Estimating Means and Covariances

- In other statistical applications, means, variances, and covariances are estimated by averaging across samples.

- In time series, we often have only one realization.

- *Stationarity* allows us to estimate moments anyway.
Mean

- If $x_t$ is stationary, $\mu_t = \mathbb{E}(x_t) \equiv \mu$, so we can estimate $\mu$ by the sample mean

$$\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t.$$ 

- We could also use a *weighted* mean

$$\sum_{t=1}^{n} w_t x_t,$$

where

$$\sum_{t=1}^{n} w_t = 1.$$ 

- Both are unbiased; usually some weighted mean has smaller variance than $\bar{x}$, but not *much* smaller.
Autocovariance

• Similarly, if $x_t$ is stationary, $\gamma(t+h,t) = \text{cov}(x_{t+h},x_t) \equiv \gamma(h)$, so we can estimate $\gamma(h)$ by

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

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

• We estimate the autocorrelation function (ACF) by

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

• $\bar{x}$ is unbiased for $\mu$.

• $\hat{\gamma}(h)$ is not unbiased for $\gamma(h)$, but

$$\frac{1}{n-h} \sum_{t=1}^{n-h} (x_{t+h} - \mu)(x_t - \mu)$$

would be. Note:

– $(n-h)$ denominator instead of $n$;

– centering at $\mu$ instead of $\bar{x}$.
Non-negative Definiteness

- The covariance matrix of \((x_1, x_2, \ldots, x_k)\) is

\[
\Gamma_k = \begin{bmatrix}
\gamma(0) & \gamma(1) & \cdots & \gamma(k-1) \\
\gamma(1) & \gamma(0) & \cdots & \gamma(k-2) \\
\vdots & \vdots & \ddots & \vdots \\
\gamma(k-1) & \gamma(k-2) & \cdots & \gamma(0)
\end{bmatrix}
\]

and, as a covariance matrix, is non-negative definite:

\[
a'\Gamma_k a = \text{var}(a_1x_1 + a_2x_2 + \cdots + a_kx_k) \geq 0
\]

for any vector of constants \(a = (a_1, a_2, \ldots, a_k)'\).

- With the above definition of \(\hat{\gamma}(h)\), \(\hat{\Gamma}_k\) is also non-negative definite; that would not be true if we divided by \((n - h)\).
Another Sampling Property

- If $x_t$ is white noise and $n$ is large and some mild conditions hold, $\hat{\rho}(h)$ is approximately normal with zero mean and standard deviation

$$\sigma_{\hat{\rho}(h)} = \frac{1}{\sqrt{n}}.$$ 

- So we can look for autocorrelations outside $\pm 2/\sqrt{n}$ as evidence of autocorrelation.
R Examples

- White noise:

  \[ \text{acf(ts(rnorm(100)))}; \]

- Southern Oscillation Index and fish recruitment:

  \[ \text{soi = scan("http://www.stat.pitt.edu/stoffer/tsa2/data/soi.dat");} \]
  \[ \text{soi = ts(soi, start = 1950, frequency = 12);} \]
  \[ \text{recruit = scan("http://www.stat.pitt.edu/stoffer/tsa2/data/recruit.dat");} \]
  \[ \text{recruit = ts(recruit, start = 1950, frequency = 12);} \]
  \[ \text{acf(soi, 50);} \]
  \[ \text{acf(recruit, 50);} \]
  \[ \text{ccf(soi, recruit, 50);} \]
  \[ \# \text{Negative lags indicate SOI leads recruitment.} \]
Interpreting the Cross-Correlation

- `help(ccf)` states: “The lag ’k’ value returned by ’ccf(x,y)’ estimates the correlation between ’x[t+k]’ and ’y[t]’.”

- So the graph shows negative correlation between SOI(t - 5 to 9 months) and recruit(t).

- That is, current recruitment is (negatively) correlated with SOI from 5 – 9 months ago.
SAS Example

- Southern Oscillation Index and fish recruitment:

  options pagesize = 80;

  data soi;
  infile 'soi.dat';
  input soi;
  run;

  data recruit;
  infile 'recruit.dat';
  input recruit;
  run;
data both;
   time +1;
   merge soi recruit;
run;

proc gplot data = both;
   symbol i = join;
   plot (soi recruit) * time;
run;

proc arima data = both;
   title 'SOI and recruitment';
   identify var = soi nlag = 50;
   identify var = recruit crosscorr = soi nlag = 50;
   /* Positive lags indicate SOI leads recruitment. */
run;

SAS program and output.
Seasonality in the SOI

- The ACF of the SOI suggests that $x_t$ has a correlation of around 0.4 with $x_{t+12}$, $x_{t+24}$, and so on.

- This “correlation” is caused by the fact that those values all fall in the same month of the year, and different months have different means.

- That is, this series has a non-constant mean function $\mu_t$.

- Since it is non-stationary in the mean, the sample ACF does not estimate the population ACF, and the graph has no meaning.
• We can estimate $\mu_t$ and subtract it, to give a series with zero mean.

• The simplest way is to subtract the mean for a given month of the year from all data for that month.

• In R (in SAS, use corresponding `proc glm`):

```r
soiSA = residuals(lm(soi ~ factor(cycle(soi))));
# transfer the time series structure of soi to soiSA:
soiSA = ts(soiSA, start = start(soi), frequency = frequency(soi));
acf(soiSA, lag = 50);
```
• The ACF graph now shows correlation dropping progressively from around 0.5 at a one month lag to zero at one year.

• The CCF of soiSA and recruit shows correspondingly simpler structure.

• Frequency-domain methods will show that the recruitment series also has some seasonality, but with much weaker effects.

• Replacing recruit with a corresponding recruitSA makes negligible changes to the ACF and CCF.
• Frequency-domain methods will also show that the seasonal effects in SOI consist largely of an annual sine wave.

• Instead of estimating 12 separate monthly means, we can fit, and remove, a three-parameter model

\[ \mu_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi t}{12}\right) + \beta_2 \sin\left(\frac{2\pi t}{12}\right). \]

• In R:

```r
soiCS = residuals(lm(soi ~ cos(2 * pi * time(soi)) +
                   sin(2 * pi * time(soi))));
soiCS = ts(soiCS, start = start(soi), frequency = frequency(soi));
acf(soiCS, lag = 50);
```