ASYMPTOTICS

1. Some basic definitions

Definition. A sequence of random variables $\{X_t,\ t = 1, \dots, \infty\}$ defined on a common probability space, so that their joint distributions are all well-defined, is a **stochastic process**.

Definition. A stochastic process $\{X_t\}$ is **stationary** if and only if for any collection of integers $t_1, \dots, t_n$ and any s the joint distribution of $X_{t_1}, \dots, X_{t_n}$ is the same as that of $X_{t_1+s}, \dots, X_{t_n+s}$.

Definition. A stationary process $\{X_t\}$ is **ergodic** if and only if for any bounded continuous function f

$$\frac{1}{T}\sum_{1}^{T} f(X_t) \xrightarrow[T\to\infty]{a.s.} E[f(X_t)]$$

Definition. Suppose $\{X_t\}_{t=t_0}^{\infty}$ is a stochastic process and $\mathcal{I}_t$ is a sequence of information sets such that, for each t, $\mathcal{I}_t$ includes $\{X_s\}_{s=t_0}^{t}$. (More generally, $\mathcal{I}_t$ can be defined as a σ-field, in which case the condition is that all X_s for $s \leq t$ are $\mathcal{I}_t$-measurable.) Then $\{X_t\}$ is a **martingale** if and only if $E[X_{t+1} \,|\, \mathcal{I}_t] = X_t$.

Note that this definition implies by the law of iterated expectations that $E_t[X_{t+s}] = X_t$, for all $s > 0$, where we are using the notation $E_t[\cdot]$ as a shorthand for $E[\cdot \,|\, \mathcal{I}_t]$.

Definition. Suppose $\{Y_t\}_{t=t_0}^{\infty}$ is a stochastic process and $\mathcal{I}_t$ is a sequence of information sets such that, for each t, $\mathcal{I}_t$ includes $\{Y_t\}_{t=t_0}^{t}$. Then Y_t is a **martingale difference** sequence if, for each t, $E_t[Y_{t+1}] = 0$.

2. Consistency of Bayes estimators and posteriors

Theorem (Martingale Convergence Theorem)**.** *If*

(a) $\{X_t\}$ *is a martingale*
(b) $X_t < Z$ *a.s. for all* t*, and*
(c) $E[|Z|] < \infty$,

then X_t *converges almost surely.*

For a proof, see the mathematical appendix to Schervish (1995) (or most advanced probability theory books).

©2005 by Christopher A. Sims. This document may be reproduced for educational and research purposes, so long as the copies contain this notice and are retained for personal use or distributed free.

Note that though we have specified an upper bound, the theorem obviously works also with a lower bound. Note also that the limit to which the martingale converges can be stochastic. The theorem asserts that with probability one the realized sequence of numbers $\{x_t\}$ converges in the ordinary calculus sense to a limit x_∞, where x_∞ can be the value of a random variable, i.e. a different number for different draws of the random $\{x_t\}$ sequence.

Theorem. *If there is a consistent estimator $\hat{\theta}_T$ of $\theta \in \Theta$, if $f(\theta)$ is continuous and bounded above and below, and if $E[f(\theta)]$ exists, then $E[f(\theta) \mid Y_T]$, the Bayesian posterior mean for $f(\theta)$, converges to $f(\theta_0)$, the true value, with probability one.*

For a careful proof, see Schervish (1995, p.429, Theorem 7.78). An outline of the idea of the proof runs as follows. We know that $E_T[f(\theta)]$ is the minimum-variance predictor of $f(\theta)$ based on the time-T information set. But boundedness of f and consistency of $\hat{\theta}_T$ imply that $E[(f(\hat{\theta}_T) - f(\theta))^2 \mid \theta] \xrightarrow[t\to\infty]{} 0$ for each θ. This implies further that the unconditional expectation $E[(f(\hat{\theta}_T) - f(\theta))^2]$ converges to zero. (The convergence of conditional probabilities is not uniform, but because the convergence occurs for each θ, the probability of the set of θ's for which the discrepancy is large must be getting steadily smaller, and the discrepancy is bounded.) But the fact that the unconditional expectation is going to zero means also that

$$\begin{aligned} &E[(f(\hat{\theta}_T) - f(\theta))^2 \mid Y_T] \\ &\qquad = (f(\hat{\theta}_T) - E[f(\theta) \mid Y_T])^2 + E[(f(\theta) - E[f(\theta) \mid Y_T])^2 \mid Y_T] \xrightarrow[T\to\infty]{P} 0\,. \end{aligned} \tag{1}$$

(The crossproduct term drops out because of the properties of conditional expectation.) Since this is the sum of two positive terms, the fact that the sum goes to zero in probability implies that each component term goes to zero in probability, and thus in particular that $E_T[f(\theta)] \xrightarrow{P} f(\theta)$. Then since the martingale convergence theorem tells us that $E_T[f(\theta)]$ converges a.s. to some limiting variable Z, it must be that $Z = f(\theta)$. (a.s. convergence implies convergence in probability, and a sequence can't converge in probabiity to more than one limiting random variable.)

