CHAPTER 15

EXERCISES FOR CANONICAL CORRELATION ANALYSIS

`Rcode.exercise.Chapter15.R`	main program
`cca.data.clim763.DJF.RData`	SST and N. American temperature for DJF
`eof.latlon.R`	auxiliary file: calculate EOFs
`plot_latlon_v4.R`	auxiliary file: plot spatial maps

In this homework you will write an R function that performs Canonical Correlation Analysis. You will need to download data and a few R programs. These files are summarized in the above table. The R code `Rcode.exercise.Chapter15.R` reads the data file `cca.data.clim763.DJF.RData` and computes the EOFs of SST and U.S. temperature, and plots out the leading EOF and PC time series. In a previous homework you investigated the relation between ENSO and U.S. temperature using the concept of field significance. However, that analysis assumed you knew the NINO3.4 index. Here, you will *derive* an index for predicting U.S. temperature based on SST.

The following exercises break up CCA into discrete steps. However, in the end, you should submit a single function that performs all the calculations. The preamble of this function should be the following:

```
cca.pca = function(x.pca,y.pca,tx=NULL,ty=NULL) {
#########################################################
## PERFORMS CANONICAL CORRELATION ANALYSIS ON X AND Y.
## X AND Y ARE ASSUMED TO BE IN THE FOLLOWING FORMS:
## X = FX %*% EXˆT   ### AND ### Y = FY %*% EYˆT
## WHERE FX AND FY HAVE COVARIANCE MATRICES = I
## FOR EXAMPLE: FROM PRINCIPAL COMPONENT ANALYSIS
## IF TX OR TY = NULL, THEN BOTH (TX,TY) SELECTED USING MIC
## INPUT:
#    X.PCA: LIST OUTPUT FROM EOF.LATLON[X-DATA]
#    Y.PCA: LIST OUTPUT FROM EOF.LATLON[Y-DATA]
#    TX: TRUNCATION FOR X (TX <= MX); IF NULL, TX IS SELECTED
#    TY: TRUNCATION FOR Y (TY <= MY); IF NULL, TY IS SELECTED
## OUTPUT LIST:
#    MIC[MX,MY]: MUTUAL INFORMATION CRITERION
#    NMIN[1,2]: VALUES OF TX,TY THAT MINIMIZES MIC
#    CAN.COR[MIN(TX,TY)]: CANONICAL CORRELATIONS
#    RX[NTOT,MIN(TX,TY)]: CANONICAL VARIATES FOR X
#    RY[NTOT,MIN(TX,TY)]: CANONICAL VARIATES FOR Y
#    PX[SX  ,MIN(TX,TY)]: CANONICAL LOADING VECTORS FOR X
#    PY[SY  ,MIN(TX,TY)]: CANONICAL LOADING VECTORS FOR Y
#    QX.TILDE[TX,MIN(TX,TY)]: WEIGHTING VECTORS FOR X-FEATURES
#    QY.TILDE[TY,MIN(TX,TY)]: WEIGHTING VECTORS FOR Y-FEATURES
#    TX, TY: SELECTED VALUES OF TX AND TY
#########################################################
```

Following the notes, the data sets are assumed to have been decomposed into the form

$$\underset{N \times S_X}{\mathbf{X}} = \underset{N \times M_X}{\mathbf{F}_X} \quad \underset{M_X \times S_X,}{\mathbf{E}_X^T} \tag{15.1}$$

and

$$\underset{N \times S_Y}{\mathbf{Y}} = \underset{N \times M_Y}{\mathbf{F}_Y} \quad \underset{M_Y \times S_Y,}{\mathbf{E}_Y^T} \tag{15.2}$$

where

$$\frac{1}{N-1}\mathbf{F}_X^T\mathbf{F}_X = \mathbf{I} \quad \text{and} \quad \frac{1}{N-1}\mathbf{F}_Y^T\mathbf{F}_Y = \mathbf{I}. \tag{15.3}$$

This decomposition is accomplished in `Rcode.exercise.Chapter15.R` using principal component analysis. You will be performing CCA on the time series matrices $\mathbf{F}_X$ and $\mathbf{F}_Y$.

