**ISYE 7405**

# Homework 6

*Jacob Aguirre* *Email: aguirre@gatech.edu*
*Instructor: Dr. Shihao Yang*

1. (Problem 11.2.6 in Mardia, Kent and Bibby) (See Bartlett, 1965) The following problem involves two multivariate normal populations with the same means but different covariance matrices. In discriminating between monozygotic and dizygotic twins of like sex on the basis of simple physical measurements such as weight, height, etc., the observations recorded are the differences $x_1, \ldots, x_p$ between corresponding measurements on each set of twins. As either twin might have been measured first, the expected mean differences are automatically zero. Let the covariance matrices for the two types of twins be denoted $\Sigma_1$ and $\Sigma_2$, and assume for simplicity that

$$\Sigma_1 = \sigma_1^2\{(1-\rho)\boldsymbol{I} + \rho\mathbf{1}\mathbf{1}'\},$$
$$\Sigma_2 = \sigma_2^2\{(1-\rho)\boldsymbol{I} + \rho\mathbf{1}\mathbf{1}'\}.$$

Under the assumption of multivariate normality, show that the ML discriminant function is proportional to

$$z_1 - \rho\{1 + (p-1)\rho\}^{-1} z_2 + \text{const.},$$

where $z_1 = x_1^2 + \cdots + x_p^2$ and $z_2 = (x_1 + \cdots + x_p)^2$. How would the boundary between the allocation regions be determined so that the two types of misclassification have equal probability?

**Solution.** Let $\Sigma = \{(1-\rho)\boldsymbol{I} + \rho\mathbf{1}\mathbf{1}'\}$. Therefore, $\Sigma_1 = \sigma_1^2\Sigma, \Sigma_2 = \sigma_2^2\Sigma$, and according to HW1,

$$\Sigma^{-1} = \frac{1}{1-\rho}\boldsymbol{I} - \frac{\rho}{(1-\rho)^2 + \rho(1-\rho)p}\mathbf{1}\mathbf{1}'.$$

Up to some constant independent of $x$, the difference between the log likelihoods of the two samples is proportional to

$$x'\Sigma^{-1}x + \text{constant} = x'\left\{\frac{1}{1-\rho}\boldsymbol{I} - \frac{\rho}{(1-\rho)^2 + \rho(1-\rho)p}\mathbf{1}\mathbf{1}'\right\}x + \text{constant}.$$

Ignoring some multiplicative constant $1/(1-\rho)$, the difference between the log likelihoods is proportional to

$$\sum_{i=1}^p x_i^2 - \frac{\rho}{1+(p-1)\rho}\left(\sum_{i=1}^p x_i\right)^2 + \text{constant} = z_1 - \rho\{1+(p-1)\rho\}^{-1}z_2 + \text{constant}.$$

From the discussion above, we know that

$$L = z_1 - \rho\{1+(p-1)\rho\}^{-1}z_2 = (1-\rho)x'\Sigma^{-1}x.$$

Without loss of generality, we assume that $\sigma_1^2 > \sigma_2^2$. And therefore the classification rule is Popualtion 1 if $L > \xi$, and Population 2 otherwise. To ensure that the two types of misclassification rates are the same, we need

$$P(L < \xi | Pop_1) = P(L > \xi | Pop_2).$$

Using the facts

$$\frac{L}{(1-\rho)\sigma_1^2} | Pop_1 \sim \chi_p^2, \quad \frac{L}{(1-\rho)\sigma_2^2} | Pop_2 \sim \chi_p^2,$$

we have

$$P\left(\frac{L}{(1-\rho)\sigma_1^2} < \frac{\xi}{(1-\rho)\sigma_1^2} | Pop_1\right) = P\left(\frac{L}{(1-\rho)\sigma_2^2} > \frac{\xi}{(1-\rho)\sigma_2^2} | Pop_2\right).$$

Therefore, $\xi$ can be determined through the following equation:

$$\text{pchisq}\left(\frac{\xi}{(1-\rho)\sigma_1^2}, df = p\right) = 1 - \text{pchisq}\left(\frac{\xi}{(1-\rho)\sigma_2^2}, df = p\right). \square$$