This theorem puts no restrictions on the prior distribution, but its "a.s." convergence assertion is with respect to the joint probability on parameters and observations induced by the prior and the model. Therefore the convergence it asserts could fail on sets with prior probability zero, and if the prior puts probability zero on big subsets of Θ, the convergence could fail on those big sets. On the other hand, if the prior puts non-zero probability on every open subset of Θ, as do distributions like multivariate t or Normal, then the sets on which convergence fails must be small, at least in a certain sense.

Proposition. *If $P[\theta \in S(Y)] \geq 1 - \delta$, Then $P\big[P[\theta \in S(Y) \mid Y] \geq 1 - \sqrt{\delta}\big] \geq 1 - \sqrt{\delta}$.*

This result implies that if a T^{-p}-consistent estimator exists, then the posterior distribution is T^{-p}-consistent "in probability":

Proposition. *If there is a T^{-p}-consistent estimator, in the sense that for any $\delta > 0$ there is an $A < \infty$ such that*

$$\liminf_{T\to\infty} P[\|\hat{\theta}_T - \theta\| < AT^{-p} \mid \theta] \geq 1 - \delta\,,$$

then for any $\delta^* > 0$ there is an $A^* < \infty$ such that

$$P\big[P[\|\theta - \theta_0\| < A^* T^{-p} \mid \mathcal{I}_T] > 1 - \delta^*\big] > 1 - \delta^*\,.$$

3. Asymptotics of likelihood shape

Suppose the likelihood has the form $\prod p(y_t, \theta)$, where y_t is a vector of observed values and θ is the parameter vector. This will obviously be true for i.i.d. observations. It also holds in some other models (particularly time series) where the data, though dependent across t, can be described as generated from i.i.d. disturbances.

Log likelihood for a sample of size T is then

$$\ell(Y_T, \theta) = \sum_{t=1}^{T} \log\big(p(y_t, \theta)\big)\,.$$

Taking a Taylor expansion around the MLE $\hat{\theta}_T$ leads to

$$\ell(Y_T, \theta) = \ell(Y_T, \hat{\theta}_T) + \tfrac{1}{2}(\theta - \hat{\theta}_T)' \sum_{t=1}^{T} \frac{\partial^2 \log(p(y_t, \bar{\theta}_T)}{\partial\theta\partial\theta'}(\theta - \hat{\theta}_T)\,,$$

where $\bar{\theta}$ lies on the line connecting θ and $\hat{\theta}$.

Consider the sequence of regions $\mathcal{N}(\hat{\theta}, A/\sqrt{T})$, where $\mathcal{N}(a, b)$ stands for the sphere of radius b about the point a. Suppose it's true that $\hat{\theta}_T$ for large enough T gets close to some limiting point θ_0 with high probability, so that $\mathcal{N}(\hat{\theta}_T, A/\sqrt{T}) \subset \mathcal{N}(\theta_0, 2A/\sqrt{T})$ with high probability.

Assume also that the law of large numbers applies to $\partial^2 \log(p)/\partial\theta\partial\theta'$ at each point in the parameter space Θ and that its expected value is continuous. The point $\bar{\theta}_T$ at which we take the second derivative in our expansion is not fixed, but it is in an ever-shrinking neighborhood of the fixed θ_0. So it is reasonable to suppose, and weak regularity conditions would let us prove, that

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 \log(p(y_t, \bar{\theta})}{\partial\theta\partial\theta'} \xrightarrow[T\to\infty]{P} E\left[\frac{\partial^2 \log(p(y_t, \theta_0)}{\partial\theta\partial\theta'}\right]\,.$$

If we now consider $\ell(\hat{\theta}+\delta/\sqrt{T})-\ell(\hat{\theta})$ and apply our Taylor expansion formula, we arrive at

$$\left(\frac{\delta}{\sqrt{T}}\right)' T\left(\frac{1}{T}\frac{\partial^2\ell(Y_T,\hat{\theta})}{\partial\theta\partial\theta'}\right)\left(\frac{\delta}{\sqrt{T}}\right)\xrightarrow[T\to\infty]{P}\delta' E\left[\frac{\partial^2\log(p(y_t,\theta)}{\partial\theta\partial\theta'}\right]\delta\,.$$

Plugging this result back in to the Taylor expansion for log likelihood, we have the conclusion that for large samples, in a neighborhood of the likelihood maximum that shrinks at the rate $1/\sqrt{T}$, the likelihood is well approximated by a Gaussian shape centered at the likelihood maximum and with a covariance matrix

$$-\left(\frac{\partial^2\ell(Y_T,\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}$$

.

