FINM 331: DATA ANALYSIS FOR FINANCE AND STATISTICS
FALL 2015
PROBLEM SET 3

The required files for all problems can be found in:

`http://www.stat.uchicago.edu/~lekheng/courses/331/hw3/`

The file name indicates which problem the file is for (`p1*.txt` for Problem 1, etc). You are welcomed to use any programming language or software packages you like.

1. (*Factor Analysis*) This is the same air quality data set we saw in the previous problem set but this time we will only take four variables X_1, X_2, X_5 and X_6 by leaving out `CO`, `NO`, and `HC` variables.

(a) Obtain the principal component solution to the factor model $\mathbf{X} = \boldsymbol{\mu} + L\mathbf{F} + \boldsymbol{\varepsilon}$ with number of factors $m = 1$ and $m = 2$ using:
   (i) the sample covariance matrix;
   (ii) the sample correlation matrix.

   In other words, you should find the matrix factor loadings $L \in \mathbb{R}^{n\times m}$, the specific variances $\psi_1, \dots, \psi_p \in \mathbb{R}$, and write down the proportions of variability (in percentages) due to the factors.

(b) Find the angle between the first factor loading in (i) and that the first factor loading in (ii).

(c) For the $m = 2$ case, compare the factor loadings obtained in (i) and that in (ii) using orthogonal Procrustes analysis.

(d) Comment on your results.

2. (*Population Canonical Correlation Analysis*) The 2×1 random vectors $\mathbf{X}$ and $\mathbf{Y}$ have joint mean vector and joint covariance matrix

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix} \in \mathbb{R}^{4\times 4}, \quad \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \in \mathbb{R}^{4\times 4},$$

where

$$\boldsymbol{\mu}_X = \begin{bmatrix} -3 \\ 2 \end{bmatrix}, \quad \boldsymbol{\mu}_Y = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and

$$\Sigma_X = \begin{bmatrix} 8 & 2 \\ 2 & 5 \end{bmatrix}, \quad \Sigma_Y = \begin{bmatrix} 6 & -2 \\ -2 & 7 \end{bmatrix}, \quad \Sigma_{YX}^{\mathsf{T}} = \Sigma_{XY} = \begin{bmatrix} 3 & 1 \\ -1 & 3 \end{bmatrix}.$$

(a) Calculate the canonical correlation ρ_1 (the largest), ρ_2 (the second largest).

(b) Find the canonical correlation variables (U_1, V_1) and (U_2, V_2) corresponding to ρ_1 and ρ_2.

(c) Let $U = [U_1, U_2]^{\mathsf{T}}$ and $V = [V_1, V_2]^{\mathsf{T}}$. Evaluate

$$E\left(\begin{bmatrix} U \\ V \end{bmatrix}\right) \quad \text{and} \quad \operatorname{Cov}\left(\begin{bmatrix} U \\ V \end{bmatrix}\right) = \begin{bmatrix} \Sigma_U & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_V \end{bmatrix}$$

(d) Comment on the correlation structure between and within U and V.

3. (*Sample canonical correlation analysis*) The data set for this problem is obtained by taking four different measures of stiffness, `shock`, `vibrate`, `static1`, `static2`, for each of $n = 30$ boards. The first measurement involves sending a shock wave down the board, the second measurement

---

Date: November 15, 2015 (Version 1.5); due: November 25, 2015.

is determined while vibrating the board, and the last two measurements are obtained from static tests. The squared distances $d_j^2 = (\mathbf{x}_j - \overline{\mathbf{x}})^\mathsf{T} S^{-1} (\mathbf{x}_j - \overline{\mathbf{x}})$ are also included as the last column in the data matrix.

Let $\mathbf{X} = [X_1, X_2]^\mathsf{T}$ be the random vector representing the dynamic measures of stiffness, and let $\mathbf{Y} = [Y_1, Y_2]^\mathsf{T}$ be the random vector representing the static measures of stiffness. Load the data matrix `p3.txt` (R command: `stiff = read.table("p3.txt")`)

(a) Perform a canonical correlation analysis of these data by computing the singular value decomposition of an appropriate matrix formed from the sample covariance matrices. You may compare your result with that obtained by your software (if you use R, it is `cancor(X1,X2)`).

