PCA for the Crimes Dataset
R.L.S.
12/01/08

Several students have asked me about carrying out the PCA for the Crimes dataset in R, since the S-Plus commands given in the notes do not appear to work.

Indeed, it appears that the R command `princomp` does not work for this dataset, since there are more variables (18) than observations (14).

However, there is an alternative command `prcomp`, which works fine.

Here are the basic commands:

```
#
# input data and define headers
#
cr<-matrix(scan(file='crimes.txt'),byrow=T,ncol=14)
names<-c('hcd','wou','hom','het','brk','rob','lar','fra','rec','inj','for',
'blc','ass','mal','rev','alc','ind','mot')
# transpose matrix so that there is one column for variable
cr<-t(cr)
nr<-length(cr[,1])
dimnames(cr)<-list(1:nr,names)
#
# PCA via prcomp command (note centering and scaling)
#
prc1<-prcomp(cr,center=T,scale=T)
#
# print SDs of PCs
prc1$sdev
```

The last statement produces the output

```
> prc1$sdev
 [1] 3.591125e+00 1.646908e+00 9.782248e-01 8.270213e-01 5.676015e-01
 [6] 4.079726e-01 3.571915e-01 2.497862e-01 1.678720e-01 1.532584e-01
[11] 9.544367e-02 7.821820e-02 7.170100e-02 2.205450e-16
```

Thus, the SDs of the first few PC2 are 3.59, 1.65, 0.98,..., exactly as in the course notes.

The screeplot and biplot can be produced simply by the commands

```
screeplot(prc1)
biplot(prc1)
```

See Figs. 1 and 2.

Here is some R code that approximately reproduces Fig. 3.6 of the course notes:

```
name0<-c('PC1','PC2','PC3','PC4')
postscript('crime3.ps',horizo=F)
par(mfrow=c(4,1),oma=c(0,0,3,0))
for(k in 1:4){
# calculate loadings corresponding to k'th PC
loa1<-prc1$rotation[,k]
```

```
# increasing order of loa1
loa1ord<-order(-abs(loa1))
# rearrange loadings and names in this order
loa2<-loa1[loa1ord]
name2<-names[loa1ord]
# construct plot
plot(1:18,loa2,type='n',xlab='Component',ylab='Loading',main=name0[k])
y0<-0.05*(max(loa1)-min(loa1))
for(i in 1:18){
lines(c(i-0.3,i-0.3),c(0,loa2[i]))
lines(c(i+0.3,i+0.3),c(0,loa2[i]))
lines(c(i-0.3,i+0.3),c(loa2[i],loa2[i]))
if(loa2[i]>0){text(x=i,y=-y0,name2[i],cex=1)}
if(loa2[i]<0){text(x=i,y=y0,name2[i],cex=1)}
}
lines(c(0.7,18.3),c(0,0))
}
mtext('Fig. 3: Loadings Plots',
line=1,side=3,cex=1.5,outer=T)
dev.off()
```

Similar for Fig. 3.7:

```
# draw plots of scores
#
# first standardize all cols to mean 0 and variance 1
cr1<-cr
for(j in 1:18){cr1[,j]<-(cr[,j]-mean(cr[,j]))/sqrt(var(cr[,j]))}
postscript('crime4.ps',horizo=F)
par(mfrow=c(4,1),oma=c(0,0,3,0))
for(k in 1:4){
# calculate scores for k th PC
sco1<- cr1%*%prc1$rotation[,k]
plot(1950:1963,sco1,xlab='Year',ylab=name0[k])
}
mtext('Fig. 4: Score Plots',
line=1,side=3,cex=1.5,outer=T)
dev.off()
```

The results are shown here as Fig. 3 and Fig. 4.

