Addenda to the Multivariate Analysis Notes

Richard L. Smith

April 24, 2019

These notes are intended to supply a few corrections and amendments to my multivariate anaysis notes that arose during the class taught in Spring 2019. My lectures this semester did not cover Chapter 5.

1 Union-Intersecton Principle for Testing Covariance Matrix

On page 18 (Case III) I define a test statistic

$$z_a^2 \;=\; (n-1)\frac{a^T S a}{a^T \Sigma_0 a}$$

and claim the null distribution for fixed a is χ_1^2. This should be χ_{n-1}^2. The proof is as follows. First, Proposition 10 of Chapter 1 shows that when $\Sigma = \Sigma_0$, $(n-1)S \sim W_p[\Sigma_0, n-1]$. Second, Proposition 4 shows that $(n-1)a^T S a \sim W_1[a^T\Sigma_0 a, n-1]$. The result then follows from Proposition 5.

2 R code for MANOVA (page 26)

The computations for the original notes were performed in S-Plus and not all the code is transferable to R. The `plot.design` function works as advertised. To get the other results on this page, use the following R code:

```
plot(Sepal.Length~Species,iris)
plot(Sepal.Width~Species,iris)
plot(Petal.Length~Species,iris)
plot(Petal.Width~Species,iris)
m1=manova(cbind(Sepal.Length,Sepal.Width,Petal.Length,Petal.Width)~Species,iris)
summary(m1,test='Pillai')
summary(m1,test='Wilks')
summary(m1,test='Hotelling-Lawley')
summary(m1,test='Roy')
```

