Final Exam - Multivariate Data Analysis

June 12, 2017

You will get no credit if you do not provide any detailed explanations for your answer. Whenever you use a new notation, definine it clearly first.

1. Look at Figure 1. Eigenvectors and eigevalues are obtained from the sample corre-

```
> head(data)
   100m  200m  400m 800m 1500m 5000m 10000m Marathon  Country
1 10.39 20.81 46.84 1.81  3.70 14.04  29.36   137.72 argentin
2 10.31 20.06 44.84 1.74  3.57 13.28  27.66   128.30 australi
3 10.44 20.81 46.82 1.79  3.60 13.26  27.72   135.90  austria
4 10.34 20.68 45.04 1.73  3.60 13.22  27.45   129.95  belgium
5 10.28 20.58 45.91 1.80  3.75 14.68  30.55   146.62  bermuda
6 10.22 20.43 45.21 1.73  3.66 13.62  28.62   133.13   brazil
```

Eigenvectors:

```
> PCres$vectors[,1:2]
           [,1]        [,2]
[1,] -0.3175565 -0.56687750
[2,] -0.3369792 -0.46162589
[3,] -0.3556454 -0.24827331
[4,] -0.3686841 -0.01242993
[5,] -0.3728099  0.13979665
[6,] -0.3643741  0.31203045
[7,] -0.3667726  0.30685985
[8,] -0.3419261  0.43896267
```

Eigenvalues:

```
> PCres$values[1:4]
[1] 6.6221461 0.8776183 0.1593211 0.1240494
>
> cumsum(PCres$values)[1:4]/sum(PCres$values)
[1] 0.8277683 0.9374706 0.9573857 0.9728919
```

Figure 1: National track records for men and PCA result (Problem 1)

lation matrix of the given dataset. Give your answers for the following questions.

- Interpret the first two principal components.
- Explain how many principal components are necessary (based on the given numbers).

2. The random vectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ have

$$
\begin{aligned}
Cov(\mathbf{x}^{(1)}) &= \Sigma_{11} \\
Cov(\mathbf{x}^{(2)}) &= \Sigma_{22} \\
Cov(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) &= \Sigma_{12} \\
Cov(\mathbf{x}) &= \Sigma
\end{aligned}
$$

where Σ is the full-rank covariance matrix of $\mathbf{x} = ((\mathbf{x}^{(1)})^T, (\mathbf{x}^{(2)})^T)^T$.
For coefficient vectors $\mathbf{a}$ and $\mathbf{b}$, form the linear combinations

$$U = \mathbf{a}^T\mathbf{x}^{(1)} \text{ and } V = \mathbf{b}^T\mathbf{x}^{(2)}.$$

Here, let ${\rho^*}_1^2$ be the largest eigenvalues of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ and denote its associated eigenvector by $\mathbf{e}_1$. And let ${\rho^*}_1^2$ be the largest eigenvalue of $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$ and denote its associated eigenvector by $\mathbf{f}_1$. Note that ${\rho^*}_1^2$ is common for the both matrices.
Show that $\max_{\mathbf{a},\mathbf{b}} Corr(U, V) = \rho_1^*$ attained by

$$U_1 = \mathbf{e}_1^T\Sigma_{11}^{-1/2}\mathbf{x}^{(1)} \text{ and } V_1 = \mathbf{f}_1^T\Sigma_{22}^{-1/2}\mathbf{x}^{(2)}.$$

3. Look at Figure 2. Give your answers for the following questions.

- Write down the fitted orthogonal factor model by using the numbers given in Fig 2.
- Interpret the first two factors.
- Explain a method to get a better interpretation for the factor analysis.

4. Consider the classification problem with two classes. The three key components of the expected cost of missclassification (ECM) are (1) prior probability (p_1, p_2), (2) misclassification cost $(c(2|1), c(1|2))$, and (3) density functions $(f_1(x), f_2(x))$. Then, ECM is defined as

$$ECM = c(2|1)\int_{R_2} f_1(x)dxp_1 + c(1|2)\int_{R_1} f_2(x)dxp_2.$$

examination scores in $p = 6$ subject areas for $n = 220$ male students (Lawley and Maxwell, 1971).

The sample correlation matrix is as follows.

```
> R
     Gaelic English History Arithmetic Algebra Geometry
[1,]  1.000   0.439   0.410      0.288   0.329    0.248
[2,]  0.439   1.000   0.351      0.354   0.320    0.329
[3,]  0.410   0.351   1.000      0.164   0.190    0.181
[4,]  0.288   0.354   0.164      1.000   0.595    0.470
[5,]  0.329   0.320   0.190      0.595   1.000    0.464
[6,]  0.248   0.329   0.181      0.470   0.464    1.000
```

```
> res.fac1<-factanal(covmat=R,factors=2,rotation="none")
> res.fac1

Call:
factanal(factors = 2, covmat = R, rotation = "none")

Uniquenesses:
    Gaelic    English    History Arithmetic    Algebra   Geometry 
     0.510      0.594      0.644      0.377      0.431      0.628 

Loadings:
           Factor1 Factor2
Gaelic      0.553   0.429 
English     0.568   0.288 
History     0.392   0.450 
Arithmetic  0.740  -0.273 
Algebra     0.724  -0.211 
Geometry    0.595  -0.132 

               Factor1 Factor2
SS loadings      2.209   0.606
Proportion Var   0.368   0.101
Cumulative Var   0.368   0.469

The degrees of freedom for the model is 4 and the fit was 0.0109
```

Figure 2: Sample correlation matrix for examination scores (6 subjects) and factor analysis result (Problem 3)

Then,

$$R_1 = \{x | \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}\}.$$

After defining necessary notations clearly, answer the following questions.

1. Suppose that the density functions are assumed to be multivariate normal distributions with a common covariance, and every parameters are known here. Show that the classification regions are linearly separated.

2. From the linear discriminant analysis, we get the following confusion matrix. See Figure 1. Compute the apprarent error rate (APER). Explain why APER is not a good measure for performance comparison and provide a better measure with a detailed algorithm.

π_1 : riding-mower owners, π_2 : nonowners

$$
\begin{array}{c}
\text{Predicted membership} \\
\text{Actual membership} \quad
\begin{pmatrix}
 & \pi_1 & \pi_2 \\
\pi_1 & 10 & 2 \\
\pi_2 & 2 & 10
\end{pmatrix}
\end{array}
$$

Figure 3: Analysis results for Riding-mower data (Problem 4)

5. Look at Figure 3. Give the dendrgoram when the single linkage is used. You should

distance matrix between pairs of five objects.

$$
\begin{pmatrix}
 & 1 & 2 & 3 & 4 & 5 \\
1 & 0 & & & & \\
2 & 9 & 0 & & & \\
3 & 3 & 7 & 0 & & \\
4 & 6 & 5 & 9 & 0 & \\
5 & 11 & 10 & 2 & 8 & 0
\end{pmatrix}
$$

Figure 4: Sample distance matrix (Problem 5)

show how the dendrogram evolves at every step.