Fig. 1: Screeplot

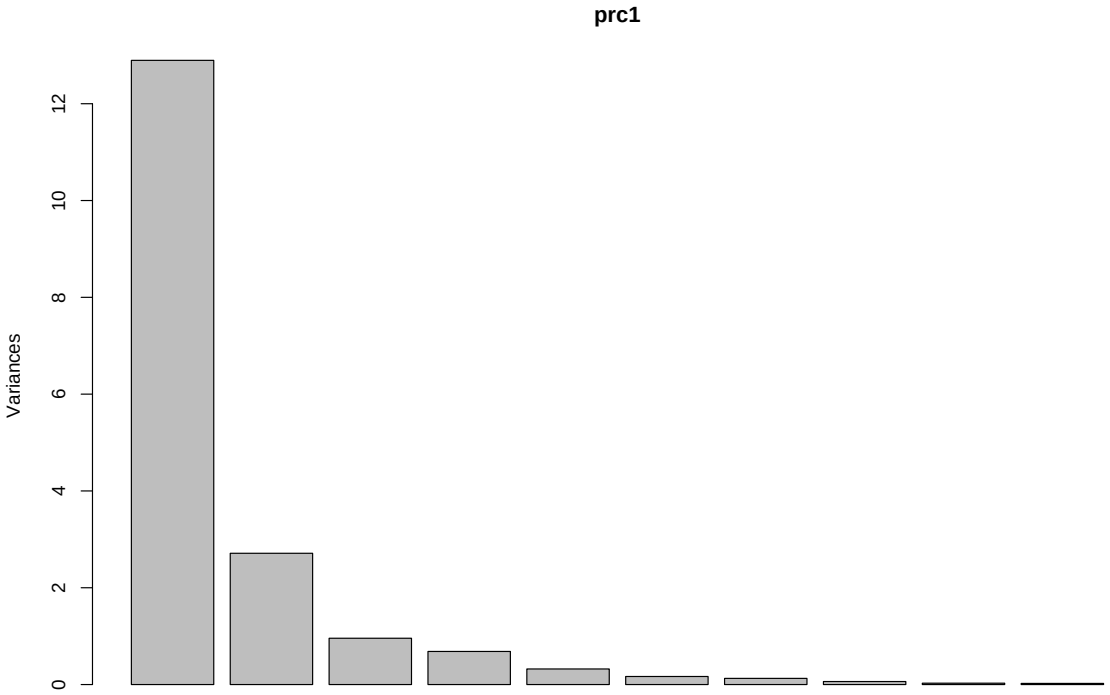

Fig. 2: Biplot

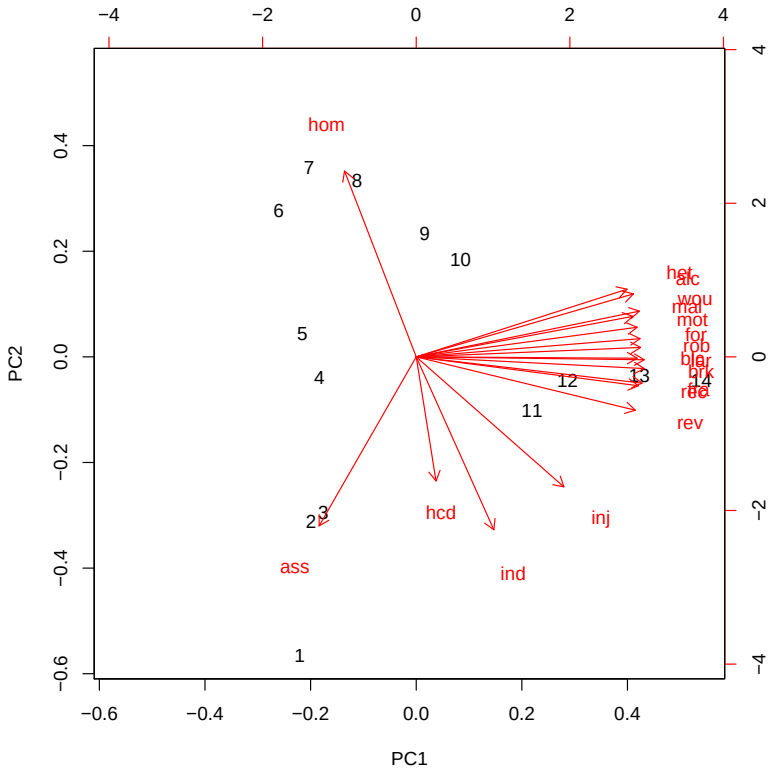