(b) Write the first canonical correlation variables U_1 and V_1 as linear combinations of `shock`, `vibrate`, `static1`, `static2`.

(c) Produce two scatterplots of the data: one in the coordinate plane of the first canonical correlation vectors, one in the plane of the second canonical correlation vectors.

(d) Based on the two plots and the values of the canonical correlations $\{\rho_1, \rho_2\}$, comment on the correlation structure captured by each canonical pair.

(e) Repeat (a) with sample correlation matrices in place of sample covariance matrices and verify that the pairs of canonical vectors obtained are related via scaling by the sample standard deviation matrix.

4. (*Canonical correlation analysis for angular measurements*) Some observations are in the form of angles. Here we will see how to deal with such data.

(a) Consider a bivariate random vector $\mathbf{X} = [X_1, X_2]^\mathsf{T}$ with a uniform distribution inside a circle of radius 1 centered at some unknown point

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \in \mathbb{R}^2.$$

Then $E(\mathbf{X}) = \boldsymbol{\mu}$. A sample of $n = 4$ is taken. The observed values are

$$\begin{bmatrix} 0.9 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}, \quad \begin{bmatrix} 0.6 \\ -0.6 \end{bmatrix}, \quad \begin{bmatrix} 0.9 \\ 0 \end{bmatrix}.$$

Compute sample mean $\overline{\mathbf{x}}$ and sample covariance matrix. Is $\overline{\mathbf{x}}$ a good estimator of $\boldsymbol{\mu}$? Why or why not?

(b) We consider an angular valued random variable θ, note that this can always be represented as a random vector $\mathbf{Y} = [\cos\theta, \sin\theta]^\mathsf{T}$ that takes value on the circle. Show that

$$\mathbf{b}^\mathsf{T}\mathbf{Y} = \sqrt{b_1^2 + b_2^2}\cos(\theta - \beta)$$

where $b_1/\sqrt{b_1^2 + b_2^2} = \cos\beta$ and $b_2/\sqrt{b_1^2 + b_2^2} = \sin\beta$. Here $\mathbf{b} = [b_1, b_2]^\mathsf{T} \in \mathbb{R}^2$ is a constant vector.

(c) Let $\mathbf{X} = X$ be a random vector with a single component, i.e., just a random variable. Here X is not angular valued. Show that the population canonical correlation is

$$\rho_1 = \max_\beta \operatorname{Corr}\big(X, \cos(\theta - \beta)\big)$$

and that selecting the population canonical correlation variable V_1 amounts to selecting a new 'origin' or 'baseline' β for the angle θ.

(d) Let X is a random variable representing ozone (O_3) levels and θ is a angular random variable representing wind direction measured from the north. We make 19 observations to obtain

the sample correlation matrix

$$R = \begin{bmatrix} R_X & R_{X\theta} \\ R_{\theta X} & R_\theta \end{bmatrix} = \begin{array}{c} \\ \mathrm{O}_3 \\ \cos\theta \\ \sin\theta \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{O}_3 & \cos\theta & \sin\theta \end{array} \\ \begin{bmatrix} 1.000 & 0.166 & 0.694 \\ 0.166 & 1.000 & -0.051 \\ 0.694 & -0.051 & 1.000 \end{bmatrix} \end{array}.$$

Find the sample canonical correlation $\widehat{\rho}_1$ and the sample canonical correlation variable $\widehat{V}_1$ representing the new origin β.

(e) Let ϕ be another angular valued random variable and let $\mathbf{X} = [\cos\phi, \sin\phi]^\mathsf{T}$. Then similar to (b), we get

$$\mathbf{a}^\mathsf{T}\mathbf{X} = \sqrt{a_1^2 + a_2^2}\cos(\phi - \alpha).$$

Now show that

$$\rho_1 = \max_{\alpha,\beta} \mathrm{Corr}\big(\cos(\phi - \alpha), \cos(\theta - \beta)\big).$$