Exercise 15.1. In a previous homework you wrote a function that computed Mutual Information Criterion (MIC). Augment that function to compute MIC for all truncations T_X and T_Y, up to some maximum number (what *is* the maximum number?). As a reminder, MIC is defined as

$$MIC = \sum_i \log(1 - \hat{\rho}_i^2) + \mathbb{P}, \tag{15.4}$$

where $\hat{\rho}_i$ is the i'th sample canonical correlation and

$$\mathbb{P} = (N+1)\left(\frac{T_X + T_Y}{N - T_X - T_Y - 2} - \frac{T_X}{N - T_X - 2} - \frac{T_Y}{N - T_Y - 2}\right). \tag{15.5}$$

Plot MIC for a range of values and identify the value of T_X and T_Y that minimizes MIC. Print out the values for the first 5 rows and columns (i.e., print `mic[1:5,1:5]`). These results should match the example in the notes. Where is the minimum value, and what is it? The location of the minimum can be found using the following R commands:

```
nmin    = which(mic == min(micc,na.rm=TRUE),arr.ind=TRUE)
tx      = nmin[1]
ty      = nmin[2]
```

□

Exercise 15.2. Compute all canonical correlations for the optimum choice of T_X and T_Y. State the values. These values should be consistent with those in the notes. □

Exercise 15.3. Write a function that computes canonical variates. Compute the leading canonical variates between SST and U.S. temperature using the same truncation parameters as above. Make a plot that shows the leading canonical variate for the two data sets. Verify that the sample covariance matrices of the canonical variates equals the identity matrix. The sample covariance matrix of `rx` can be obtained using the R command `cov(rx)`. □

Exercise 15.4. Write a function that computes canonical loading vectors. Compute the leading canonical loadings between SST and U.S. temperature using the same truncation parameters as above. Make a plot that shows the leading canonical loadings for the two data sets. □

Exercise 15.5. Write a function to compute the fraction of variance explained by the canonical component. This is somewhat tricky because (1) the area weighting needs to be included, (2) missing data should be skipped, and (3) the total variance needs to be computed to compute the fraction. All of this information is available from the output of `eof.latlon`. To help you out, here is the way to compute it for the EOFs of SST:

```
var.x.tot = sum(sst.eof$sval^2)/(nyrs-1)
exp.var.x = rep(NA,dim=tx)
for ( n in 1:tx) exp.var.x[n] = sum(px[!sst.eof$lbad,n]^2 *
   sst.eof$weight[!sst.eof$lbad]^2)/var.x.tot
```

Verify that the sum of the fractional variances equals the sum of the fractional variances of the first T_X EOFs `sum(sst.eof$fexpvar[1:tx])`. You might have to compute *all* of the singular vectors using the command `svd(cov.mat,nu=tx,nv=ty)`. State the explained variances of the canonical components for SST and for U.S. Temperature. These variances should match those in the notes. □

Exercise 15.6. Write a *separate* code that computes the 5% significance levels of the canonical correlations based on 5000 trials of Monte Carlo experiments. State the critical values for each canonical correlation based on the optimum choice of T_X and T_Y. Use these results to decide whether the canonical correlations computed from data are significant. Clearly state your conclusion. (Coding advice: first use only 100 trials until you have debugged your code, and then change to 5000 trials when the code is working.) □

Exercise 15.7. Repeat the above steps, except this time *do not detrend the data*. To prevent detrending, set `npoly =0` at the top of the code. How do the results differ? Explain why this makes sense. □

Exercise 15.8. Show mathematically that the error covariance matrix for the model (15.80) is

$$\tilde{\mathbf{\Sigma}}_{\epsilon Y} = \mathbf{I} - \hat{\mathbf{S}}_{\rho}^{2}. \tag{15.6}$$

This result shows that the error covariance matrix is diagonal, and that the diagonal elements equal $1 - \hat{\rho}_k^2$, which can be interpreted as the variance explained by the canonical variate. □