Fig. 3: Loadings Plots

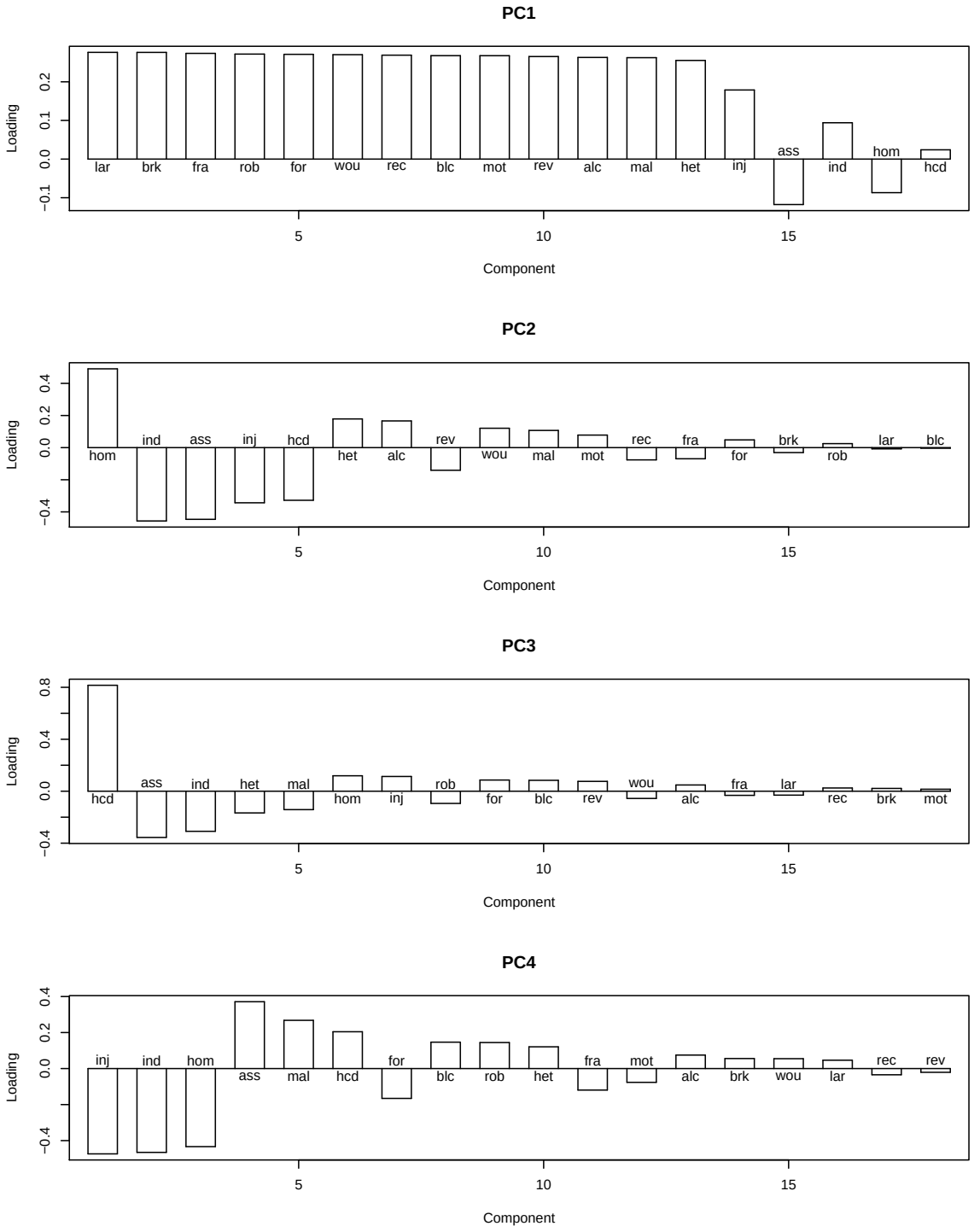

Fig. 4: Score Plots