TA section 7

May 22nd, 2009 by Yang

1. Optimal Forecast for Gaussian Processes

Time series processes are Gaussian if they have a jointly normal distribution.[1] For Guassian processes, the optimal forecast is actually linear.

Theorem. Let Y_{t+1} and X_t be jointly Gaussian processes as follows.

$$\begin{pmatrix} Y_{t+1} \\ X_t \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} \right)$$

Then,

$$Y_{t+1}|X_t \sim N\left(\mu_1 + \Omega_{12}\Omega_{22}^{-1}(X_t - \mu_2), \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right)$$

Proof. See Theorem 3.5.3. in Hogg et al. (2005, p.175), and Chapter 4.6 in Hamilton (1994, p.100).

Applying the above theorem, it follows that

$$\begin{aligned} E[Y_{t+1}|X_t] &= \mu_1 + \Omega_{12}\Omega_{22}^{-1}(X_t - \mu_2) \\ &= (\mu_1 - \Omega_{12}\Omega_{22}^{-1}\mu_2) + \Omega_{12}\Omega_{22}^{-1}X_t \\ &= \begin{pmatrix} \mu_1 - \Omega_{12}\Omega_{22}^{-1}\mu_2 & \Omega_{12}\Omega_{22}^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ X_t \end{pmatrix} \end{aligned}$$

which means that $E[Y_{t+1}|X_t]$ is a linear function of 1 and X_t. Since $E[Y_{t+1}|X_t]$ minimizes the mean squared error among all forecasts, it is the optimal linear forecast when it is a linear function.

2. HAC Estimation

Consider the model

$$y_t = x_t'\theta_0 + u_t, \qquad t = 1, \cdots, T$$

where

$$Ex_tu_t = 0$$

One of the consistent estimators of θ_0 is

$$\widehat{\theta}_{OLS} = \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t y_t$$

Then, $\widehat{\theta}_{OLS}$ has the following asymptotic distribution.

$$\sqrt{T}(\widehat{\theta}_{OLS} - \theta_0) \xrightarrow{d} N\left(0, (Ex_tx_t')^{-1}\Omega(Ex_tx_t')^{-1}\right)$$

[1] In other words, their joint pdf is a Gaussian function.

where

$$\Omega := \lim_{T\to\infty} var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t u_t\right)$$

We are interested in estimating Ω when the errors are heteroskedastic and the data are autocorrelated. As technical conditions, we need stationary and mixing $x_t u_t$. Since $Ex_t u_t = 0$, we have

$$\Omega = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} Ex_t u_t x_s' u_s$$

Define

$$\Omega_T := \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} Ex_t u_t x_s' u_s$$

then by definition, $\Omega_T \to \Omega$ as $T \to \infty$. Let $v_t := x_t u_t$ and define

$$\Gamma_T(j) := \begin{cases} \frac{1}{T}\sum_{t=j+1}^{T} Ev_t v_{t-j}' & j \geq 0 \\ \frac{1}{T}\sum_{t=-j+1}^{T} Ev_{t+j} v_t' & j < 0 \end{cases}$$

then,

$$\Omega_T = \sum_{j=-T+1}^{T-1} \Gamma_T(j)$$

Now we know that $Ev_t v_{t-j}$ is invariant to t, so if $j \geq 0$ is relatively small,

$$\frac{1}{T-j}\sum_{t=j+1}^{T} v_t v_{t-j}' \xrightarrow{p} Ev_t v_{t-j}'$$

by the (dependent and nonidentical version of) law of large numbers. This ensures that for such j,

$$\frac{1}{T}\sum_{t=j+1}^{T} v_t v_{t-j}' - \Gamma_T(j) = \frac{T-j}{T}\frac{1}{T-j}\sum_{t=j+1}^{T} v_t v_{t-j}' - \frac{T-j}{T}Ev_t v_{t-j}' \xrightarrow{p} 0$$

The same property holds for $j < 0$ relatively small in absolute value. Define

$$\widetilde{\Gamma}_T(j) := \begin{cases} \frac{1}{T}\sum_{t=j+1}^{T} v_t v_{t-j}' & j \geq 0 \\ \frac{1}{T}\sum_{t=-j+1}^{T} v_{t+j} v_t' & j < 0 \end{cases}$$

Let S_T be relatively small. For any j such that $|j| \leq S_T$, we have $\widetilde{\Gamma}_T(j) - \Gamma_T(j) \xrightarrow{p} 0$. For any j such that $|j| > S_T$, we hope $\Gamma_T(j)$ is so small that we may ignore. This is the idea that leads to the HAC estimator of Ω.

Theorem. Suppose (1) $k : \mathbb{R} \to [-1, 1]$ be a kernel such that $k(0) = 1$, $\int_{-\infty}^{\infty} k^2(x)dx < \infty$ and k is symmetric, continuous at 0 and continuous up to a finite number of points, (2) $S_T \to \infty$ and $\frac{S_T}{T} \to 0$ as $T \to \infty$, (3) $x_t u_t$ is mean 0, stationary and mixing, and (4) $\sqrt{T}(\widehat{\theta}_{OLS} - \theta_0) = O_p(1)$. Then, the HAC estimator defined by

$$\widehat{\Omega}_{T,HAC} := \sum_{j=-T+1}^{T+1} k\left(\frac{j}{S_T}\right) \widetilde{\Gamma}_T(j)$$

is consistent for Ω.

