CHAPTER 16

EXERCISES: COVARIANCE DISCRIMINANT ANALYSIS

Exercise 16.1. Prove (16.14). Hint, substitute (16.12) into the outer product and expanding gives

$$(\mathbf{x}_t - \hat{\boldsymbol{\mu}}_Y)(\mathbf{x}_t - \hat{\boldsymbol{\mu}}_Y)^T = \mathbf{A} + \mathbf{B} + \mathbf{C}, \tag{16.1}$$

where

$$\mathbf{A} = (\mathbf{u}_t^* - \hat{\boldsymbol{\mu}}_Y)(\mathbf{u}_t^* - \hat{\boldsymbol{\mu}}_Y)^T \tag{16.2}$$

$$\mathbf{B} = \mathbf{f}_t \mathbf{f}_t^T \tag{16.3}$$

$$\mathbf{C} = (\mathbf{u}_t^* - \hat{\boldsymbol{\mu}}_Y)\mathbf{f}_t^T + \mathbf{f}_t(\mathbf{u}_t^* - \hat{\boldsymbol{\mu}}_Y)^T. \tag{16.4}$$

Show that $\mathbb{E}[\mathbf{C}] = \mathbf{0}$. What is $\mathbb{E}[\mathbf{A}]$? □

Exercise 16.2. Suppose $d(\sigma_X^2, \sigma_Y^2)$ is some measure of the difference between the variances σ_X^2, σ_Y^2. Suppose further that the function $d(,)$ is invariant to an invertible linear transformation of X and Y. Prove that $d(,)$ can depend only on the ratio of variances

$$d(\sigma_X^2, \sigma_Y^2) = f\left(\frac{\sigma_X^2}{\sigma_Y^2}\right) \tag{16.5}$$

□

Exercise 16.3 (Discriminant analysis via SVD)**.** Given two data matrices $\mathbf{X}$ and $\mathbf{Y}$, solve discriminant analysis based on the singular value decomposition (SVD) (i.e., without solv-

ing an eigenvalue problem). Show how the discriminant ratios, variates, and loading vectors can be derived from the results of the SVD. Hint: compute the SVD of $\mathbf{Y}$ to construct a whitening transformation, then compute the SVD of the whitened data matrix $\mathbf{X}$. □

Exercise 16.4 (Loading Patterns)**.** Show that the matrix $\mathbf{P}$ that minimizes

$$\gamma_Y = E\left[\|\mathbf{y} - E[\mathbf{y}] - \mathbf{P}\mathbf{r}_Y\|_W^2\right], \tag{16.6}$$

is

$$\mathbf{P} = \mathrm{cov}[\mathbf{y}, \mathbf{r}_Y]\left(\mathrm{cov}[\mathbf{r}_Y, \mathbf{r}_Y]\right)^{-1}, \tag{16.7}$$

thereby proving (16.64). Also, show that the matrix $\mathbf{P}$ that minimizes

$$\dot{\gamma}_Y = \|\dot{\mathbf{Y}} - \mathbf{R}_Y\mathbf{P}^T\|_W^2. \tag{16.8}$$

is

$$\dot{\mathbf{P}} = \dot{\mathbf{Y}}^T\mathbf{R}_Y\left(\mathbf{R}_Y^T\mathbf{R}_Y\right)^{-1}, \tag{16.9}$$

thereby proving (16.94). Note that $\dot{\mathbf{P}}$ and $\mathbf{P}$ are merely the sample and population versions of each other.

□

Exercise 16.5. Prove that if $\mathbf{\Sigma}_X \neq \mathbf{\Sigma}_Y$, then there exists a $\mathbf{q}$ such that $\lambda \neq 1$, where

$$\lambda = \frac{\mathbf{q}^T\mathbf{\Sigma}_X\mathbf{q}}{\mathbf{q}^T\mathbf{\Sigma}_Y\mathbf{q}}. \tag{16.10}$$

□

Exercise 16.6. Define the sum total variance of $\dot{\mathbf{Y}}$ as

$$\|\dot{\mathbf{Y}}\|_W^2 = \frac{1}{N_Y - 1}\,\mathrm{tr}\left[\dot{\mathbf{Y}}\mathbf{W}\dot{\mathbf{Y}}^T\right], \tag{16.11}$$

where $\mathbf{W}$ is a positive definite matrix defining how different points are weighted. If $\dot{\mathbf{Y}} = \mathbf{R}_Y\dot{\mathbf{P}}_Y^T$, use the properties of CDA to show explicitly that the sum total variance can be written as

$$\|\dot{\mathbf{Y}}\|_W^2 = \sum_{k=1}^{T}\mathbf{p}_k^T\mathbf{W}\mathbf{p}_k. \tag{16.12}$$

