CHAPTER 10

EXERCISES FOR MODEL SELECTION

In this problem set you will estimate the annual cycle of the NINO3.4 index and apply model selection criteria to determine the number of terms in the Fourier series that should be fitted. The NINO3.4 index can be downloaded from

```
1 http://www.esrl.noaa.gov/psd/data/correlation/nina34.data
```

An R-code for reading this data set is `Rcode.exercise.Chapter10.R`, which can be downloaded from the class website. You will need to modify the variables `dir.Rlib` and `dir.data` to point to the directories containing your R-functions and ENSO time series, respectively. The variable `nharm` specifies the maximum number of Fourier harmonics to be considered. When you run this code, it should produce a figure showing the *raw* NINO3.4 time series (without annual cycle removed).

The exercises below will ask you to write a function to evaluate a variety of model selection criteria. Each exercise will focus on a single criterion, but in the end you should submit a single function that calculates all these criteria. This function should have the following name and preamble:

```
modsel.loo.simple = function(y,x,kmax=dim(x)[2]+1) {
#### EVALUATES VARIOUS MODEL SELECTION CRITERIA FOR THE MODEL
#### Y = X B + E
#### NOTE: THE CONSTANT (INTERCEPT) SHOULD NOT BE INCLUDED IN X;
####   IT IS INSERTED BY THIS PROGRAM
#
# INPUT:
#   Y[N]: PREDICTAND, WHERE N = NUMBER OF INDEPENDENT SAMPLES
#   X[N,K]: PREDICTOR MATRIX, K = NUMBER OF PREDICTORS
#      (NOT INCLUDING CONSTANT TERM)
#   KMAX: MAXIMUM NUMBER OF PREDICTORS (INCLUDING THE CONSTANT)
#      (DEFAULT: ALL PREDICTORS)
# OUTPUT:
#   SE.SQR[KMAX]: UNBIASED ERROR VARIANCE OF EACH MODEL 1, 2, ..., KMAX
#   NSE.SQR[KMAX]: NORMALIZED UNBIASED ERROR VARIANCE OF EACH MODEL
#   CVMSE[KMAX]: LOO CROSS-VALIDATED MEAN SQUARE ERROR FOR EACH MODEL
#   CVSTD[KMAX]: STANDARD ERROR OF LOO CROSS-VALIDATED SQUARED ERRORS
#   AIC[KMAX],AICC[KMAX],BIC[KMAX]: INFORMATION CRITERIA FOR EACH MODEL
#   IC.STD[KMAX]: STANDARD ERROR OF THE INFORMATION CRITERIA
```

Exercise 10.1. We want to fit the regression model

$$y_n = \beta_0 + \sum_{h=1}^{H} \left(c_h \cos\left(2\pi n h/12\right) + s_h \sin\left(2\pi n h/12\right) \right) \tag{10.1}$$

where y_n is the NINO3.4 index and $n = 1, 2, \ldots, N$. Show that this model can be written in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{10.2}$$

What is $\mathbf{X}$? What is $\boldsymbol{\beta}$? Hint: in R, the predictor matrix containing the Fourier harmonics can be generated as

```
x.pred = NULL
for ( nh in 1:nharm) x.pred =
  cbind(x.pred,cos(2*pi*nh*nino34.time),sin(2*pi*nh*nino34.time))
```

□

Exercise 10.2. Write a function that computes the unbiased error variance of the regression model (10.1) as a function of the number of harmonics H. The unbiased error variance is

$$s_e^2 = \frac{SSE}{N - M}. \tag{10.3}$$

The input to this function is the time series $\mathbf{y}$ and the predictor matrix $\mathbf{X}$ *without the constant term*. Internally, you can insert the constant function as follows:

```
ntot    = length(y)
x.const = cbind(rep(1,ntot),x); # ADD COLUMN OF ONES IN PREDICTOR MATRIX
```

If k is the number of columns of `x.const`, then the following commands will fit the regression model and produce fitted values and residuals.

```
for ( k in 1:kmax) {
        x.pred      = x.const[,1:k]
        xy.lm       = lm(y~x.pred-1)
        yhat        = fitted(xy.lm)
        yresid      = residuals(xy.lm)
}
```

Add to this code by computing s_e^2 and the normalized error variance

$$\frac{s_e^2}{\hat{\sigma}_Y^2}, \tag{10.4}$$

where $\hat{\sigma}_Y^2$ is the sample variance of $\mathbf{y}$. Make a plot of the normalized error variance as a function of the number of predictors. Where does the minimum occur? How many Fourier harmonics does this correspond to? How much variance is explained by the annual cycle? How much variance is unexplained by the annual cycle?