**2.** We discussed critical angles between a $q$-dimensional linear space $\mathcal{L}_A$ and a $p$-dimensional linear space $\mathcal{L}_B$ in class ($p \leq q$). We start from assuming $\mathcal{L}_B = \mathcal{L}_{row}(X)$, where $X_{p\times n}$ is of full rank $p$, and obtain that the critical angles are formed between the pairs $(Y_k, \widehat{Y}_k)$ with $Y_k \in \mathcal{L}_B$ and $\widehat{Y}_k = Y_k P_A \in \mathcal{L}_A$. Now suppose that instead of $X$, we begin with $\tilde{X} = MX$, where $M$ is nonsingular so that $\mathcal{L}_{row}(\tilde{X}) = \mathcal{L}_{row}(X)$. Show that the pairs $(\tilde{Y}_k, \widehat{\tilde{Y}}_k)$ are the same as $(Y_k, \widehat{Y}_k)$. In other words, it does not matter how the linear space is described.

**Solution.** Under the new basis, a vector $\widetilde{u}^\top = \widetilde{g}^\top \widetilde{X} \in \mathcal{L}_B$ projects into $\mathcal{L}_A$ as $\widehat{\widetilde{u}}^\top = \widetilde{g}^\top \widetilde{X} P_A$. Then the $\cos^2$ of the angle between $\widetilde{u}$ and $\widehat{\widetilde{u}}$ is

$$\cos^2\theta = \frac{||\widehat{\widetilde{u}}||^2}{||\widetilde{u}||^2} = \frac{\widetilde{g}^\top \widetilde{X} P_A \widetilde{X}^\top \widetilde{g}}{\widetilde{g}^\top \widetilde{X}\widetilde{X}^\top \widetilde{g}} = \frac{\widetilde{g}^\top M X P_A X^\top M^\top \widetilde{g}}{\widetilde{g}^\top M X X^\top M^\top \widetilde{g}} \equiv \frac{g^\top A g}{g^\top B g},$$

where

$$g^\top = \widetilde{g}^\top M, \quad A = X P_A X^\top, \quad B = XX^\top.$$

Then the maximization problem reduce to the this problem under the old basis $X$. Due to the transformation above, we have

$$\Xi^\top = \widetilde{\Xi}^\top M.$$

Therefore, under the new basis $\widetilde{X}$, we have

$$\widetilde{Y} = \widetilde{\Xi}^\top \widetilde{X} = \Xi^\top M^{-1} M X = \Xi^\top X = Y, \quad \widehat{\widetilde{Y}} = \widetilde{Y} P_A = Y P_A = \widehat{Y}. \square$$

**3.** Let $R$ be a $p$-dimensional random vector, and $S$ be a $q$-dimensional random vector. $\Sigma_{RR} > 0$ and $\Sigma_{SS} > 0$.

1. Show that for any fixed $g \in \mathcal{R}^p$, the member of $\mathcal{L}(S)$ most highly correlated with $g^\top R$ is $g^\top \Sigma_{RS}\Sigma_{SS}^{-1} S$.

   **Solution.** Since we have easily change the sign of correlation, we only need to consider the correlation squared. For $h \in \mathcal{L}(S)$, we have

   $$Corr^2(h^\top S, g^\top R) = \frac{h^\top \Sigma_{SR} g g^\top \Sigma_{RS} h}{h^\top \Sigma_{SS} h \times g^\top \Sigma_{RR} g} \propto \frac{h^\top \Sigma_{SR} g g^\top \Sigma_{RS} h}{h^\top \Sigma_{SS} h}.$$

   According to the re-stated version of the fundamental lemma, we need to find the eigen-vector corresponding to the largest eigen-value of the matrix

   $$\Sigma_{SS}^{-1/2} \Sigma_{SR} g g^\top \Sigma_{RS} \Sigma_{SS}^{-1/2}.$$

   It is a matrix of rank one, and its maximal eigen-value is the only nonzero eigen-value, which must be the same as its trace:

   $$\mathrm{tr}(\Sigma_{SS}^{-1/2} \Sigma_{SR} g g^\top \Sigma_{RS} \Sigma_{SS}^{-1/2}) = \mathrm{tr}(g^\top \Sigma_{RS} \Sigma_{SS}^{-1} \Sigma_{SR} g) = g^\top \Sigma_{RS} \Sigma_{SS}^{-1} \Sigma_{SR} g \equiv \lambda_1.$$