(f) Let ϕ and θ be two angular random variables representing wind directions at 6:00 A.M. and at 12:00 P.M. We make 21 measurements of $\mathbf{X}$ and $\mathbf{Y}$ (related to ϕ and θ as in (b) and (d)). We obtain the sample correlation matrix

$$R = \begin{bmatrix} R_X & R_{XY} \\ R_{YX} & R_Y \end{bmatrix} = \begin{array}{c} \\ \cos\phi \\ \sin\phi \\ \cos\theta \\ \sin\theta \end{array} \begin{array}{c} \begin{array}{cccc} \cos\phi & \sin\phi & \cos\theta & \sin\theta \end{array} \\ \begin{bmatrix} 1.000 & -0.291 & 0.440 & 0.372 \\ -0.291 & 1.000 & -0.205 & 0.243 \\ 0.440 & -0.205 & 1.000 & 0.181 \\ 0.372 & 0.243 & 0.181 & 1.000 \end{bmatrix} \end{array}.$$

Find the sample canonical correlation $\widehat{\rho}_1$ and sample canonical correlation variables $\widehat{U}_1$ and $\widehat{V}_1$.

5. (*Proofs behind* CCA) Let $A \in \mathbb{R}^{p\times p}$ and $B \in \mathbb{R}^{q\times q}$ be symmetric positive definite matrices and $C \in \mathbb{R}^{p\times q}$. Let

$$G = A^{-1/2}CB^{-1/2} \in \mathbb{R}^{p\times q}.$$

We shall write $\lambda_{\max}(M)$ for the largest eigenvalue of a matrix M.

(a) Suppose $p = q$. Show that eigenvalues of $B^{-1}A$, $B^{-1/2}AB^{-1/2}$, and AB^{-1} are all equal. What are the relations between the eigenvectors?

(b) Suppose $p = q$. Show that

$$\max_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\mathsf{T}A\mathbf{x} : \mathbf{x}^\mathsf{T}B\mathbf{x} = 1\} = \max_{\mathbf{y}\in\mathbb{R}^p}\{\mathbf{y}^\mathsf{T}B^{-1/2}AB^{-1/2}\mathbf{y} : \mathbf{y}^\mathsf{T}\mathbf{y} = 1\}.$$

By using Problem **7** in Homework **2**, deduce that

$$\max_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\mathsf{T}A\mathbf{x} : \mathbf{x}^\mathsf{T}B\mathbf{x} = 1\} = \lambda_{\max}(B^{-1/2}AB^{-1/2}),$$

$$\operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\mathsf{T}A\mathbf{x} : \mathbf{x}^\mathsf{T}B\mathbf{x} = 1\} = \mathbf{q}_{\max},$$

where $\mathbf{q}_{\max} \in \mathbb{R}^p$ is the eigenvector corresponding to the largest eigenvalue.

(c) Show that if we fix $\mathbf{x} \in \mathbb{R}^p$ and just maximize over all $\mathbf{y} \in \mathbb{R}^q$, then

$$\max_{\mathbf{y}\in\mathbb{R}^q}\{(\mathbf{x}^\mathsf{T}C\mathbf{y})^2 : \mathbf{y}^\mathsf{T}B\mathbf{y} = 1\} = \max_{\mathbf{y}\in\mathbb{R}^q}\{\mathbf{y}^\mathsf{T}[C^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}C]\mathbf{y} : \mathbf{y}^\mathsf{T}B\mathbf{y} = 1\}$$

and deduce that from (a) and (b) that

$$\max_{\mathbf{y}\in\mathbb{R}^q}\{(\mathbf{x}^\mathsf{T}C\mathbf{y})^2 : \mathbf{y}^\mathsf{T}B\mathbf{y} = 1\} = \lambda_{\max}(B^{-1}C^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}C).$$

Show that the largest eigenvalue of a rank-1 matrix $\mathbf{a}\mathbf{b}^\mathsf{T}$ is $\mathbf{b}^\mathsf{T}\mathbf{a}$ and deduce that

$$\max_{\mathbf{y}\in\mathbb{R}^q}\{(\mathbf{x}^\mathsf{T}C\mathbf{y})^2 : \mathbf{y}^\mathsf{T}B\mathbf{y} = 1\} = \mathbf{x}^\mathsf{T}CB^{-1}C^\mathsf{T}\mathbf{x}.$$

(d) Using (a), (c), and Problem **7** in Homework **2**, show that

$$\max_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{(\mathbf{x}^\mathsf{T}C\mathbf{y})^2 : \mathbf{x}^\mathsf{T}A\mathbf{x} = 1,\ \mathbf{y}^\mathsf{T}B\mathbf{y} = 1\} = \lambda_{\max}(GG^\mathsf{T}).$$