□

Exercise 10.3. Write a function that computes the leave-one-out cross validated mean square error. Hint: use "negative" arguments in R arrays to leave data out. For instance:

```
yhat = numeric(ntot)
for ( n in 1:ntot) {
        y.pred  = y[-n]
        x.pred  = x.const[-n,1:k]
        yhat[n] = sum(coefficients(lm(y.pred~x.pred-1)) * x.const[n,1:k])
}
```

State the values of the leave-one-out cross validated mean square error (after normalizing by the variance of y). □

Exercise 10.4. The above approach to leave-one-out cross validation is computationally intensive. A faster numerical method is explained in exercise 11.2. You do not have to prove exercise 11.2, but you will use the result. Write a function that computes the leave-one-out cross validated mean square error based on the numerical approach discussed in exercise 11.2. Verify that the result is identical to what you found in the previous exercise. Discuss how much faster the new function is relative to the old.

□

Exercise 10.5. Compute the standard error of the cross validated mean square error. Indicate the standard error as error bars on the figure produced in the previous problem. Error bars can be plotted using the `arrows` command as follows:

```
modsel.list = modsel.loo.simple(nino34.ts,x.pred)
x0 = 0:(2*nharm)
y0 = modsel.list$cvmse - modsel.list$cvstd
y1 = modsel.list$cvmse + modsel.list$cvstd
yrange = range(y0,y1)
plot(x0,modsel.list$cvmse,type='b',pch=19,xlab='number of Fourier terms',
  ylab='unbiased error variance',ylim=yrange)
arrows(x0,y0,x0,y1,length=0.1,angle=90,code=3,lwd=2)
```

What model does the "one-standard-deviation rule" select?

□

Exercise 10.6. Write a function to evaluate AIC, BIC, and AICC. Plot these values as a function of the number of predictors in the annual cycle model. State the numerical values of these quantities. What model does these criteria select?

□

Exercise 10.7. Write a function that computes the standard error of the information criteria. Show the standard error in the plot generated in the previous exercise. What model does the "one-standard-error" rule select? □

Exercise 10.8. Based on all of the above results, you should have found that either 1 or 2 harmonics should be used to fit the annual cycle of NINO3.4. Let us decide to use 2 harmonics (so this means you have 5 predictors: 1 intercept term, cos/sine for the first harmonic, and cos/sine for the second harmonic). Now subtract this annual cycle from the NINO3.4 index and plot the residual. This is called the NINO3.4 *anomaly*. You should recognize the 1982, 1998, and 2015 El Niño events.

□

Exercise 10.9 (Apparently Different Information Criteria)**.** Different definitions of AIC, AICC, BIC appear in the literature. However, these definitions differ by either a factor of N or an additive constant, neither of which affect the *location* of the minimum value. For example, some papers define AIC_c as

$$AIC_c = N \log\left(\frac{SSE}{N}\right) + \frac{2KN}{N-K-1}, \tag{10.5}$$

where K is the total number of "parameters" in the regression model *including the error variance* σ_ϵ^2; that is, $K = M + 1$ in terms of the model defined in (**??**). Show that this definition differs from the AIC_c defined (10.28) by a constant and therefore the two expressions (10.28) and (10.5) are equivalent model selection criteria. □

Exercise 10.10 (Shortcut for Leave-One-Out Cross Validation)**.** Show that the leave-one-out cross validated error of the regression model (**??**) can be written as

$$\epsilon_k^{LOO} = y_k - \mathbf{x}_k \boldsymbol{\beta}_k = \frac{y_k - \mathbf{x}_k \hat{\boldsymbol{\beta}}}{1 - \alpha_k}, \tag{10.6}$$

where $\hat{\boldsymbol{\beta}}$ is the least-squares estimate of $\boldsymbol{\beta}$ using all data and α_k is the scalar

$$\alpha_k = \mathbf{x}_k \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_k^T. \tag{10.7}$$

Interestingly, the cross validated error $y_n - \mathbf{x}_n \boldsymbol{\beta}_n$ is merely an *inflated* version of the residual of the least squares model derived from the entire data set $y_n - \mathbf{x}_n \hat{\boldsymbol{\beta}}$. Equation (10.6) also shows that the leave-one-out cross validated mean square error can be determined from the traditional least squares solution, without re-fitting the model N separate times. Although the method requires computing the inverse of $\mathbf{X}^T\mathbf{X}$ once, this inverse is already available since it is required to compute $\hat{\boldsymbol{\beta}}$. Hint: note that

$$\boldsymbol{\beta}_k = \left(\mathbf{X}^T\mathbf{X} - \mathbf{x}_k^T\mathbf{x}_k\right)^{-1}\left(\mathbf{X}^T\mathbf{y} - \mathbf{x}_k^T y_k\right), \tag{10.8}$$

and apply the Sherman-Morrison-Woodbury formula.

□