3 R code for PCA and Factor Analysis (pages 32–43 and 51–55)

The following R code will reproduce the tables and plots of these two sections, with minor variations from the original S-Plus output. Note that some of the computations produce loadings of opposite sign to those given in the original notes.

```
# Chapter 3 - PCA
#
# use your own local path for file names.
exams=read.table('C:/Users/rsmith/feb08/UNC/s754/dataandprograms/PCA/exams.txt',header=F)
dimnames(exams)=list(1:nrow(exams),c('Vectors','Mechanics','Algebra','Analysis','Statistics'))
exams.prc=prcomp(exams,scale=F)
print(summary(exams.prc))
# Adjust graphical parameters as needed - cex parameter controls size of type
# or in some cases whether the type appears at all
par(mfrow=c(1,1),cex=0.9)
screeplot(exams.prc)
screeplot(exams.prc,type='lines')
# The following replaces the "plot.loadings" commands in my earlier S-Plus notes
par(mfrow=c(3,2),cex=1)
barplot(exams.prc$rotation[order(abs(exams.prc$rotation[,1]),decreasing=T),1],main='PC 1')
barplot(exams.prc$rotation[order(abs(exams.prc$rotation[,2]),decreasing=T),2],main='PC 2')
barplot(exams.prc$rotation[order(abs(exams.prc$rotation[,3]),decreasing=T),3],main='PC 3')
barplot(exams.prc$rotation[order(abs(exams.prc$rotation[,4]),decreasing=T),4],main='PC 4')
barplot(exams.prc$rotation[order(abs(exams.prc$rotation[,5]),decreasing=T),5],main='PC 5')
# print the standard deviations
print(exams.prc$sdev)
# Prediction plots
pr<-predict(exams.prc)
par(mfrow=c(2,2),cex=1.2)
plot(1:88,pr[,1],xlab='Student',ylab='Component 1',pch=20)
plot(1:88,pr[,2],xlab='Student',ylab='Component 2',pch=20)
plot(1:88,pr[,3],xlab='Student',ylab='Component 3',pch=20)
plot(1:88,pr[,4],xlab='Student',ylab='Component 4',pch=20)
# Predictions for new dataset
exam2<-matrix(scan(file='C:/Users/rsmith/feb08/UNC/s754/dataandprograms/PCA/exam2.txt'),
byrow=T,ncol=5)
dimnames(exam2)=list(1:nrow(exam2),c('Vectors','Mechanics','Algebra','Analysis','Statistics'))
print(predict(exams.prc,newdata=exam2))
# Biplot
par(mfrow=c(1,1),cex=1.1)
biplot(exams.prc)
#
# crimes dataset
# PCA - note use of "prcomp" rather than "princomp" though both could be used
crime<-matrix(scan(file='C:/Users/rsmith/feb08/UNC/s754/dataandprograms/PCA/crimes.txt'),
```

```
byrow=T,ncol=14)
crime<-t(crime)
nr<-length(crime[,1])
dimnames(crime)<-list(1:nr,c("Hcd","Wou","Hom","Het","Brk","Rob","Lar","Fra","Rec","Inj",
"For","Blc","Ass","Nal","Rev","Alc","Ind","Mot"))
crime.prc<-prcomp(crime,scale=T)
# two forms of screeplot
screeplot(crime.prc,type='barplot')
screeplot(crime.prc,type='lines')
# to see the numerical values on the scree plot type
summary(crime.prc)
# plot of loadings
par(mfrow=c(2,2),cex=0.7)
barplot(crime.prc$rotation[order(abs(crime.prc$rotation[,1]),decreasing=T),1],main='PC 1')
barplot(crime.prc$rotation[order(abs(crime.prc$rotation[,2]),decreasing=T),2],main='PC 2')
barplot(crime.prc$rotation[order(abs(crime.prc$rotation[,3]),decreasing=T),3],main='PC 3')
barplot(crime.prc$rotation[order(abs(crime.prc$rotation[,4]),decreasing=T),4],main='PC 4')
# Prediction plot
par(mfrow=c(2,2),cex=1.1)
plot(1950:1963,predict(crime.prc)[,1],ylab='PC 1',xlab='Year',pch=20)
plot(1950:1963,predict(crime.prc)[,2],ylab='PC 2',xlab='Year',pch=20)
plot(1950:1963,predict(crime.prc)[,3],ylab='PC 3',xlab='Year',pch=20)
plot(1950:1963,predict(crime.prc)[,4],ylab='PC 4',xlab='Year',pch=20)
#Biplot
par(mfrow=c(1,1),cex=1.1)
biplot(crime.prc)
#
# Chapter 4: Factor analysis
#
exams.fa=factanal(exams,factors=1,scores='regression')
# alternative would be
# exams.fa=factanal(exams,factors=1,scores='Bartlett')
plot(1:88,exams.fa$scores,pch=20)
# same for two-factor analysis
exams.fa=factanal(exams,factors=2,scores='regression')
par(mfrow=c(1,2))
plot(1:88,exams.fa$scores[,1],pch=20,xlab='Student',ylab='Score for Factor 1')
plot(1:88,exams.fa$scores[,2],pch=20,xlab='Student',ylab='Score for Factor 2')
# "biplot" function in R doesn't appear to work for factor analysis so do this:
# approximation to the biplot (without the principal factor arrows)
par(mfrow=c(1,1),cex=1.1)
plot(exams.fa$scores[,1],exams.fa$scores[,2],xlab='Factor 1',ylab='Factor 2',type='n')
text(exams.fa$scores[,1],exams.fa$scores[,2],labels=as.character(1:88),cex=0.8)
# plot of loadings
par(mfrow=c(1,2),cex=0.7)
```

```
barplot(exams.fa$loadings[order(abs(exams.fa$loadings[,1]),decreasing=T),1],main='Factor 1')
barplot(exams.fa$loadings[order(abs(exams.fa$loadings[,2]),decreasing=T),2],main='Factor 2')
```

4 A Clarification about Factor Analysis Scores

In Section 4.4.2 (pp. 49–50), I refer to two formulas to calculate factor analysis scores, denoted $\hat{f}$ (equation (4.10)) and f^* (equation (4.11). The sample R code provided above also mentioned that there are two options within the R program `factanal`, either `scores='Bartlett'` or `scores='regression'`.

In fact, $\hat{f}$ is the formula for Bartlett scores and f^* is the formula for regression scores (also known as Thompson scores). See Mardia, Kent and Bibby (1979), *Multivariate Analysis*, Academic Press, pp. 273–275.

Having reviewed the comments that I first made twenty years ago, I would like to reiterate my preference for f^*. The factor analysis model could also be interpreted as a random effects model in which, like random effect analysis of variance, there are unobserved random variables $f = (f_1, ..., f_k)^T$ which have a prescribed distribution given the model parameters. The estimator f^* is the conditional mean of the random effects given the observations, the same algorithm as is used in random effects analysis, see e.g. Pinheiro, J.C., and Bates, D.M. (2000), *Mixed-Effects Models in S and S-PLUS*, Springer, esp. pp. 100, 461.