(e) Let $\sigma_1, \dots, \sigma_p \in \mathbb{R}$, $\mathbf{u}_1, \dots, \mathbf{u}_p \in \mathbb{R}^p$, $\mathbf{v}_1, \dots, \mathbf{v}_p \in \mathbb{R}^q$ be the singular values and left/right singular vectors of G. By Problem **7** in Homework **2**, show that

$$\max_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\mathsf{T}GG^\mathsf{T}\mathbf{x} : \mathbf{x}^\mathsf{T}\mathbf{x} = 1,\ \mathbf{u}_i^\mathsf{T}\mathbf{x} = 0,\ i = 1, \dots, k-1\} = \sigma_k^2,$$

$$\operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\mathsf{T}GG^\mathsf{T}\mathbf{x} : \mathbf{x}^\mathsf{T}\mathbf{x} = 1,\ \mathbf{u}_i^\mathsf{T}\mathbf{x} = 0,\ i = 1, \dots, k-1\} = \mathbf{u}_k,$$

for $k = 1, \dots, p$. Hence deduce that

$$\max_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{\mathbf{x}^\mathsf{T}C\mathbf{y} : \mathbf{x}^\mathsf{T}A\mathbf{x} = 1,\ \mathbf{y}^\mathsf{T}B\mathbf{y} = 1,\ \mathbf{u}_i^\mathsf{T}A\mathbf{x} = 0,\ i = 1, \dots, k-1\} = \sigma_k,$$

$$\operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{\mathbf{x}^\mathsf{T}C\mathbf{y} : \mathbf{x}^\mathsf{T}A\mathbf{x} = 1,\ \mathbf{y}^\mathsf{T}B\mathbf{y} = 1,\ \mathbf{u}_i^\mathsf{T}A\mathbf{x} = 0,\ i = 1, \dots, k-1\} = (A^{-1/2}\mathbf{u}_k, B^{-1/2}\mathbf{v}_k),$$

for $k = 1, \dots, p$. Finally show that

$$\max_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{\mathbf{x}^\mathsf{T}C\mathbf{y} : \mathbf{x}^\mathsf{T}A\mathbf{x} = 1,\ \mathbf{y}^\mathsf{T}B\mathbf{y} = 1,\ \mathbf{u}_i^\mathsf{T}A\mathbf{x} = 0,\ \mathbf{v}_i^\mathsf{T}B\mathbf{y} = 0,\ i = 1, \dots, k-1\} = \sigma_k,$$

$$\operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{\mathbf{x}^\mathsf{T}C\mathbf{y} : \mathbf{x}^\mathsf{T}A\mathbf{x} = 1,\ \mathbf{y}^\mathsf{T}B\mathbf{y} = 1,\ \mathbf{u}_i^\mathsf{T}A\mathbf{x} = 0,\ \mathbf{v}_i^\mathsf{T}B\mathbf{y} = 0,\ i = 1, \dots, k-1\} = (A^{-1/2}\mathbf{u}_k, B^{-1/2}\mathbf{v}_k),$$

for $k = 1, \dots, p$.

6. (*Linear discriminant analysis*) The admissions committee of a business school used GPA and GMAT scores to make admission decisions. The values for the variable `admit = 1,2,3` correspond to admission decisions of `yes`, `no`, `waitlist`. Lable the data set `p6.txt` (R commands: `gsbdata = read.table("p6.txt")`; `colnames(gsbdata)=c("GPA", "GMAT","admit")`).

(a) Calculate $\overline{\mathbf{x}}_i, \overline{\mathbf{x}}$ and S_{pool}.

(b) Calculate the sample within groups matrix W, its inverse W^{-1}, and the sample between groups matrix B. Find the eigenvalues and eigenvectors of $W^{-1}B$. (R command for A^{-1} is `solve(A)`).

(c) Use the linear discriminants derived from these eigenvectors to classify the two new observations $\mathbf{x} = [3.21, 497]^\mathsf{T}$ and $\mathbf{x} = [3.22, 497]^\mathsf{T}$.

(d) Scatterplot the original data set on the plane of the first two discriminants, labeled by admission decisions. Comment on the results in (c). Is this a good admission policy?