Stats531 Homework1

Liangqi Tang

Question 1.1 Recall the basic properties of covariance, $\mathrm{Cov}(X,Y) = E\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big]$, following the convention that upper case letters are random variables and lower case letters are constants:

P1. $\mathrm{Cov}(Y,Y) = \mathrm{Var}(Y)$,

P2. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$,

P3. $\mathrm{Cov}(aX,bY) = ab\,\mathrm{Cov}(X,Y)$,

P4. $\mathrm{Cov}\left(\sum_{m=1}^{M} Y_m, \sum_{n=1}^{N} Y_n\right) = \sum_{m=1}^{M}\sum_{n=1}^{N} \mathrm{Cov}(Y_m, Y_n)$.

Let $Y_{1:N}$ be a covariance stationary time series model with autocovariance function γ_h and constant mean function, $\mu_n = \mu$. Consider the sample mean as an estimator of μ,

$$\hat{\mu}(y_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} y_n.$$

Show how the basic properties of covariance can be used to derive the expression,

$$\mathrm{Var}(\hat{\mu}(Y_{1:N})) = \frac{1}{N}\gamma_0 + \frac{2}{N^2}\sum_{h=1}^{N-1}(N-h)\gamma_h.$$

Proof:

$$
\begin{aligned}
\mathrm{Var}(\hat{\mu}(Y_{1:N})) &= \mathrm{Var}(\frac{1}{N}\sum_{n-1}^{N} y_n) \\
&= \frac{1}{N^2}\mathrm{Var}(\sum_{n=1}^{N} y_n) \\
&= \frac{1}{N^2}\mathrm{Cov}(y_1 + \cdots + y_n, y_1 + \cdots + y_n) \\
&= \frac{1}{N^2}[\sum_{n=1}^{N}\mathrm{Cov}(y_n, y_n) + \sum_{n=2}^{N}\mathrm{Cov}(y_1, y_n) + \cdots + \sum_{n=N-1}^{N}\mathrm{Cov}(y_{N-2}, y_n) \\
&\quad + \mathrm{Cov}(y_{N-1}, y_N) + \mathrm{Cov}(y_2, y_1) + \sum_{n=1}^{2}\mathrm{Cov}(y_3, y_n) + \cdots + \sum_{n=1}^{N-1}\mathrm{Cov}(y_N, y_n)] \\
&= \frac{1}{N^2}[N\gamma_0 + (N-1)\gamma_{-1} + (N-2)\gamma_{-2} + \cdots + \gamma_{-(N-1)} \\
&\quad + (N-1)\gamma_1 + (N-2)\gamma_2 + \cdots + \gamma_{N-1}] \\
&= \frac{1}{N^2}[N\gamma_0 + 2(N-1)\gamma_1 + 2(N-2)\gamma_2 + \cdots + 2\gamma_{N-1}] \\
&= \frac{1}{N}\gamma_0 + \frac{2}{N^2}\sum_{h=1}^{N-1}(N-h)\gamma_h
\end{aligned}
$$

Question 1.2 The sample autocorrelation is perhaps the second most common type of plot in time series analysis, after simply plotting the data. We investigate how R represents chance variation in the plot of the sample autocorrelation function produced by the `acf` function. We seek to check what R actually does when it constructs the dashed horizontal lines in this plot. What approximation is being made? How should the lines be interpreted statistically?

If you type `acf` in R, you get the source code for the acf function. You'll see that the plotting is done by a service function `plot.acf`. This service function is part of the package, and is not immediately accessible to you. Nevertheless, you can check the source code as follows:

1. Notice, either from the help documentation `?acf` or the last line of the source code `acf` that this function resides in the package `stats`.
2. Now, you can access this namespace directly, to list the source code, by

```
stats:::plot.acf
```

3. Now we can see how the horizontal dashed lines are constructed. The critical line of code seems to be

```
clim0 <- if (with.ci) qnorm((1 + ci)/2)/sqrt(x$n.used)
```

This appears to correspond to a normal distribution approximation for the sample autocorrelation estimator, with mean zero and standard deviation $1/\sqrt{N}$.

A. This question investigates the use of $1/\sqrt{N}$ as an approximation to the standard deviation of the sample autocorrelation estimator under the null hypothesis that the time series is a sequence of independent, identically distributed (IID) mean zero random variables.

