STAT 185 - Problem Set #3

NAME

This assignment covers Singular Value Decomposition and Canonical Correlation Analysis through a mixture of theory and coding. You will need the file *pset3.Rdata* available on Canvas. The commands for loading the data assume this file, *Homework3.Rmd*, and the Rdata file are in the same folder.

You can use the commands *svd(A)* to calculate the reduced SVD of matrix A and *cc(X,Y)* for the CCA of data matrices X and Y. Note *out* $\leftarrow$ *svd(A)* stores the left singular vectors (as columns) in *out$u*, the right singular vectors (as columns) in *out$v*, and the singular values (as a vector) in *out$d*.

Using the command *out* $\leftarrow$ *cc(X,Y)* for CCA saves the correlations in *out$cor*, the canonical correlation vectors in *out$xcoef* and *out$ycoef*, and the canonical correlation variables in *out$scores*.

Complete the following questions and submit your solutions as a *pdf* via Canvas by 11:59 PM on Tuesday October 8th.

1. Let $\vec{x}$ and $\vec{y}$ be random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively. Assume $\vec{x}$ and $\vec{y}$ have means μ_x and μ_y and covariances $\Sigma_x \in \mathbb{R}^{p\times p}$ and $\Sigma_y \in \mathbb{R}^{q\times q}$ respectively and cross-covariance

$$\Sigma_{XY} = E[(\vec{x}-\mu_x)(\vec{y}-\mu_y)^T] \in \mathbb{R}^{p\times q}.$$

   We assume that both Σ_X and Σ_Y are invertible.

   a. Show that for fixed, nonzero $\vec{b} \in \mathbb{R}^q$

$$\max_{\vec{a}\in\mathbb{R}^d} \frac{(\vec{a}^T\Sigma_{XY}\vec{b})^2}{(\vec{a}^T\Sigma_X\vec{a})(\vec{b}^T\Sigma_Y\vec{b})} = \frac{\vec{b}^T\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\vec{b}}{\vec{b}^T\Sigma_Y\vec{b}}$$

   where $\Sigma_{YX} = \Sigma_{XY}^T$. (Hint: for any two vectors $\vec{u}$ and $\vec{w}$, recall $(\vec{u}^T\vec{w})^2 = \|\vec{u}\|^2\|\vec{v}\|^2\cos^2\theta$ where θ is the angle between $\vec{u}$ and $\vec{v}$.)

   b. Show that the maximum of

$$\frac{\vec{b}^T\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\vec{b}}{\vec{b}^T\Sigma_Y\vec{b}}$$

   over all $\vec{b} \in \mathbb{R}^q$ is given by the largest eigenvalue of $\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}$.

2. Suppose that $\vec{x} = (x_1, x_2)^T$ and $\vec{y} = (y_1, y_2)^T$ are $\mathcal{N}(\vec{0}, I)$ random vectors such that

$$x_1 + x_2 = y_1 + y_2.$$

   a. Find the cross-covariance of $\vec{x}$ and $\vec{y}$.

   b. Find the canonical correlations and canonical correlation vectors $\vec{x}$ and $\vec{y}$.

3. Let $\vec{x}$ and $\vec{y}$ be vectors in $\mathbb{R}^2$. Suppose

$$\Sigma_X = \Sigma_Y = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \text{ and } \Sigma_{XY} = \begin{bmatrix} \rho & \rho \\ \rho & \rho \end{bmatrix}$$

   for some $\rho \in (-1, 1)$. Find the canonical correlation vectors and coefficients.

4. Suppose that random vectors $\vec{x} \in \mathbb{R}^d$ and $\vec{y} \in \vec{R}^d$ have canonical correlations $\rho_1, \ldots, \rho_d$.

a. If $A, B \in \mathbb{R}^{d\times d}$ are both invertible matrices and $\vec{v}$ and $\vec{w}$ are constant vectors in $\mathbb{R}^d$ respectively, show that the canonical correlations of $A\vec{x}+\vec{u}$ and $B\vec{y}+\vec{w}$ are still $\rho_1, \ldots, \rho_d$. (Hint: you may find it more straightforward to investigate the eigenvalues of $\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}$ and the same expression for the transformed variables.)

b. Suppose that $\vec{x}$ and $\vec{y}$ are standardized so that each of their coordinates are mean zero and variance 1. Using part (a), show that the canonical correlation cofficients are unchanged.

