Stat 537: Homework 6

Brandon Fenton and Kenny Flagg

Due Tuesday, March 8 at 5:00 PM

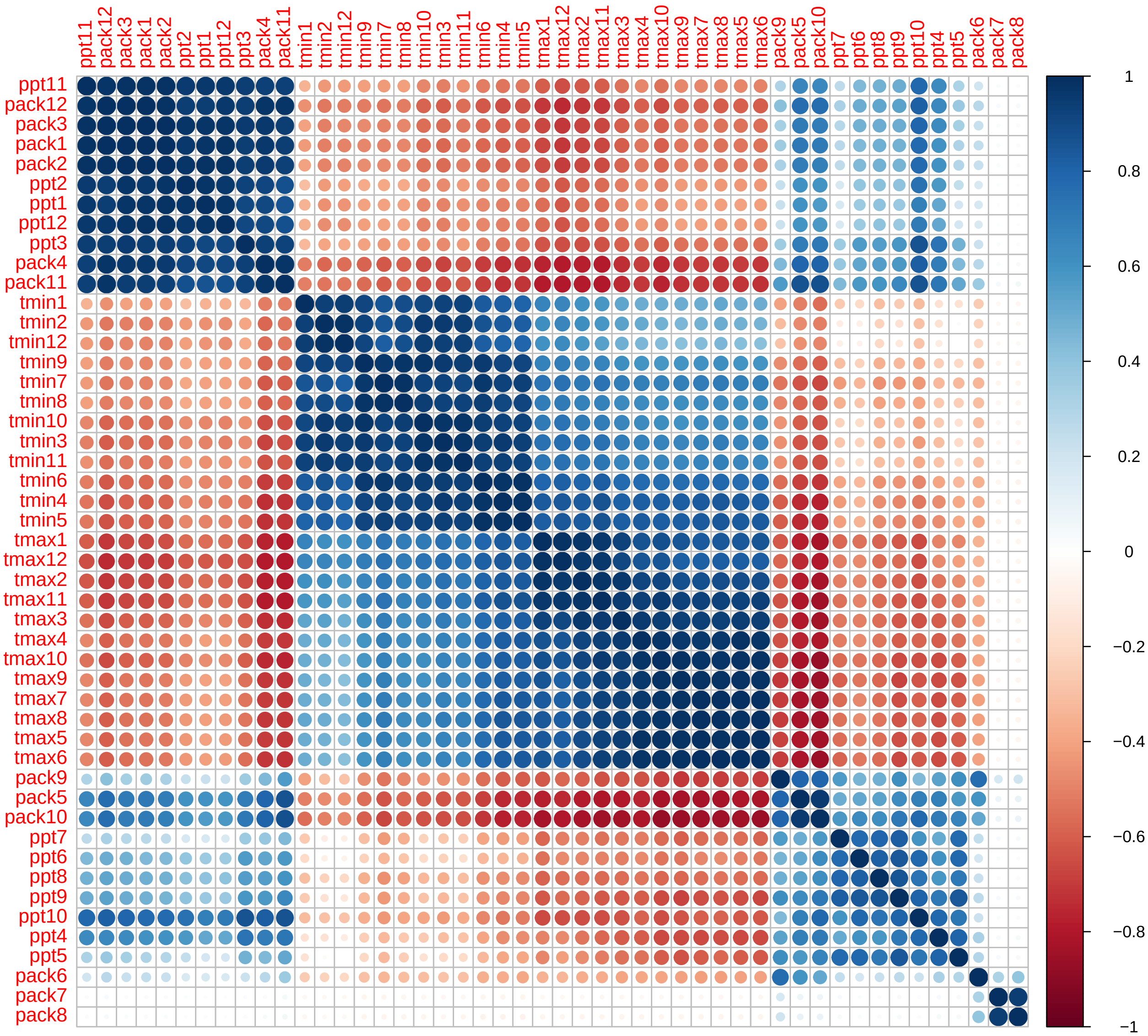
The following will involve working with a data set related to spatial variation in a suite of potential predictor variables and then, eventually, for building a predictive model for the presence/absence of whitebark pine in the greater Yellowstone Ecosystem.

For this work, we will focus on the historic climate and water balance data only (read the related sections carefully for variable names and definitions) in http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0111669

Read the article as we will find all of the paper interesting before the end of the semester. Initially, we are interested in doing a PCA of the monthly 1950 to 1980 average minimum and maximum temperature, precipitation, and snow pack (Q=48). Note that the number of each variable is the month of the year from January to December (1 to 12). You can use tc1_r below for this first analysis as I subset the entire data set for you.