   We can easily verify that

   $$\Sigma_{SS}^{-1/2} \Sigma_{SR} g g^\top \Sigma_{RS} \Sigma_{SS}^{-1/2} \cdot \Sigma_{SS}^{-1/2} \Sigma_{SR} g = \lambda_1 \Sigma_{SS}^{-1/2} \Sigma_{SR} g,$$

   and therefore, $\Sigma_{SS}^{-1/2} \Sigma_{SR} g$ is the corresponding eigen-vector. Finally, we know that $h^* = \Sigma_{SS}^{-1/2} \Sigma_{SS}^{-1/2} \Sigma_{SR} g = \Sigma_{SS}^{-1} \Sigma_{SR} g$ maximizes the correlation, and the corresponding element in $\mathcal{L}(S)$ is $h^{*\top} S = g^\top \Sigma_{RS} \Sigma_{SS}^{-1} S$. $\square$

   **Remark:** We can also use Cauchy-Schwarz Inequality to prove this.

   $$Corr(h^T S, g^T R) = \frac{g^T \Sigma_{RS} h}{\sqrt{h^T \Sigma_{SS} h}\sqrt{g^T \Sigma_{RR} g}} = \frac{< (g^T \Sigma_{RS} \Sigma_{SS}^{-1})^T, h >_{\Sigma_{SS}}}{\sqrt{< h, h >_{\Sigma_{SS}}}\sqrt{g^T \Sigma_{RR} g}}$$

   is maximized when $h = (g^T \Sigma_{RS} \Sigma_{SS}^{-1})^T$, therefore, $h^T S = g^T \Sigma_{RS} \Sigma_{SS}^{-1} S$ maximizes this correlation.

2. If $p = 1$, then the greatest possible correlation square between $R$ and a linear combination of $S$ is $\lambda = \Sigma_{RS}\Sigma_{SS}^{-1}\Sigma_{SR}/\Sigma_{RR}$.

**Solution.** The correlation squared is

$$Corr^2(h^\top S, R) = \frac{h^\top \Sigma_{SR}\Sigma_{RS}h}{h^\top \Sigma_{SS}h \times \Sigma_{RR}},$$

with maximal value being

$$\frac{\lambda_1(\Sigma_{SS}^{-1/2}\Sigma_{SR}\Sigma_{RS}\Sigma_{SS}^{-1/2})}{\Sigma_{RR}}.$$

Since $\Sigma_{SS}^{-1/2}\Sigma_{SR}\Sigma_{RS}\Sigma_{SS}^{-1/2}$ has rank one, it only has one nonzero eigen-value, which must be the same as its trace:

$$\mathrm{tr}(\Sigma_{SS}^{-1/2}\Sigma_{SR}\Sigma_{RS}\Sigma_{SS}^{-1/2}) = \mathrm{tr}(\Sigma_{RS}\Sigma_{SS}^{-1}\Sigma_{SR}) = \Sigma_{RS}\Sigma_{SS}^{-1}\Sigma_{SR}.$$

Therefore, the maximal correlation squared is $\Sigma_{RS}\Sigma_{SS}^{-1}\Sigma_{SR}/\Sigma_{RR}$.

**4.** $\mathcal{L}$ is a $p$-dimensional subspace of $R^n (p \le n)$ with linearly independent basis vectors $\alpha_1, \alpha_2, \ldots, \alpha_p$. Let $A$ be a $p \times n$ matrix with rows $\alpha_1^\top, \ldots, \alpha_p^\top$. Show that the $p$-dimensional volume in $\mathcal{L}$ of the parallelepiped $C$ formed by $\alpha_1, \ldots, \alpha_p$ satisfied $\mathrm{vol}^2(C) = \det(AA^\top)$.
**Solution.** We use the following intuitive definition of the volume of a parallelepiped in a $p$-dim subspace of $n$-dim space:

$$\mathrm{vol}(C) = ||\alpha_1|| \times ||P_1^\perp \alpha_2|| \times \cdots \times ||P_{1:(p-1)}^\perp \alpha_p||,$$

