Introduction to Time Series Analysis. Lecture 24.

1. Review: parametric spectral estimation
2. Lagged regression models.
3. Cross-covariance function, sample CCF.
4. Lagged regression in the time domain: prewhitening.

Review: Parametric spectral density estimation

Given data $x_1, x_2, \ldots, x_n$,

1. Estimate the AR parameters $\phi_1, \ldots, \phi_p, \sigma_w^2$ (for example, using Yule-Walker or maximum likelihood),
   and choose a suitable model order p (for example, using $\text{AIC}_c = (n+p)/(n-p-2)$ or $\text{SIC} = p \log n/n$).

2. Use the estimates $\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\sigma}_w^2$ to compute the estimated spectral density:

$$\hat{f}_y(\nu) = \frac{\hat{\sigma}_w^2}{\left|\hat{\phi}\left(e^{-2\pi i \nu}\right)\right|^2}.$$

Review: Parametric spectral density estimation

For large n,

$$\mathrm{Var}(\hat{f}(\nu)) \approx \frac{2p}{n} f^2(\nu).$$

Bias-variance trade-off:

As p increases, the bias decreases—we can model more complex spectra—but the variance increases.

Advantage over nonparametric: better *frequency resolution* of a small number of peaks. This is especially important if there is more than one peak at nearby frequencies.

Disadvantage: inflexibility (bias).

Lagged regression models

Consider a lagged regression model of the form

$$Y_t = \sum_{h=-\infty}^{\infty} \beta_h X_{t-h} + V_t,$$

where X_t is an observed input time series, Y_t is the observed output time series, and V_t is a stationary noise process.

This is useful for

- Identifying the (best linear) relationship between two time series.
- Forecasting one time series from the other.
  (We might want $\beta_h = 0$ for $h < 0$.)

Lagged regression models

$$Y_t = \sum_{h=-\infty}^{\infty} \beta_h X_{t-h} + V_t.$$

In the SOI and recruitment example, we might wish to identify how the values of the recruitment series (the number of new fish) is related to the Southern Oscillation Index.
Or we might wish to predict future values of recruitment from the SOI.

Lagged regression models: Agenda

- Multiple, jointly stationary time series in the time domain: cross-covariance function, sample CCF.
- Lagged regression in the time domain: model the input series, extract the white time series driving it ('prewhitening'), regress with transformed output series.
- Multiple, jointly stationary time series in the frequency domain: cross spectrum, coherence.
- Lagged regression in the frequency domain: Calculate the input's spectral density, and the cross-spectral density between input and output, and find the transfer function relating them, in the frequency domain. Then the regression coefficients are the inverse Fourier transform of the transfer function.

Cross-covariance

Recall that the autocovariance function of a stationary process $\{X_t\}$ is

$$\gamma_x(h) = \mathbf{E}\left[(X_{t+h} - \mu_x)(X_t - \mu_x)\right].$$

The *cross-covariance function* of two jointly stationary processes $\{X_t\}$ and $\{Y_t\}$ is

$$\gamma_{xy}(h) = \mathbf{E}\left[(X_{t+h} - \mu_x)(Y_t - \mu_y)\right].$$

(Jointly stationary = constant means, autocovariances depending only on the lag h, and cross-covariance depends only on h.)

Cross-correlation

The *cross-correlation function* of jointly stationary $\{X_t\}$ and $\{Y_t\}$ is

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}.$$

Notice that $\rho_{xy}(h) = \rho_{yx}(-h)$ (but $\rho_{xy}(h) \not\equiv \rho_{xy}(-h)$).

Example: Suppose that $Y_t = \beta X_{t-\ell} + W_t$ for $\{X_t\}$ stationary and uncorrelated with $\{W_t\}$, and W_t zero mean and white. Then $\{X_t\}$ and $\{Y_t\}$ are jointly stationary, with $\mu_y = \beta\mu_x$,

$$\gamma_{xy}(h) = \beta\gamma_x(h+\ell).$$

If $\ell > 0$, we say x_t *leads* y_t.
If $\ell < 0$, we say x_t *lags* y_t.

