CHAPTER 14

EXERCISES FOR MULTIVARIATE LINEAR REGRESSION

Exercise 14.1. It is natural to wonder whether a "better" estimate of $\mathbf{B}$ can be obtained by weighting the SSE toward certain variables. To explore this possibility, suppose w_s is the weight for the s'th spatial location. Then, the SSE in (14.20) can be modified as

$$SSE = \sum_{s=1}^{S} \sum_{n=1}^{N} w_s \left(Y_{ns} - \sum_{m=1}^{M} X_{nm} B_{ms} \right)^2 . \tag{14.1}$$

More generally, one can specify a positive-definite weight matrix $\mathbf{W}$ and define

$$SSE = \sum_{s'=1}^{S} \sum_{s=1}^{S} \sum_{n=1}^{N} W_{ss'} \left(Y_{ns} - \sum_{m=1}^{M} X_{nm} B_{ms} \right) \left(Y_{ns'} - \sum_{m=1}^{M} X_{nm} B_{ms'} \right), \tag{14.2}$$

which can be written equivalently as

$$SSE_W = \mathrm{tr} \left[(\mathbf{Y} - \mathbf{XB}) \, \mathbf{W} \, (\mathbf{Y} - \mathbf{XB})^T \right]. \tag{14.3}$$

Show that the matrix $\mathbf{B}$ that minimizes SSE_W is independent of $\mathbf{W}$ and equal to the least squares estimate (14.25). This means that there is nothing to gain by weighting different spatial locations differently. Explain why this is so. □

Computational Exercise for Multivariate Linear Regression

In this homework, you will derive a multivariate linear prediction model for Sea Surface Temperature.

Getting the Data The data set will be the ERSSTv5,[2] which is on the COLA servers.[3] However, you might prefer to download the data set directly on your laptop (I would recommend this). To download this data using R, grab the following files on the class website:

```
download.ersstv5.R
download_ftp_file.R
create_directory.R
```

Then type `source('download.ersstv5.R')`. The code will take a few minutes to run and will show the file names being downloaded. It also will generate a warning message about creating directories, which can be ignored.

Reading the Data To make this assignment easier, I have written an R code called `Rcode.exercise.Chapter14.R` that does the following things:

1. read the NetCDF files for ERSSTv5
2. load the data into the array `sst.full[space,month,year]`
3. regress out the mean and linear trend from each grid point
4. define the Pacific Ocean domain to be between $30°$S and $60°$N
5. compute the EOFs of monthly SST in the prescribed domain

This code also requires the following files from the class website:

```
eof.latlon.R
mask.define.R
```

It is good practice to put all of your data in a single directory (say "data"), and all your auxiliary programs into a single directory that can be referred to in future R programs. You should change `dir.Rlib` to point to the directory of your R functions, and `dir.ersst` to point to the directory of your ERSSTv5 data.

After running `Rcode.exercise.Chapter14.R`, you should see a plot of the leading EOF and PC. If the code crashes or gives an error message, please see me. Otherwise, you are ready! After the code runs to completion, you should have the following data arrays available.

[2] https://www.ncdc.noaa.gov/data-access/marineocean-data/extended-reconstructed-sea-surface-temperature-ersst-v5

[3] /shared/obs/sst/ncdc/ersst-v5

```
# 1) THE GRIDDED DATA SET IS IN SST.FULL[SPACE,MONTH,YEAR]
# 2) EOFS ARE IN EOF.LIST
#   (E.G., EOF.LIST$EOF[SPACE,NEOF], EOF.LIST$PC[TIME,NEOF])
# 3) LAT/LON IS LON.SST, LAT.SST
# 4) MON.INIT IS THE CALENDAR MONTH OF THE INITAL CONDITION
# 5) MON.TARG IS THE CALENDAR MONTH OF THE TARGET
```

Temperature over the ocean is set to `NA`. Also, temperature is read as an array with dimension `nlon, nlat, time`, but eventually is "reshaped" into an array with dimension `nlon*nlat, 12, nyrs`. If you do not understand what this means, please see me!

Exercise 14.2. Write a R function to compute MIC given two data arrays. The function also should compute the 5% significance threshold for MIC. The preamble should be as follows

```
mic.gaussian = function(x,y,equal.dim=TRUE,alpha=0.05) {
### THIS FUNCTION COMPUTES CORRECTED MUTUAL INFORMATION CRITERION (MIC)
###     FOR Y = XB + E
### INPUT:
##    X[NSAMP,XDIM]: X DATA ARRAY, OFTEN FORMATTED AS [TIME,EOF]
##    Y[NSAMP,YDIM]: Y DATA ARRAY, OFTEN FORMATTED AS [TIME,EOF]
##    EQUAL.DIM: LOGICAL INDICATING WHETHER
##      (TRUE) EQUAL NUMBER OF X'S, Y'S CHOSEN: MIC[MIN(XDIM,YDIM)];
##      (FALSE) MIC FOR ALL TRUNCATIONS ARE COMPUTED: MIC[XDIM,YDIM]
### OUTPUT: LIST (DIMENSIONS DEPEND ON EQUAL.DIM)
##    $MIC: MIC VALUES
##    $PENALTY: THE PENALTY TERM IN MIC
##    $CRIT: SIGNIFICANCE THRESHOLD OF MIC
```

□

Exercise 14.3. Consider the case of predicting December SST based on June SST. Calculate the MIC for EOF truncations at least for 1-25. The MIC values should be the following:

```
> micc
 [1]   -0.93434963   -1.23487398   -1.61938679   -1.67394748
 [5]   -1.84103698   -1.50387241   -1.19900423   -0.69182271
 [9]    0.05407005    0.33422799    1.48966123    2.50855073
[13]    4.09619446    5.50819524    7.95854904   11.03580560
[17]   14.07453324   18.15159799   23.34799814   30.02343607
[21]   38.16355309   48.57419513   62.29384427   79.60195564
[25]  101.97693930  131.36865590  172.06833168  227.00353490
[29]  312.44543534  452.13622542  710.10017386 1323.74697742
[33] 4435.93164492            NA
```

Plot the MIC values and indicate the minimum, which should be 5 EOFs. Next, do the same thing, except for predicting December SST based on September initial condition. Print the values of MIC (as above). Decide how many EOFs should be used and state your answer. □

Exercise 14.4. Use multivariate regression to make a prediction of December SSTs based on June SSTs using 5 EOFs. Submit your code for doing this. Make sure your results are consistent with the notes. Then, make a prediction for December SSTs using September SSTs and 5 EOFs. Submit plots of your predictions. □