Lecture 8

Estimation of Correlation Part 2

Arnab Hazra

Sample autocovariance function (Recap)

- Suppose the realizations are $x_1, \ldots, x_T$.
- The sample autocovariance function is defined as

$$\hat{\gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (x_{t+h} - \bar{x})(x_t - \bar{x})$$

with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for $h = 0, 1, \ldots, T-1$.

- Why not just divide by $T - h$ instead of T?
- Hint: Ensure that $\widehat{\mathrm{Var}}(a_1 X_1 + \ldots + a_T X_T)$ is also non-negative.

Sample ACF

- The sample ACF is defined as

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

- Large-Sample Distribution of the ACF: If X_t are white noise with finite fourth moment, then for T large, the sample ACF, $\hat{\gamma}(h)$, for $h = 1, 2, \dots, H$, where H is fixed but arbitrary, is approximately normally distributed with zero mean and standard deviation given by

$$\sigma_{\hat{\rho}}(h) = 1/\sqrt{T}.$$

- We obtain a rough method of assessing whether peaks in $\hat{\rho}(h)$ are significant by determining whether the observed peak is outside the interval $\pm 2/\sqrt{T}$.

Example

- Suppose X_t are IID with $P(X_t = 1) = 0.5$ and $P(X_t = -1) = 0.5$.
- We construct $Y_t = 5 + X_t - 0.7X_{t-1}$
- Calculate $\rho_Y(1)$ and compare

```
set.seed(101010)
x1 = 2*rbinom(11, 1, .5) - 1     # simulated sequence of coin tosses
x2 = 2*rbinom(101, 1, .5) - 1
y1 = 5 + filter(x1, sides=1, filter=c(1,-.7))[-1]
y2 = 5 + filter(x2, sides=1, filter=c(1,-.7))[-1]
plot.ts(y1, type='s'); plot.ts(y2, type='s')    # plot both series (not shown)
c(mean(y1), mean(y2))                # the sample means
   [1] 5.080   5.002
acf(y1, lag.max=4, plot=FALSE)   # 1/√10 = .32
   Autocorrelations of series 'y1', by lag
       0      1      2      3      4
   1.000 -0.688  0.425 -0.306 -0.007
acf(y2, lag.max=4, plot=FALSE)   # 1/√100 = .1
   Autocorrelations of series 'y2', by lag
       0      1      2      3      4
   1.000 -0.480 -0.002 -0.004  0.000
```

Sample cross-covariance function

- Two time series X_t and Y_t are said to be jointly stationary if they are each stationary, and the cross-covariance function

$$\gamma_{X,Y}(h) = \mathrm{Cov}(X_{t+h}, Y_t) = E[(X_{t+h} - \mu_X)(Y_t - \mu_Y)]$$

  is a function only of lag h.

- Suppose the realizations are $x_1, \ldots, x_T$ and $y_1, \ldots, y_T$.
- The sample cross-covariance function is defined as

$$\hat{\gamma}_{X,Y}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (x_{t+h} - \bar{x})(y_t - \bar{y})$$

  with $\hat{\gamma}_{X,Y}(-h) = \hat{\gamma}_{Y,X}(h)$ for $h = 0, 1, \ldots, T-1$.

Sample CCF

- The cross-correlation function (CCF) of jointly stationary time series X_t and Y_t is defined as

$$\rho_{X,Y}(h) = \frac{\gamma_{X,Y}(h)}{\sqrt{\gamma_X(0) \times \gamma_Y(0)}}.$$

- The sample CCF is

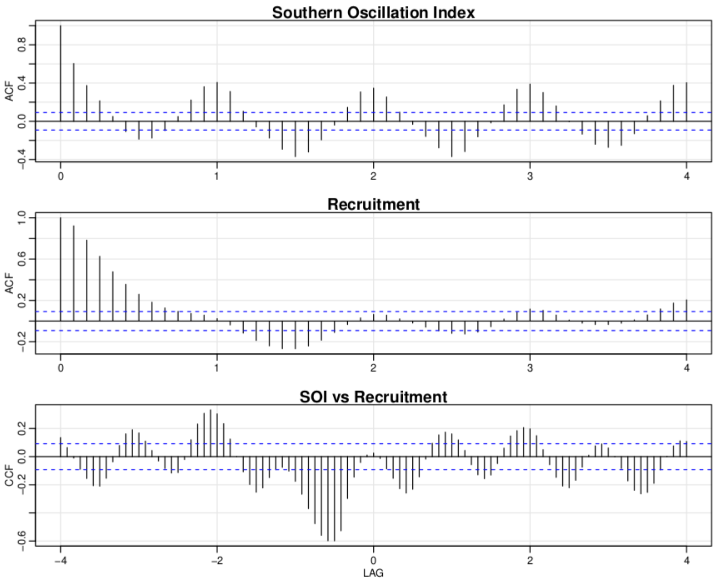
$$\hat{\rho}_{X,Y}(h) = \frac{\hat{\gamma}_{X,Y}(h)}{\sqrt{\hat{\gamma}_X(0) \times \hat{\gamma}_Y(0)}}.$$

- Large-Sample Distribution of sample CCF: The large sample distribution of $\hat{\rho}_{X,Y}(h)$ is normal with mean zero and

$$\sigma_{\hat{\rho}_{X,Y}} = 1/\sqrt{T}$$

if at least one of the processes is independent white noise.