where $P_{1:(j-1)}^\perp \alpha_j$ is the projection of $\alpha_j$ onto the space orthogonal to the linear space spanned by $\alpha_1, \cdots, \alpha_{j-1}$.
This definition is closely related to QR decomposition of a matrix. We have

$$A_{p\times n} = L_{p\times p}\Gamma_{p\times n},$$

where $L$ is a lower-triangular matrix, and $\Gamma$ has orthonormal row vectors. By the construction of the QR decomposition, we know that the diagonal elements of $L$ are the lengths of $P_{1:(j-1)}^\perp \alpha_j$ for $j = 1, \ldots, p$. Therefore,

$$\mathrm{vol}(C) = \prod_{i=1}^p |l_{ii}| = \det(L) = \sqrt{\det(LL^\top)} = \sqrt{\det(L\Gamma\Gamma^\top L^\top)} = \sqrt{\det(AA^\top)}.\square$$

**5.** (Problem 10.2.10 in Mardia, Kent and Bibby) (Hotelling, 1936) Four examinations in reading speed, reading power, arithmetic speed, and arithmetic power were given to $n = 148$ children. The question of interest is whether reading ability is correlated with arithmetic ability. The correlations are given by

$$R_{11} = \begin{pmatrix} 1 & 0.6328 \\ 0.6328 & 1 \end{pmatrix}, \quad R_{22} = \begin{pmatrix} 1 & 0.4248 \\ 0.4248 & 1 \end{pmatrix}, \quad R_{12} = \begin{pmatrix} 0.2412 & 0.0586 \\ -0.0553 & 0.0655 \end{pmatrix}.$$

Using Exercise 10.2.9, verify that the canonical correlations are given by

$$\rho_1 = 0.3945, \quad \rho_2 = 0.0688.$$

**Solution.** Problem 10.2.9 seems silly. Using R we can directly obtain the following:

```
> r11<-matrix(c(1,0.6328,0.6328,1),2,2)
> r22<-matrix(c(1,0.4248,0.4248,1),2,2)
> r12<-matrix(c(0.2412,-0.0553,0.0586,0.0655),2,2)
>
> b<-sum(r12^2)+2*(r12[1,1]*r12[2,2]+r12[1,2]*r12[2,1])*r11[1,2]*r22[1,2]
-2*(r12[1,1]*r12[2,1]+r12[1,2]*r12[2,2])*r11[1,2]
-2*(r12[1,1]*r12[1,2]+r12[2,2]*r12[2,1])*r22[1,2]
>
```

```
> c<-det(r12)^2*(1+r11[1,2]^2*r22[1,2]^2-r11[1,2]^2-r22[1,2]^2)
>
> lambda1<-(b+sqrt(b^2-4*c))/(2*(1-r11[1,2]^2)*(1-r22[1,2]^2))
> lambda2<-(b-sqrt(b^2-4*c))/(2*(1-r11[1,2]^2)*(1-r22[1,2]^2))
>
> eg.matrix<-eigen(solve(r11)%*%r12%*%solve(r22)%*%t(r12))
>
> sqrt(eg.matrix$values)
[1] 0.39450592 0.06884787
> sqrt(c(lambda1,lambda2))
[1] 0.39450592 0.06884787
>
```

**6.** (Problem 10.2.13 in Mardia, Kent and Bibby)

1. Using the data matrix for the open/close book data in Example 10.2.5 and Table 1.2.1, show that the scores of the first eight individuals on the first canonical correlation variables are as follows:

| Subject | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\eta_1$ | 6.25 | 5.68 | 5.73 | 5.16 | 4.90 | 4.54 | 4.80 | 5.16 |
| $\phi_1$ | 6.35 | 7.44 | 6.67 | 6.00 | 6.14 | 6.71 | 6.12 | 6.30 |