Instead of studying the full autocorrelation estimator, you are asked to analyze a simpler situation where we take advantage of the knowledge that the mean is zero and consider

$$\hat{\rho}_h(Y_{1:N}) = \frac{\frac{1}{N}\sum_{n=1}^{N-h} Y_n\, Y_{n+h}}{\frac{1}{N}\sum_{n=1}^{N} Y_n^2}$$

where $Y_1, \ldots, Y_N$ are IID random variables with zero mean and finite variance. Specifically, find the mean and standard deviation for $\hat{\rho}_h(Y_{1:N})$ when N becomes large.

The actual autocorrelation estimator subtracts a sample mean, and you can analyze that instead if you want an additional challenge.

You will probably want to make an argument based on linearization. You can reason at whatever level of math stat formalization you're happy with. According to *Mathematical Statistics and Data Analysis* by John Rice, a textbook used for the undergraduate upper level Math Stats course, STATS 426,

"When confronted with a nonlinear problem we cannot solve, we linearize. In probability and statistics, this method is called **propagation of errors** or the δ method. Linearization is carried out through a Taylor Series expansion."

Rice then proceeds to describe the delta method in a way very similar to the Wikipedia article on this topic. In summary, suppose X is a random variable with mean μ_X and small variance σ_X^2, and $g(x)$ is a nonlinear function with derivative $g'(x) = dg/dx$. To study the random variable $Y = g(X)$ we can make a Taylor series approximation,

$$Y \approx g(\mu_X) + (X - \mu_X)g'(\mu_X).$$

This approximates Y as a linear function of X, so we have

1. $\mu_Y = \mathbb{E}[Y] \approx g(\mu_X)$.
2. $\sigma_Y^2 = \mathrm{Var}(Y) \approx \sigma_X^2\{g'(\mu_X)\}^2$.
3. If $X \sim N[\mu_X, \sigma_X^2]$, then Y approximately follows a $N[g(\mu_X), \sigma_X^2\{g'(\mu_X)\}^2]$ distribution.

For this question, we have a two-dimensional situation, where $Y = g(U, V)$ and the Taylor series approximation becomes

$$Y \approx g(\mu_U, \mu_V) + (U - \mu_U)\frac{\partial}{\partial u}g(\mu_U, \mu_V) + (V - \mu_V)\frac{\partial}{\partial v}g(\mu_U, \mu_V)$$

with $U = \hat{\gamma}_h(Y_{1:N}) = \frac{1}{N}\sum_{n=1}^{N-h} Y_n\, Y_{n+h}$ and $V = \hat{\gamma}_0(Y_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} {Y_n}^2$. Finding the mean and variance of U and V requires similar techniques to Question 1.1.

Solution:

Suppose:

$$\mathbb{E}[Y] = 0, \quad \mathrm{Var}(Y) = \sigma^2$$

Denote:

$$U = \hat{\gamma}_h(Y_{1:N}) = \frac{1}{N}\sum_{n=1}^{N-h} Y_n Y_{n+h}$$
$$V = \hat{\gamma}_0(Y_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} Y_n^2$$

then:

$$\begin{aligned}\mu_U = \mathbb{E}[U] &= \frac{1}{N}\sum_{n=1}^{N-h}\mathbb{E}[Y_n Y_{n+h}] \\ &= \frac{1}{N}\sum_{n=1}^{N-h}\mathbb{E}[Y_n]\mathbb{E}[Y_{n+h}] \\ &= 0\end{aligned}$$

$$\begin{aligned}\mu_V = \mathbb{E}[V] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[Y_n^2] \\ &= \frac{1}{N}\sum_{n=1}^{N}(\mathrm{Var}(Y_n) + \mathbb{E}[Y_n]) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sigma^2 \\ &= \sigma^2\end{aligned}$$