5. Suppose that $\vec{x} \in \mathbb{R}^p$ has mean $\vec{0}$ and covariance I. Let $A \in \mathbb{R}^{q\times p}$ be a full rank matrix with $p > q$ and let $\vec{y} = A\vec{x}$ be a vector in $\mathbb{R}^q$.

   a. Find Σ_y and Σ_{xy}.

   b. Find the canonical correlations of $\vec{x}$ and $\vec{y}$.

   c. Let A be a 5×10 matrix such that

$$A_{ij} = \begin{cases} 1 & i \leq j \\ 0 & i > j. \end{cases}$$

   This matrix is included in *pset3.Rdata*. Generate 1000 independent samples, $\vec{x}_1, \ldots, \vec{x}_{1000}$ from a $\mathcal{N}(\vec{0}_{10}, I_{10})$ distribution and set $\vec{y}_i = A\vec{x}_i$. Compute the canonical correlations and print them to verify your answer from part (b).

6. Let A and $\vec{x}$ be as in problem 5. Now let $\vec{y} = A\vec{x} + \vec{\epsilon}$ where $\vec{\epsilon} \sim \mathcal{N}(0, \rho^2 I_q)$ is independent of $\vec{x}$.

   a. Find Σ_y and Σ_{xy}.

   b. Find the canonical correlations between $\vec{x}$ and $\vec{y}$. Express your answer in terms of ρ^2 and the singular values of A.

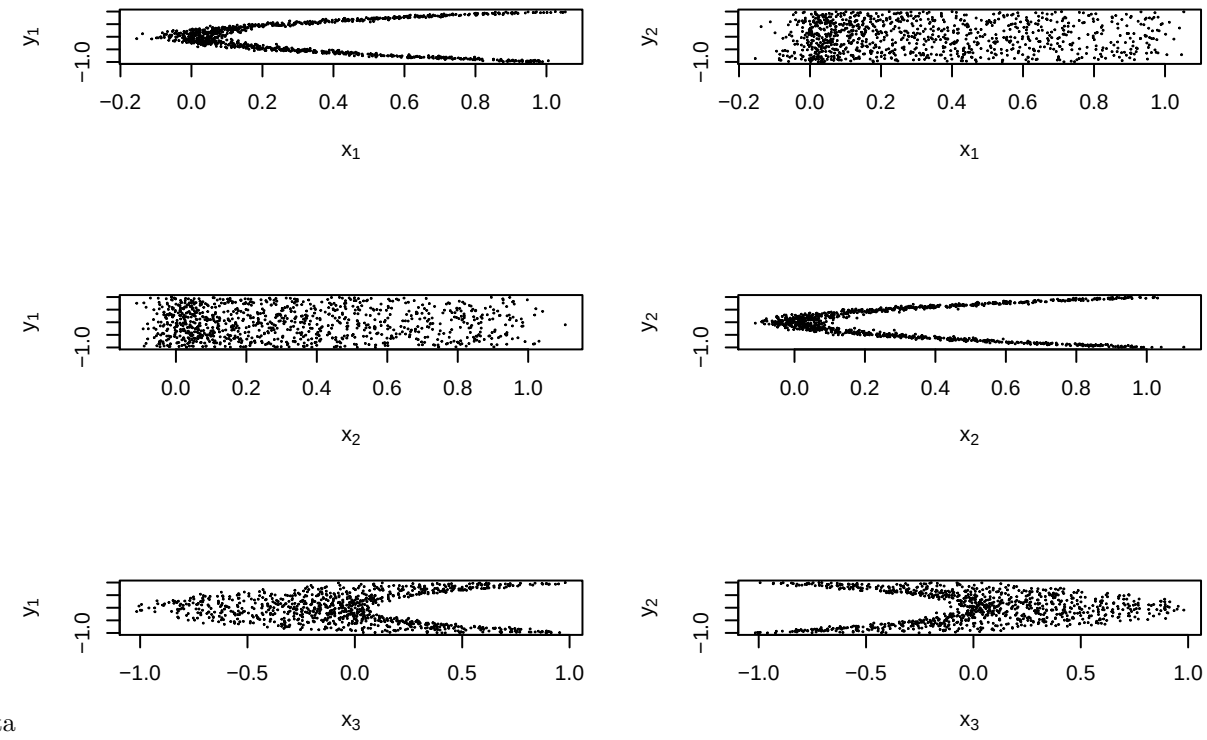
   c. Suppose A is the same as specified in 5c. Suppose 1000 independent vectors $\vec{x}_1, \ldots, \vec{x}_{1000}$ were generated from a $\mathcal{N}(\vec{0}, I)$ distribution and saved in the data matrix *problem6.X*. The vectors $\vec{y}_i = A\vec{x}_i + \vec{\epsilon}_i$ were then generated and saved in the data matrix *problem6.Y*. Compute the sample canonical correlations and use your answer from part (b) to estimate ρ^2.