**Solution.**. I got slightly different results from the book:

```
> R<-cov(dat)
> R11<-R[1:2,1:2]
> R22<-R[3:5,3:5]
> R12<-R[1:2,3:5]
> R21<-R[3:5,1:2]
> library(expm)
> K<-sqrtm(solve(R11))%*%R12%*%sqrtm(solve(R22))
> eg1<-eigen(K%*%t(K))
> eg2<-eigen(t(K)%*%(K))
>
> open<-dat[,3:5]
> close<-dat[,1:2]
> s11<-solve(sqrtm(matrix(cov(scale(close)),nrow=2,ncol=2)))
> s12 <- matrix(cov(scale(close),scale(open)),nrow = 2, ncol=3)
> s22 <- solve(sqrtm(matrix(cov(scale(open)),nrow = 3, ncol=3)))
> K1 <- matrix(s11 %*% s12 %*% s22, nrow=2, ncol=3)
> #svd(K1)
>
> score1<-as.matrix(dat[1:8,1:2])%*%sqrtm(solve(R11))%*%eg1$vectors
> score2<-as.matrix(dat[1:8,3:5])%*%sqrtm(solve(R22))%*%(-eg2$vectors)
>
>
> a<-sqrtm(solve(R11))%*%eg1$vectors
> b<-sqrtm(solve(R22))%*%(-eg2$vectors)
>
> eta1<-score1[,1]
> phi1<-score2[,1]
>
> a[,1]
[1] 0.02583319 0.05145928
```

```
> b[,1]
[1] 0.081909496 0.008020362 0.003454856
> rbind(eta1,phi1)
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
eta1 6.208816 5.641315 5.694017 5.125894 4.869425 4.508175 4.765264 5.126308
phi1 6.305144 7.394028 6.624761 5.956654 6.103198 6.662991 6.080371 6.260580
```

2. Plot the above eight points on a scattergram.

   **Solution.**. See Figure 1.

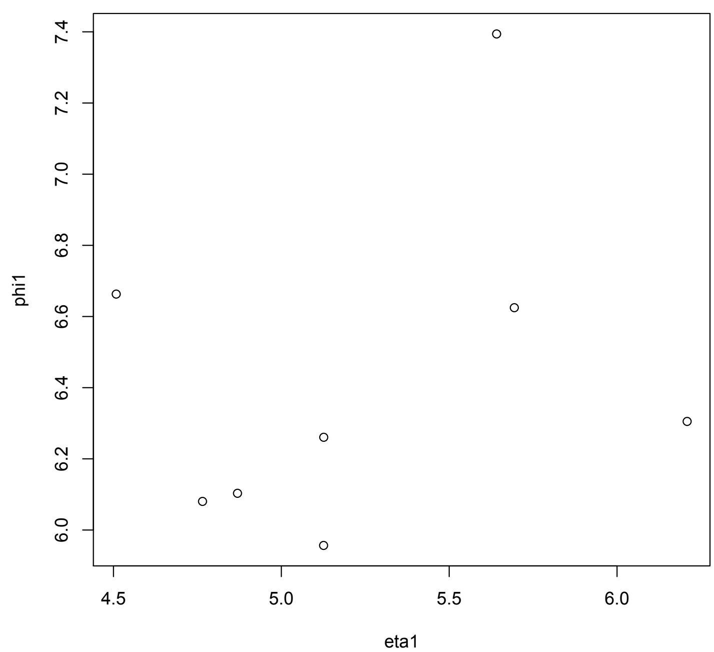


Figure 1: CCA

3. Repeat the procedure for the second canonical correlation variable and analyze the difference in the correlations. (The second canonical correlation is $r_2 = 0.041$ and the corresponding loading vectors are given by

$$\boldsymbol{a}_2' = (-0.064, 0.076), \quad \boldsymbol{b}_2' = (-0.091, 0.099, -0.014).)$$

   **Solution.**. See the code below.

```
> a[,2]
[1] -0.06361496  0.07544314
> b[,2]
[1] -0.09035660  0.09840149 -0.01433057
> rbind(eta2,phi2)
           [,1]      [,2]       [,3]      [,4]      [,5]      [,6]       [,7]      [,8]
eta2  1.2879860  1.876823  0.7362276 1.9330836 0.7451757  1.230439  1.8103277  1.527737
phi2 -0.6217682 -1.501199 -1.0815961 0.2211601 0.1120998 -1.254111 -0.4515606 -0.845868
>
> sqrt(eg1$values)
[1] 0.66305211 0.04094594
```