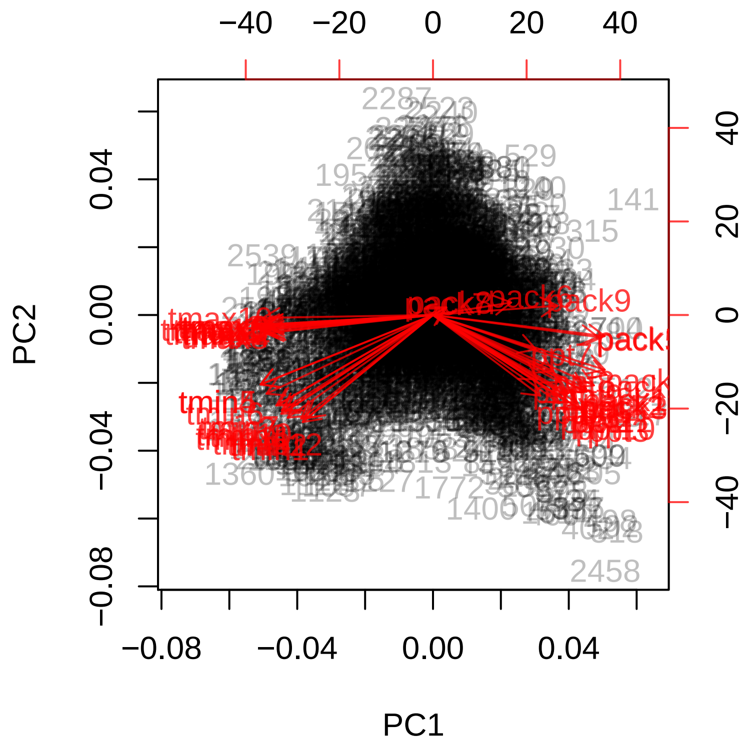
The provided code will source in a modified version of corrplot.mixed that I will discuss in class. The short version is that it orders the variables based on a hierarchical cluster analysis using a dissimilarity measure that treats positive and negative correlations equally (two variables that have r=0.5 are just as similar as two variables that have r=-0.5).

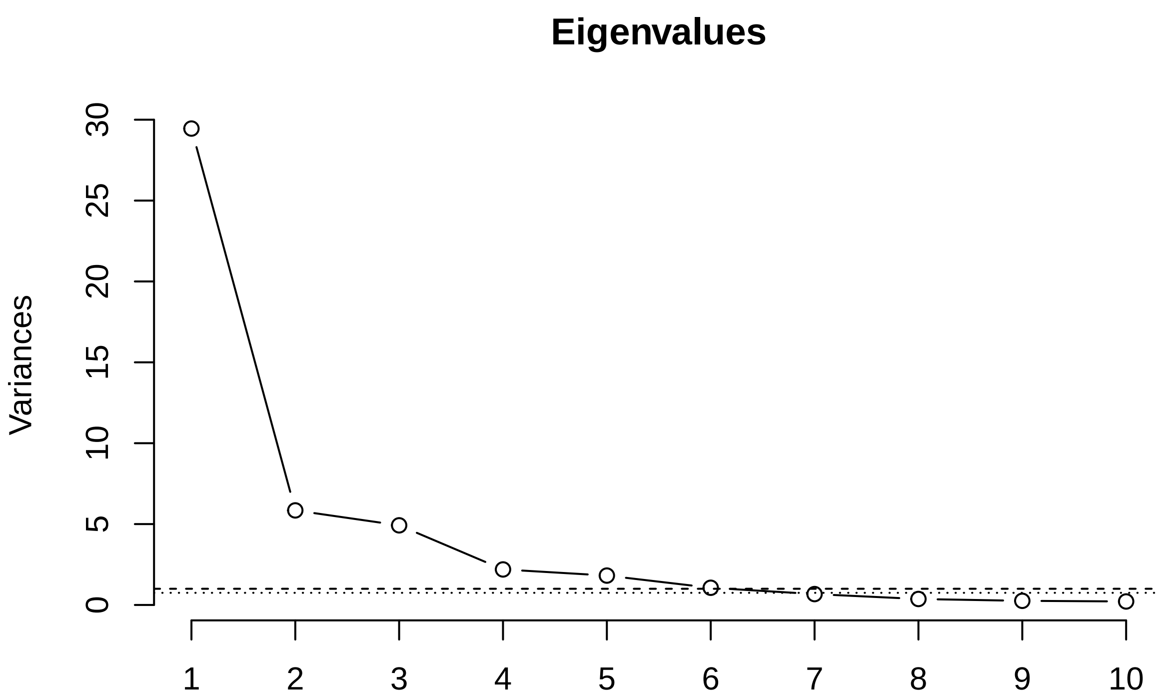
```
tc1<-read.csv("https://montana.box.com/shared/static/m5tv7r4ce7mw3w0vyqu0q7i1f8jflkkc.csv",header=T)
tc1$responsef<-factor(tc1$response)
tc1_r<-tc1[,c(4:39,64:75)]
cor1<-cor(tc1_r)

require(corrplot)
source("https://montana.box.com/shared/static/7ydjuwpraqpuovf7r1ovorsp828ckzs0.r")
corrplot_mg(cor1,order="hclust",tl.pos="lt")
```