$$
\begin{aligned}
\sigma_U^2 = \mathrm{Var}(U) &= \frac{1}{N^2}\mathrm{Var}(\sum_{n=1}^{N-h} Y_n Y_{n+h}) \\
&= \frac{1}{N^2}\sum_{i=1}^{N-h}\sum_{j=1}^{N-h}\mathrm{Cov}(Y_i Y_{i+h}, Y_j Y_{j+h}) \\
&= \frac{1}{N^2}\sum_{i=1}^{N-h}\sum_{j=1}^{N-h}(\mathbb{E}[Y_i Y_j Y_{i+h} Y_{j+h}] - \mathbb{E}[Y_i Y_j]\mathbb{E}[Y_j Y_{j+h}]) \\
&= \frac{1}{N^2}\sum_{k=1}^{N-h}\mathbb{E}[Y_k^2 Y_{k+h}^2] \\
&= \frac{1}{N^2}\sum_{k=1}^{N-h}\mathbb{E}[Y_k^2]\mathbb{E}[Y_{k+h]}^2] \\
&= \frac{1}{N^2}\sum_{k=1}^{N-h}(\mathrm{Var}(Y_k) + \mathbb{E}[Y_k])(\mathrm{Var}(Y_{k+h}) + \mathbb{E}[Y_{k+h}]) \\
&= \frac{1}{N^2}\sum_{k=1}^{N-h}\sigma^4 \\
&= \frac{N-h}{N^2}\sigma^4
\end{aligned}
$$

$$
\sigma_V^2 = \mathrm{Var}(V) = \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{Var}(Y_n^2)
$$

$$
\begin{aligned}
\sigma_{UV}^2 = \mathrm{Cov}(U,V) &= \frac{1}{N^2}\sum_{i=1}^{N-h}\sum_{j=1}^{N}\mathrm{Cov}(Y_i Y_{i+h}, Y_j^2) \\
&= \frac{1}{N^2}\sum_{i=1}^{N-h}\sum_{j=1}^{N}\mathbb{E}[Y_i Y_{i+h} Y_j^2] \\
&= 0
\end{aligned}
$$

Since:

$$
\hat{\rho}_h(Y_{1:N}) = \frac{\frac{1}{N}\sum_{n=1}^{N-h} Y_n\, Y_{n+h}}{\frac{1}{N}\sum_{n=1}^{N} Y_n^2} = \frac{U}{V}
$$

Using Taylor series approximation, let $g(U,V) = \frac{U}{V}$, then:

$$\begin{aligned}\hat{\rho}_h(Y_{1:N}) &\approx g(\mu_U,\mu_V) + (U-\mu_U)\frac{\partial}{\partial U}g(\mu_U,\mu_V) + (V-\mu_V)\frac{\partial}{\partial V}g(\mu_U,\mu_V)\\ &= \frac{0}{\sigma^2} + (U-0)\cdot\frac{1}{\sigma^2} + (V-\sigma^2)\cdot(-\frac{0}{\sigma^4})\\ &= \frac{U}{\sigma^2}\end{aligned}$$

$$\mathbb{E}[\hat{\rho}_h(Y_{1:N})] = \frac{1}{\sigma^2}\mathbb{E}[U] = 0$$

$$\mathrm{Sd}(\hat{\rho}_h(Y_{1:N})) = \sqrt{\frac{1}{\sigma^4}\sigma_U^2} = \sqrt{\frac{N-h}{N^2}}$$

when $n \to \infty$:

$$\mathrm{Sd}(\hat{\rho}_h(Y_{1:N})) = \sqrt{\frac{N-h}{N^2}} \to \frac{1}{\sqrt{N}}$$

B. It is often asserted that the horizontal dashed lines on the sample ACF plot represent a confidence interval. For example, in the documentation produced by `?plot.acf` we read

```
ci: coverage probability for confidence interval.
Plotting of the confidence interval is suppressed if 'ci' is zero or negative.
```

Use a definition of a confidence interval to explain how these lines do, or do not, construct a confidence interval.

Solution:

The definition for the confidence interval here is that if we assume that the time series data we plot is just a set of white noise, then the sample ACF(autocorrelation function) will be in the interval with the probability of 95% given n(sample size) is large enough.

We can use the lines to construct a confidence interval and test our hypothesis like this: under the white noise hypothesis, approximately 95% of our sample ACF dots will be between the lines. If there are obvious more than 5% of our sample ACF dots appear beyond the lines, then we can reject our hypothesis that the time series data is just white noise.

Also, I have referred to *Time Series Analysis and Its Applications* by Robert H. Shumway and David S. Stoffer.

"Based on the previous result, 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 $\pm 2/\sqrt{n}$ (or plus/minus two standard errors); for a white noise sequence, approximately 95% of the sample ACFs should be within these limits."