4. Conditions for asymptotic normality of the likelihood

- No particular assumptions about the shape of p were needed. The result comes from the *local* properties of the likelihood around θ_0.
- θ_0 as a limiting value must be present, but the argument did not depend on p being the actual pdf of the data or on θ being the "true" value. Of course this is not much comfort, since statements about posterior probability based on the likelihood will not be correct if the likelihood does not correspond to the true probability distribution.
- This result applies to likelihood-based inference. A different kind of result is needed for a Bayesian interpretation of estimators that cannot be connected to the likelihood.
- The result depends on the assumption that $\sqrt{T}(\hat{\theta}_T-\theta)$ remains bounded in probability, which requires that a $T^{-1/2}$-consistent estimator exist.

5. Central limit theorems

There is no uniquely most general CLT. One has to limit time-dependence in the summands, and one has to limit the degree of non-stationarity. Here is the most general theorem we will apply:

Theorem (Martingale Central Limit Theorem)**.** *If $\{X_t\}$ is a stationary, ergodic martingale-difference process with mean zero and finite variance matrix Σ, then*

$$\frac{1}{\sqrt{T}}\sum_1^T X_t \xrightarrow[T\to\infty]{\mathcal{D}} N(0,\Sigma)\,.$$

6. OLS

The propositions below can all be proved under much more general assumptions. The proofs here are chosen to be simple and to apply the theorems we've listed above.

Theorem (Asymptotic Normality of OLS)**.** *Suppose $y_t = X_t\beta + \varepsilon_t$ for all t and that X_t is predetermined, meaning $E[\varepsilon_t \mid X_{t-s}, \varepsilon_{t-s-1}, s > 0] = 0$. Suppose also that X_t, ε_t are jointly stationary and ergodic, with $E[X_t'X_t] = \Sigma_X < \infty$, $|\Sigma_X| \neq 0$, and* $\mathrm{Var}(\varepsilon_t \mid X_t) = \sigma^2 < \infty$*. Then*

$$\sqrt{T}(\hat{\beta}_{OLS} - \beta) \xrightarrow[T\to\infty]{\mathcal{D}} N(0, \sigma^2(X'X)^{-1})\,.$$

Proof.

$$T^{1/2}(\hat{\beta}_{OLS} - \beta) = \left(\frac{X'X}{T}\right)^{-1} \frac{1}{\sqrt{T}} X'\varepsilon\,.$$

The $(1/T)X'X$ is a time average that by ergodicity converges in probability to Σ_X. The second term meets the criteria of the Martingale Central Limit Theorem (be sure you know why the terms in the sum have a finite covariance matrix) and hence is asymptotically $N(0, \Sigma_X)$. We established some weeks ago, that if f is continuous and $Z_t \xrightarrow{P} Z_\infty$, $f(Z_t) \xrightarrow{P} f(Z_\infty)$, and that if f is continuous and bounded in two arguments, $X_t \xrightarrow{P} X_\infty$, and $W_t \xrightarrow{\mathcal{D}} W_\infty$, then $f(X_t, W_t) \xrightarrow{P} f(X_\infty, W_\infty)$, Applying these results gives us our conclusion. □

- If the errors are non-normal, but the second derivative of the log pdf of the disturbance matches that of the normal at $\varepsilon = 0$, (as, e.g., for t-distributed errors), the asymptotic likelihood shape will be the same as that of OLS with normal errors and the same second-derivative matrix of the log likelihood. A $t_n(\Sigma)$ distribution has a second derivative of the log likelihood at its peak that matches that of the normal with the same Σ. So our theorem asserts that, despite the fact that a $t_p(\sigma^2)$ distributed variable has a larger variance matrix than the $N(0, \sigma^2)$, if the correct likelihood is used estimates of β in a regression model with $t_p(\sigma^2)$-distributed errors are in large samples as good as would have been obained with $N(0, \sigma^2)$ errors.
- However in this case the usual OLS statistics are not generally sufficient, and the pdf of the parameters conditional on the OLS statistics will be more dispersed than the conditional pdf given the full sample, even in large samples.

7. "Flipping" non-Bayesian asymptotic results

Usually,

$$\left\{\sqrt{T}(\hat{\theta} - \theta) \mid \theta\right\} \xrightarrow{\mathcal{D}} N(0, \Sigma) \Rightarrow \left\{\sqrt{T}(\hat{\theta} - \theta) \mid \hat{\theta}\right\} \xrightarrow{\mathcal{D}} N(0, \Sigma)\,.$$

In other words, non-Bayesian asymptotic distribution results can be "flipped" to become Bayesian asymptotic distribution results of the same form. However, the Bayesian results are for distributions conditioned on the estimators themselves, not on the entire sample. There will usually be information loss in conditioning on estimators alone. Nonetheless this type of result allows in many cases a Bayesian interpretation of reported estimates and standard errors that have been based on non-Bayesian asymptotics. A paper that lays out the needed regularity conditions is Kwan (1998).

References

KWAN, Y. K. (1998): "Asymptotic Bayesian analysis based on a limited information estimator," *Journal of Econometrics*, 88, 99–121.

SCHERVISH, M. J. (1995): *Theory of Statistics*, Springer Series in Statistics. Springer, New York.