Sample cross-covariance and sample CCF

$$\hat{\gamma}_{xy}(h) = \frac{1}{n} \sum_{i=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y})$$

for $h \geq 0$ (and $\hat{\gamma}_{xy}(h) = \hat{\gamma}_{yx}(-h)$ for $h < 0$).

The sample CCF is

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}.$$

Sample cross-covariance and sample CCF

If either of $\{X_t\}$ or $\{Y_t\}$ is white, then $\hat{\rho}_{xy}(h) \sim AN(0, 1/\sqrt{n})$.

Notice that we can look for peaks in the sample CCF to identify a leading or lagging relation. (Recall that the ACF of the input series peaks at $h = 0$.)

Example: CCF of SOI and recruitment (Figure 1.14 in text) has a peak at $h = -6$, indicating that recruitment at t has its strongest correlation with SOI at time $t - 6$. Thus, SOI leads recruitment (by 6 months).

Lagged regression in the time domain (Section 5.6)

Suppose we wish to fit a lagged regression model of the form

$$Y_t = \alpha(B)X_t + \eta_t = \sum_{j=0}^{\infty} \alpha_j X_{t-j} + \eta_t,$$

where X_t is an observed input time series, Y_t is the observed output time series, and η_t is a stationary noise process, uncorrelated with X_t.

One approach (pioneered by Box and Jenkins) is to fit ARIMA models for X_t and η_t, and then find a simple rational representation for $\alpha(B)$.

Lagged regression in the time domain

$$Y_t = \alpha(B)X_t + \eta_t = \sum_{j=0}^{\infty} \alpha_j X_{t-j} + \eta_t,$$

For example:

$$X_t = \frac{\theta_x(B)}{\phi_x(B)} W_t,$$

$$\eta_t = \frac{\theta_\eta(B)}{\phi_\eta(B)} Z_t,$$

$$\alpha(B) = \frac{\delta(B)}{\omega(B)} B^d.$$

Notice the delay B^d, indicating that Y_t lags X_t by d steps.

Lagged regression in the time domain

How do we choose all of these parameters?

1. Fit $\theta_x(B), \phi_x(B)$ to model the input series $\{X_t\}$.

2. *Prewhiten* the input series by applying the inverse operator $\phi_x(B)/\theta_x(B)$:

$$\tilde{Y}_t = \frac{\phi_x(B)}{\theta_x(B)} Y_t = \alpha(B) W_t + \frac{\phi_x(B)}{\theta_x(B)} \eta_t.$$

Lagged regression in the time domain

3. Calculate the cross-correlation of $\tilde{Y}_t$ with W_t,

$$\gamma_{\tilde{y},w}(h) = \mathbf{E}\left(\sum_{j=0}^{\infty} \alpha_j W_{t+h-j} W_t\right) = \sigma_w^2 \alpha_h,$$

   to give an indication of the behavior of $\alpha(B)$ (for instance, the delay).

4. Estimate the coefficients of $\alpha(B)$ and hence fit an ARMA model for the noise series η_t.

Lagged regression in the time domain

Why prewhiten?
The prewhitening step inverts the linear filter $X_t = \theta_x(B)/\phi_x(B)W_t$. Then the lagged regression is between the transformed Y_t and a white series W_t. This makes it easy to determine a suitable lag.

For example, in the SOI/recruitment series, we treat SOI as an input, estimate an AR(1) model, prewhiten it (that is, compute the inverse of our AR(1) operator and apply it to the SOI series), and consider the cross-correlation between the transformed recruitment series and the prewhitened SOI. This shows a large peak at lag -5 (corresponding to the SOI series leading the recruitment series). Examples 5.7 and 5.8 in the text then consider $\alpha(B) = B^5/(1 - \omega_1 B)$.

Lagged regression in the time domain

This sequential estimation procedure (ϕ_x, θ_x, then α, then ϕ_η, θ_η) is rather ad hoc. State space methods offer an alternative, and they are also convenient for vector-valued input and output series.