7. Suppose $\vec{x} \in \mathbb{R}^p$ and $\vec{y} \in \mathbb{R}^q$ are random vectors with covariance matrices Σ_x and Σ_y respectively. Suppose $\vec{x}$ and $\vec{y}$ have cross-covariance matrix Σ_{xy} with rank $k < \min\{p, q\}$ and canonical correlation variable $\eta_1, \ldots, \eta_k$ and $\xi_1, \ldots, \xi_k$ with canonical correlations $\sigma_1 \geq \cdots \geq \sigma_k > 0$.

   a. Find the covariance matrices of the random vectors $\vec{\eta} = (\eta_1, \ldots, \eta_k)^T$ and $\vec{\xi} = (\xi_1, \ldots, \xi_k)^T$.

   b. Find the cross-covariance of $\vec{\eta}$ and $\vec{\xi}$.

8. Consider the 1000×3 data matrix *problem8.X* and the 1000×2 data matrix *problem8.Y* where the rows of *problem8.X* and *problem8.Y* are paired observations of random variables $\vec{x}, \vec{y}$ drawn from the joint distribution $F(\vec{x}, \vec{y})$. In the following figures, we show coordinate-wise scatterplots, $\{(x_{ij}, y_{ik})\}_{i=1}^{1000}$, of these



data

a. There an apparent functional relationship between the coordinates of X and those of Y. Based on these relationships, do you expect the canonical correlations between X and Y to be large (near 1) or small (near 0)? Explain your reasoning.

b. Compute and print canonical correlations between X and Y. Compare these results to your predictions from part (a).

9. The COMBO-17 galaxy data is a publicly available subset of data obtained by Wolf et al. (2004) of the Chadra Deep Field South field which have been classified as galaxies. We can import these data containing 3462 observations of 65 variables from the Penn State Astrostatistics group.

```
galaxy <- read.csv(url("https://astrostatistics.psu.edu/datasets/COMBO17.csv"))
```

In the Izenman text (Chapter 7), the author conducts CCA on these data. More details on the variables can be found in that text as well. The following code removes redundant variables and partitions the galaxy data into x and y variables following the analysis in the Izenman text.

```
problem9.X <- galaxy[,c("UjMAG","BjMAG","VjMAG","usMAG","gsMAG","rsMAG",
                        "UbMAG","BbMAG","VnMAG","S280MAG","W420FE","W462FE",
                        "W485FD","W518FE","W571FS","W604FE","W646FD","W696FE",
                        "W753FE","W815FS","W856FD","W914FD","W914FE")]
problem9.Y <- galaxy[,c("Rmag","ApDRmag","mumax","Mcz","MCzml","chi2red")]
```

a. Compute and print the canonical correlations of *problem9.X* and *problem9.Y*.

b. Generate scatterplots of the first 6 canonical correlation variables $\{(\eta_{ij}, \xi_{ij})\}_{i=1}^{3462}$ for $j = 1, \ldots, 6$. See Figure 7.8 of Izenman for reference. Note, the notation therein for the canonical variables in different from the notation introduced in class.

c. Compute the sample covariance matrices of the vectors of the canonical correlation variables $\vec{\eta} = (\eta_1, \ldots, \eta_6)^T$ and $\vec{\xi} = (\xi_1, \ldots, \xi_6)^T$. Round to 2 decimal places. Do these matrices agree with your analysis from problem 7?

d. Compute the sample cross-covariance matrix of the vector of canonical correlation variables. Round

to 2 decimal places. Do this matrix agree with your analysis from problem 7?