Sample ACF and CCF

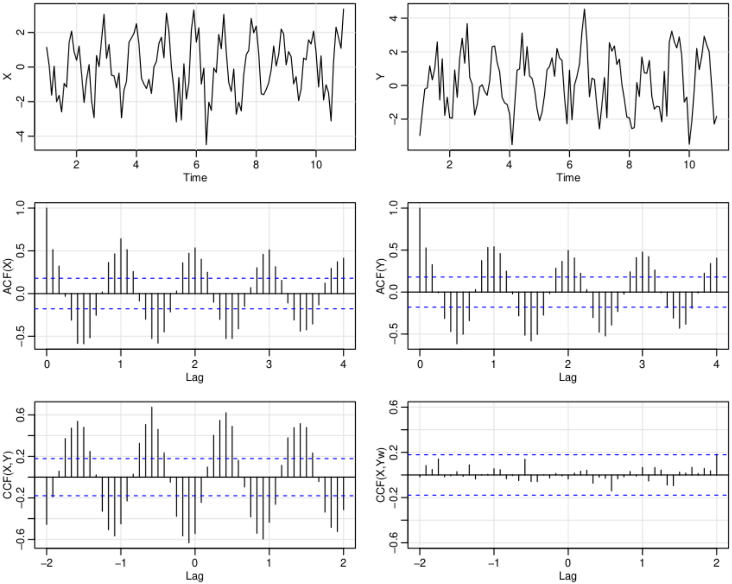


Prewhitening

By prewhtiening Y_t, we mean that the signal has been removed from the data by running a regression of Y_t on $\cos(2\pi t)$ and $\sin(2\pi t)$ and then putting $\tilde{Y}_t = Y_t - \hat{Y}_t$.

```
set.seed(1492)
num=120; t=1:num
X = ts(2*cos(2*pi*t/12) + rnorm(num), freq=12)
Y = ts(2*cos(2*pi*(t+5)/12) + rnorm(num), freq=12)
Yw = resid( lm(Y~ cos(2*pi*t/12) + sin(2*pi*t/12), na.action=NULL) )
par(mfrow=c(3,2), mgp=c(1.6,.6,0), mar=c(3,3,1,1) )
plot(X)
plot(Y)
acf(X,48, ylab='ACF(X)')
acf(Y,48, ylab='ACF(Y)')
ccf(X,Y,24, ylab='CCF(X,Y)')
ccf(X,Yw,24, ylab='CCF(X,Yw)', ylim=c(-.6,.6))
```

Prewhitening example



Vector-valued time series

- We frequently encounter situations in which the relationships between a number of jointly measured time series are of interest.
- For example, we considered discovering the relationships between the SOI and Recruitment series.
- A vector time series $\boldsymbol{X}_t = (X_{t1}, X_{t2}, \ldots, X_{tp})'$ contains p univariate time series as its components.
- For the stationary case, the p-length mean vector is $E[\boldsymbol{X}_t] = \boldsymbol{\mu}$ and $p \times p$ covariance matrix
$$\boldsymbol{\Gamma}(h) = E[(\boldsymbol{X}_{t+h} - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})']$$
- Here $\boldsymbol{\Gamma}(h) = [E[(X_{t+h,i} - \mu_i)(X_{t,j} - \mu_j)], i, j = 1, \ldots, n]$ and $\boldsymbol{\Gamma}(-h) = \boldsymbol{\Gamma}(h)'$.

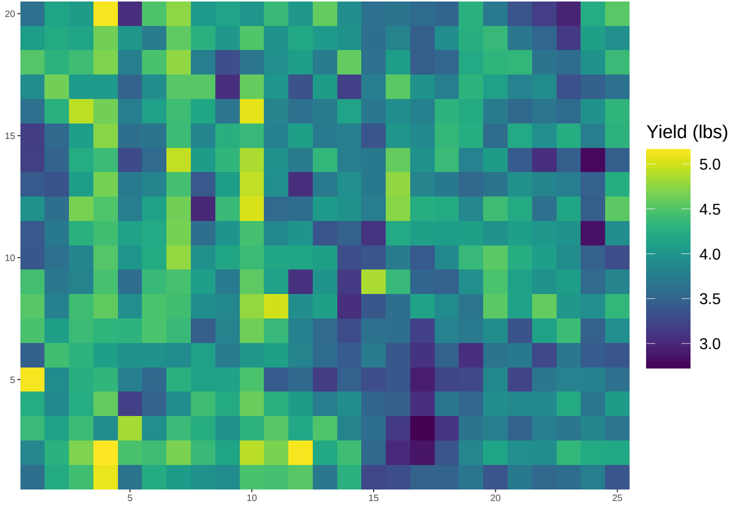
Sample autocovariance matrix

- Suppose the realizations are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$.
- The sample autocovariance matrix of the vector series $\boldsymbol{X}_t$ is the $p \times p$ matrix of sample cross-covariances, defined as

