SPURIOUS CORRELATION IN TIME SERIES

In time series analysis, we explore the correlation structure between two jointly-stationary time series[1] $X = \{X_t\}$ and $Y = \{Y_t\}$ and their lead-lag relationship using the cross-covariance function, where the cross-covariance function is given by $\gamma_{XY}(t,s) = cov(X_t, Y_s)$ for each pair of integers t and s. For jointly stationary processes, the **cross-correlation function** between X and Y at lag k can then be defined by

$$\rho_{XY}(k) = corr(X_t, Y_{t-k}) = corr(X_{t+k}, Y_t).$$

The coefficient $\rho_{XY}(0)$ measures the contemporaneous linear association between X and Y, whereas $\rho_{XY}(k)$ measures the linear association between X_t and that of Y_{t-k}[2].

Consider a time series regression model

$$Y_t = \beta_0 + \beta_1 X_{t-d} + e_t, \qquad (1)$$

where the X's are independent, identically distributed random variables with variance σ_X^2 and and the e's are also white noise with variance σ_e^2 and are independent of the X's. It can be checked that the cross-correlation function (CCF) $\rho_{XY}(k)$ is zero except for lag $k = -d$, where

$$\rho_{XY}(-d) = \frac{\beta_1 \sigma_X}{\sqrt{\beta_1^2 \sigma_X^2 + \sigma_e^2}}. \qquad (2)$$

In this case, the theoretical CCF is nonzero only at lag $-d$, reflecting the fact that X is "leading" Y by d units of time. The CCF can be estimated by the **sample cross-correlation function** (sample CCF) defined by

$$r_{XY}(k) = \frac{\sum(X_t - \bar{X})(Y_{t-k} - \bar{Y})}{\sqrt{\sum(X_t - \bar{X})^2}\sqrt{\sum(Y_t - \bar{Y})^2}}, \qquad (3)$$

where the summations are done over all data where the summands are available.

According to eqn. (1), the covariate X is independent of Y if and only if $\beta_1 = 0$, in which case the sample autocorrelation $r_{XY}(k)$ is approximately normally distributed with zero mean and variance $1/n$, where n is the sample size—the number of pairs of (X_t, Y_t) available. Sample

[1] Stationarity of a univariate time series can be easily extended to the case of multivariate time series. For example, X and Y are jointly (weakly) stationary if their means are constant and the covariance $\gamma_{XY}(t,s)$ is independent of time and a function of the time difference $t - s$.

[2] Unlike the autocorrelation function, the cross-correlation function is generally not an even function since $corr(X_t, Y_{t-k})$ need not equal $corr(X_t, Y_{t+k})$.

cross-correlations that are larger than $1.96/\sqrt{n}$ in magnitude are then deemed significantly different from zero. However, the above inference is valid only under the condition that X and Y are mutually independent white noise processes.

If X and Y are not white noise (even if the processes X and Y are independent of each other i.e. $\beta_1 = 0$), the sample CCF between X and Y is no longer approximately $N(0,1/n)$. Specifically, under the assumption that both X and Y are stationary and that they are independent of each other, $\sqrt{n}\, r_{XY}(k)$ is asymptotically normal with mean zero and variance

$$1 + 2\sum_{j=1}^{\infty} \rho_X(j)\rho_Y(j), \qquad (4)$$

where $\rho_X(k)$ is the autocorrelation of X at lag k. For refinement of this asymptotic result, see Brockwell and Davis (1991, p.410).

Suppose X and Y are both AR(1) processes with AR(1) coefficients ϕ_X and ϕ_Y, respectively. Then $\sqrt{n} \cdot r_{XY}(k)$ is approximately normally distributed with zero mean, but the variance is now approximately equal to

$$\frac{1 + \phi_X\phi_Y}{(1 - \phi_X\phi_Y)}. \qquad (5)$$

When both AR(1) coefficients are close to 1, the ratio of the sampling variance of $r_{XY}(k)$ to the nominal value of $1/\mathrm{n}$ approaches infinity. Thus, the unquestioned use of the $1/n$ rule in deciding the significance of the sample CCF may lead to many more false positives than the nominal 5% error rate, even though the response and covariate time series are independent of each other. The problem of inflated variance of the sample cross-correlation coefficients becomes more acute for nonstationary data.

Prewhitening and Transfer function noise models

In the preceding section, we found that it is difficult to assess the dependence between the two serially correlated processes. Thus, it is pertinent to disentangle the linear association between *X* and *Y*, say, from their autocorrelation. A useful device for doing this is *prewhitening* which was first introduced in the section of the Transfer Function Noise modeling.

For example, consider a single input transfer function noise model

$$y_t = \nu(B)x_t + n_t, \quad (6)$$

where x_t and n_t are mutually independent and both follow stationary ARMA process. In particular, we consider

$$\phi_X(B)x_t = \theta_X(B)\alpha_t, \quad \alpha_t{\sim}WN(0,\sigma_\alpha^2). \quad (7)$$

Prewhitening applies the filter $\phi_X(B)/\theta_X(B)$ on both sides of eqn. (6). After prewhitening, eqn. (7) becomes

$$\underbrace{\frac{\phi_X(B)}{\theta_X(B)}y_t}_{\beta_t} = \nu(B)\underbrace{\frac{\phi_X(B)}{\theta_X(B)}x_t}_{\alpha_t} + \frac{\phi_X(B)}{\theta_X(B)}n_t,$$

where α_t and β_t denote the prewhitened series for X and Y, respectively. Note that by design α_t is white noise but β_t need not be white noise.

If β_t is stationary, then (i) statistical significance of the sample CCF of the prewhitened data may be assessed using the cutoff $1.96/\sqrt{n}$, and (ii) the theoretical counterpart of the CCF so estimated is proportional to certain regression coefficients. Specifically, we have

$$\nu_j = \rho_{\beta\alpha}(j)\frac{\sigma_\beta}{\sigma_\alpha},$$

where σ_β and σ_α denote the standard deviation of β_t and α_t, respectively.