We use the following kernels.

1. Truncated kernel
$$k_{TR}(x) = \begin{cases} 1 & |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$
2. Bartlett kernel
$$k_{BT}(x) = \begin{cases} 1 - |x| & |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$
3. Parzen kernel
$$k_{PR}(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & |x| \leq 1/2 \\ 2(1 - |x|)^3 & 1/2 \leq |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

For example, if we use a truncated kernel,

$$\widehat{\Omega}_{T,TR} := \sum_{j=-S_T}^{S_T} \widetilde{\Gamma}_T(j)$$

If we use Bartlett kernel,

$$\widehat{\Omega}_{T,BT} := \sum_{j=-S_T}^{S_T} \left(1 - \frac{|j|}{S_T}\right) \widetilde{\Gamma}_T(j)$$

Remark. With the above kernels, we drop all $v_t v'_{t-j}$ for $j > |S_T|$ in constructing $\widehat{\Omega}_{T,HAC}$, so the HAC estimator is not unbiased. But if S_T is close to T, $\widehat{\Omega}_T$ is not consistent for Ω. So choosing S_T is a tradeoff between bias and variance. Note that both the asymptotic bias and the nonshrinking variance lead to inconsistency of the estimator. The condition (2) in the above theorem ensures that $\Omega_{T,HAC}$ is consistent for Ω.

Remark. Bartlett kernel downweighs $v_t v'_{t-j}$ for j close to S_T in absolute value, while a truncated kernel does not. Although both give a consistent estimator, Bartlett kernel may produce more biased estimator than a truncated kernel. But Bartlett kernel always produces an estimated matrix that is positive definite, while a truncated kernel does not. Parzen kernel also always yields a positive definite matrix. Choice of a kernel is a less significant problem than that of a bandwidth S_T.

3. Bootstrap Methods

Nonparametric bootstrap is nothing but resampling the same size of observations from the data with replacement. With the resampled set of observations, we calculate the statistics we are interested in. Do this B times, then we have B statistics.

Example. How to obtain t-statistic with bootstrap method.
Let $(W_i)_{i=1}^n$ be the original data.

1. Fix b, and draw n random numbers $U_{bi}^* \sim U[0,1]$, $i = 1, \cdots, n$.
2. Let $W_{bi}^* := W_{\ulcorner nU_{bi}^* \urcorner}$ where $\ulcorner a \urcorner$ is a ceiling function that gives the smallest number no less than a. Then, it follows that
$$\Pr(W_{bi}^* = W_j) = \Pr(j-1 < nU_{bi}^* \leq j) = \frac{1}{n}$$
for all j, and that W_{bi}^* is independent over b and i since U_{bi}^* is.
3. From $(W_{bi}^*)_{i=1}^n$, calcualte $\widehat{\theta}_{nb}^*$ and $\widehat{\omega}_{nb}^*$ in the same way as $\widehat{\theta}_n$ and $\widehat{\omega}_n$. Form
$$T_{nb}^* = \frac{\sqrt{n}(\widehat{\theta}_{nb}^* - \widehat{\theta}_n)}{\widehat{\omega}_{nb}^*}$$
4. Repeat 1-3 for $b = 1, \cdots, B$.

We claim that those statistics have the same distribution as the original statistic from the data does. Theoretical justification for this is that the data tell very much about its distribution, so the empirical distribution approximates the original distribution well. Then, resampling from the data with replacement is approximately equivalent to sampling from the original distribution. In the above example, T_n would be distirubted in the approximately same way as T_{nb}^*, where

$$T_n = \frac{\sqrt{n}(\widehat{\theta}_n - \theta_0)}{\widehat{\omega}_n}$$

is the original statistic. As we have B such statistics, those approximate the distribution of the original statistic. So we use the critical value obtained from $(T_{nb}^*)_{b=1}^B$, instead of that from the exact distribution of T_n, when we test, or construct CI. See the lecture note for details on two-sided symmetric CI and two-sided equal-tailed CI.

Remark. We calcualte $T_{nb}^* = \frac{\sqrt{n}(\widehat{\theta}_{nb}^* - \widehat{\theta}_n)}{\widehat{\omega}_{nb}^*}$ for several reasons. First, $\widehat{\theta}_n$ is the true parameter in the bootstrap world, and also believed to be consistent for the original true parameter. Second, when we want to construct CI, there is no information on the true parameter. Third, even when we test the null hypothesis $H_0 : \theta = \theta_0$, using the above formula gives better size and power properties. If we use critical values obtained from $\widetilde{T}_{nb}^* = \frac{\sqrt{n}(\widehat{\theta}_{nb}^* - \theta_0)}{\widehat{\omega}_{nb}^*}$, the power of the test is very poor, although the null rejection probability may (or may not) converge to the nomial size as n grows.