□

Numerical Exercises

`Rcode.exercise.Chapter16.CCSM4.R`	program for reading the data sets
tas_Amon_CCSM4_historical_r1i1p1_185001-200512.nc	20C, 1st member
tas_Amon_CCSM4_historical_r2i1p1_185001-200512.nc	20C, 2nd member
tas_Amon_CCSM4_piControl_r1i1p1_080001-130012.nc	PI simulation

In this homework you will write an R function to perform covariance discriminant 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.Chapter16.CCSM4.R` reads data files for the 20C and PI simulations, combines them, and computes EOFs. The following exercises break up the discriminant analysis into discrete steps, but you should submit a single function that performs all the calculations. The preamble of this function should be the following:

```
cda.eof = function(xdata,ydata,eof.list) {
### PERFORMS COVARIANCE DISCRIMINANT ANALYSIS ON X AND Y
### INPUT:
###   XDATA[NX,MDIM]
###   YDATA[NY,MDIM]
###   EOF.LIST: LIST FROM EOF CALCULATION
### OUTPUT:
###   MIC[NEOF]: MIC AS A FUNCTION OF NUMBER OF PCS
###   NMIN: LOCATION OF MINIMUM MIC
###   DISCR.RATIO[NEOF]: DISCRIMINANT RATIOS VS. NUMBER OF PCS
###   RX[NX,NMIN]: VARIATE TIME SERIES FOR X
###   RY[NY,NMIN]: VARIATE TIME SERIES FOR Y
###   PMAT[SPACE,NMIN]: LOADING VECTOR
```

In the following calculations, you should include *both* ensemble members from 20C. A trick for doing this is to reshape the array so that the time series looks twice as long:

```
### RESHAPE PC.20C[TIME,NENS,NEOF] TO PC.20C[TIME*NENS,NEOF]
dim(pc.20c) = c(tdim.20c*nens,neof)
```

Then, when you want individual ensemble members, you can reshape the array back to [time,ensemble,eof]

Exercise 16.7. Write a function that evaluates Mutual Information Criterion (MIC) for comparing covariance matrices. MIC is defined as

$$\text{MIC} = \frac{1}{N_T} \log \left(\frac{|\overline{\mathbf{\Sigma}}_X|^{N_X} |\overline{\mathbf{\Sigma}}_Y|^{N_Y}}{|\overline{\mathbf{\Sigma}}_T|^{N_T}} \right) + \mathbb{P}, \tag{16.13}$$

where

$$\mathbb{P} = \frac{T}{N_T} \left(\frac{N_X(N_X+1)}{N_X - P - 2} + \frac{N_Y(N_Y+1)}{N_Y - P - 2} - \frac{N_T(N_T+1)}{N_T - P - 2} \right),$$

and

$$\hat{\mathbf{\Sigma}}_T = \frac{N_X \overline{\mathbf{\Sigma}}_X + N_Y \overline{\mathbf{\Sigma}}_Y}{N_T} \quad \text{and} \quad N_T = N_X + N_Y. \tag{16.14}$$

The function should evaluate MIC over all possible EOF truncations T. Plot MIC for a range of values and identify the value of T that minimizes MIC. Print out the first 5 values of MIC. These results should match the example in the notes. □

Exercise 16.8. Compute the discriminant ratios for the optimum choice of T. State the values. These values should be consistent with those in the notes □

Exercise 16.9. Write a function that computes discriminant variates. Compute the variates for the optimum choice of T and plot them. Verify that the sample covariance matrix of the PI variates equals the identity matrix. Verify that the sample covariance matrix of 20C variates is diagonal, with diagonal elements equal to the discriminant ratios. □

Exercise 16.10. Write a function that computes the loading vectors. Plot the leading loading vector. □

Exercise 16.11. Write a *separate* code that computes the 5% significance levels of the discriminant ratios based on 5000 trials of Monte Carlo experiments. State the 95% percentile for all discriminant ratios for the optimum choice of T. Are your discriminant ratios significant or not? □

Exercise 16.12. What is the 5% significance level of the *univariate* F-test for equality of variances for sample sizes $N_X = 51$, $N_Y = 51$? Using your Monte Carlo code, show a plot of the 5% and 95% percentiles of *the leading discriminant ratio* as a function of the truncation parameter T. In this exercise, let $N_X = 51$, $N_Y = 51$, and the number of trials = 1000. Also, let the maximum dimension be 30. What happens to the ratios as T increases? Explain why this happens. The 95% percentile from the Monte Carlo experiments should be close to the univariate F-test for $T = 1$. □