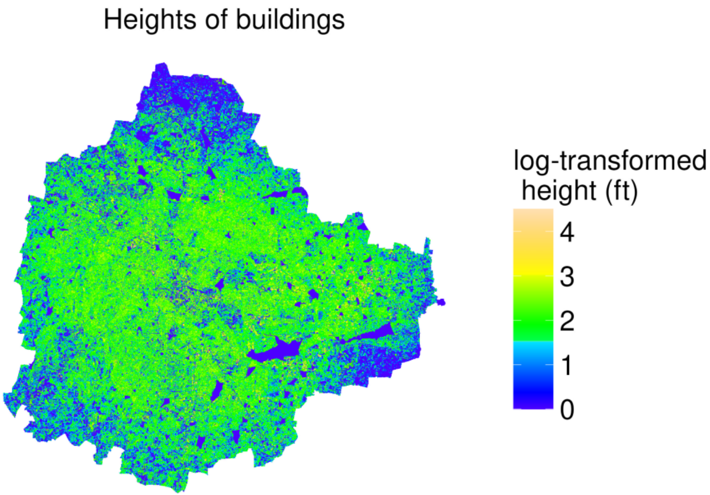
$$\hat{\Gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (\boldsymbol{x}_{t+h} - \bar{\boldsymbol{x}})(\boldsymbol{x}_t - \bar{\boldsymbol{x}})'.$$

with $\bar{\boldsymbol{x}} = \sum_{t=1}^{T} \boldsymbol{x}_t$ and $\hat{\Gamma}(-h) = \hat{\Gamma}(h)'$.

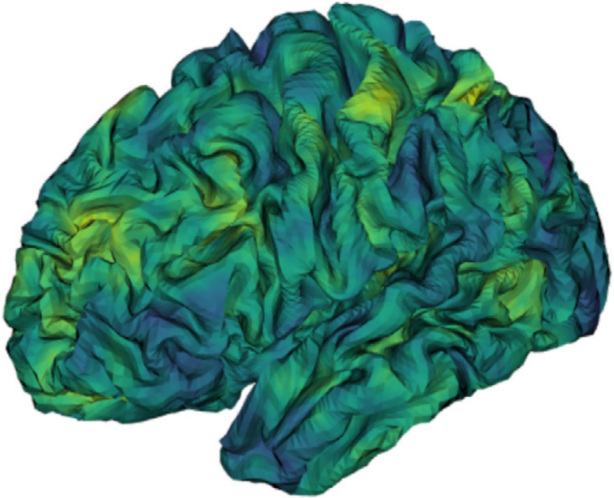
Multidimensional process: 2-D Example for regular domain



Multidimensional process: 2-D Example for irregular domain



Multidimensional process: 3-D Example for irregular domain



Stationary multidimensional process: regular domain

- We can define a multidimensional process $X_{\boldsymbol{s}}$ as a function of the $r \times 1$ vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_r)'$, where s_i denotes the coordinate of the ith index.
- Assuming stationarity, the autocovariance function of $X_{\boldsymbol{s}}$ can be defined as a function of the multidimensional lag vector, say, $\boldsymbol{h} = (h_1, h_2, \ldots, h_r)'$, as

$$\gamma(\boldsymbol{h}) = \mathrm{Cov}(X_{\boldsymbol{s}+\boldsymbol{h}}, X_{\boldsymbol{s}}) = E[(X_{\boldsymbol{s}+\boldsymbol{h}} - \mu)(X_{\boldsymbol{s}} - \mu)]$$

  where $\mu = E(X_{\boldsymbol{s}})$.
- The multidimensional sample autocovariance function is defined as

$$\hat{\gamma}(\boldsymbol{h}) = (S_1 S_2 \cdots S_r)^{-1} \sum_{s_1} \sum_{s_2} \ldots \sum_{s_r} (x_{\boldsymbol{s}+\boldsymbol{h}} - \bar{x})(x_{\boldsymbol{s}} - \bar{x})$$

  where $\bar{x} = (S_1 S_2 \cdots S_r)^{-1} \sum_{s_1} \sum_{s_2} \ldots \sum_{s_r} x_{\boldsymbol{s}}$.

Stationary multidimensional process: irregular domain

- A standard measure of dependence is variogram given by

$$2V_X(\boldsymbol{h}) = \mathrm{Var}(X_{\boldsymbol{s}+\boldsymbol{h}} - X_{\boldsymbol{s}}).$$

- Here V_X is called semivariance and twice of it is called variogram.

- A sample estimator is

$$2\widehat{V}_X(\boldsymbol{h}) = \frac{1}{N(h)} \sum_{\boldsymbol{s}} (X_{\boldsymbol{s}+\boldsymbol{h}} - X_{\boldsymbol{s}})^2.$$

- Here $N(h)$ denotes both the number of points located within h, and the sum runs over the points in the neighborhood.

Thank you!