat{\beta}$ for the least squares residuals, then

$$s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k} \xrightarrow{P} \sigma^2 \,.$$

Since X_t is i.i.d. and we've assumed $E[X_t'X_t] = \Sigma_X$ is finite, the law of large numbers tells us

$$\frac{X'X}{T} = \frac{\sum_t X_t'X_t}{T} \xrightarrow{P} \Sigma_X$$

What we do in practice

This means, using the properties of convergence in probability and distribution and the fact that all the elements of Σ_X^{-1} are continuous functions of Σ_X, we see that we can plug in consistent estimators of Σ_X and σ^2 to get

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, s^2(X'X/n)^{-1})$$

which means that we treat $\hat{\beta}$ as $N(\beta, s^2(X'X)^{-1})$.

A slightly awkward point: This suggests that using $s^2(X'X)^{-1}$ as the covariance matrix is an approximation to the $\sigma^2\Sigma_X^{-1}$ that appears in the asymptotic distribution. But in fact, the small-sample theory of the SNLM tells us that $\sigma^2(X'X)^{-1}$ is the true small-sample covariance matrix (under normality), and that $\sigma^2(\Sigma_X \cdot n)^{-1}$ is only an approximation.

Bayesian interpretation of the asymptotics

Since $\hat{\beta}$ becomes approximately $N(\beta, s^2(X'X)^{-1}$ in large samples, and we can see $\hat{\beta}$ but do not know β, we can treat the approximate pdf for $\hat{\beta}$ like a likelihood function, using its behavior as a function of β to trace out a pdf for β under a flat prior.

This leads us to

$$\beta \mid \hat{\beta}, s^2, X'X \sim N(\hat{\beta}, s^2(X'X)^{-1}) \, .$$

In other words, in large samples, a Bayesian given only the second moments of the data (not the full data set), will have approximately a posterior that treats the distribution of β around $\hat{\beta}$ exactly as the frequentist theory treats the distribution of $\hat{\beta}$ around β.

Making probability statements about coefficients

We know that the conditional distribution of $\beta \mid Y, X, \sigma^2$ under a flat prior is $N(\hat{\beta}, \sigma^2(X'X)^{-1}$, where $\hat{\beta}$ is the OLS estimator. We can therefore construct a pdf for each individual coefficient β_i from a normal distribution with mean $\hat{\beta}_i$ and variance σ_{xii}, where σ_{xii} is the $i'th$ diagonal element of $\sigma^2(X'X)^{-1}$. As we have discussed, in fairly large samples we can just plug in to this formula s^2, the estimated residual variance, in place of σ^2, the true residual variance, with little loss of accuracy.

Standard regression output (e.g. from `summary(lmout)` in R) gives coefficient estimates and estimated standard deviations. This allows determining posterior pdf's. The output also usually computes "significance levels" for the coefficients by finding the probability that the absolute value of $\hat{\beta}_i$ would exceed the value observed in this sample if in fact the true

β were zero. The posterior probability interpretation of this "significance level" is just that it is twice the probability in the tail of the posterior distribution for the coefficient that lies below zero (if $\hat{\beta}_i > 0$) or above zero (if $\hat{\beta}_i < 0$).

Joint statements about several coefficients

It may also be interesting to ask whether a set of coefficients might all be zero. The traditional way to check this is with an F test. One calculates an F or χ-squared statistic and concludes that the set of coefficients is unlikely all to be zero if the statistic is in its α-probability upper tail, where α is the significance level. Here is the posterior probability interpretation of the statistic and its significance level α. We construct the level curves of the marginal posterior pdf over the parameters being tested; they are ellipsoids centered on the estimated value of the parameter vector. We find the level curve on which the zero point in this space of tested parameters lies. If the probability inside that ellipsoid is $1 - \alpha$, we call α the significance level of the test. A diagram for the case of 2 tested coefficients is on the next slide.

(0,0) is not in the 95% ellipsoid