1) *Discuss the pattern in the correlation matrix.*

2) *Perform a PCA of these variables based on the correlation matrix, report a biplot and scree plot. No discussion, just plots.*





3) *Interpret the first and fourth PCs based on the eigenvector coefficients.*

	PC1	PC4
tmin1	-0.1271	0.01232
tmin2	-0.1286	0.005188
tmin3	-0.1529	0.01738

	PC1	PC4
tmin4	-0.1683	0.01441
tmin5	-0.1679	0.0102
tmin6	-0.1624	0.01466
tmin7	-0.1505	0.02024
tmin8	-0.144	0.01358
tmin9	-0.1417	0.008078
tmin10	-0.1475	0.0006024
tmin11	-0.148	0.00945
tmin12	-0.1235	0.004401
tmax1	-0.1684	0.03742
tmax2	-0.1695	0.03018
tmax3	-0.1656	0.03162
tmax4	-0.161	0.03944
tmax5	-0.162	0.03472
tmax6	-0.163	0.03107
tmax7	-0.1621	0.02456
tmax8	-0.1628	0.02425
tmax9	-0.1622	0.03272
tmax10	-0.166	0.03176
tmax11	-0.1718	0.03426
tmax12	-0.1687	0.03231
ppt1	0.1304	0.02685
ppt2	0.1317	0.02036
ppt3	0.1445	0.003057
ppt4	0.1234	0.01119
ppt5	0.1018	-0.03667
ppt6	0.1047	-0.06689
ppt7	0.1003	-0.06889
ppt8	0.1175	-0.06821
ppt9	0.1239	-0.05269
ppt10	0.1423	-0.03936
ppt11	0.1401	0.02064
ppt12	0.1317	0.01912
pack1	0.1494	0.0254
pack2	0.1471	0.02367
pack3	0.1498	0.01702
pack4	0.1675	-0.001529
pack5	0.1611	0.09026
pack6	0.07624	0.3788
pack7	0.01087	0.6143
pack8	0.01306	0.6288
pack9	0.1216	0.1854
pack10	0.1675	0.0479
pack11	0.1692	0.02985
pack12	0.157	0.02553

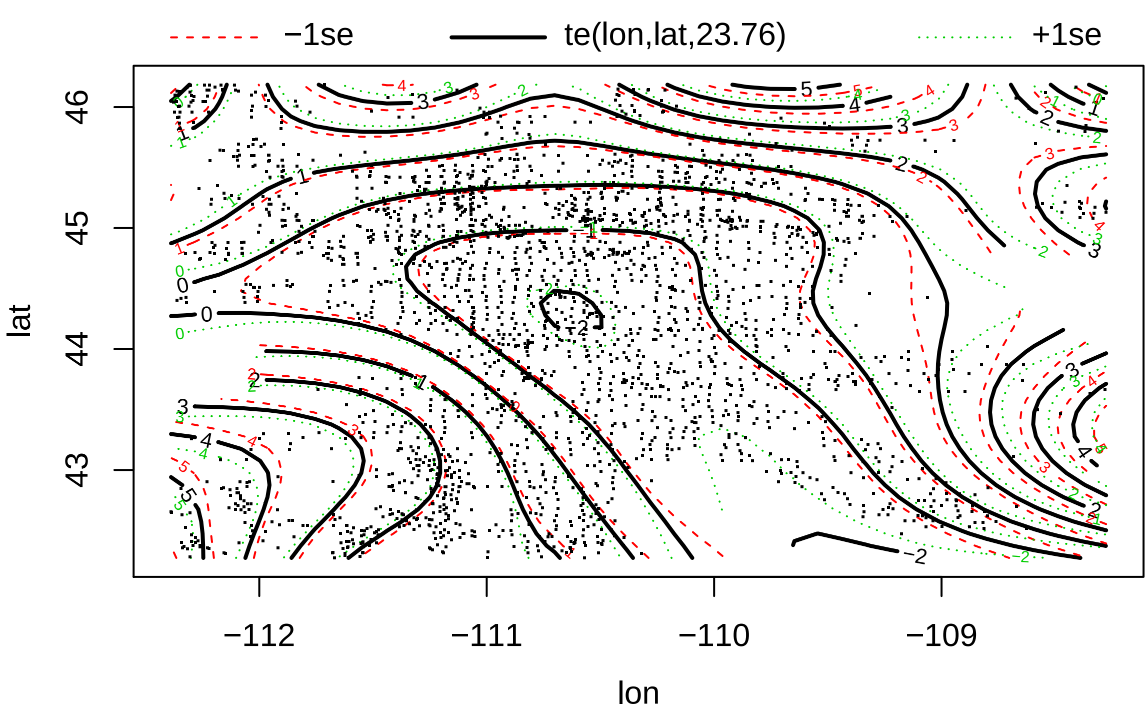
4) *Calculate the fourth PC using the predict function. Then replicate that calculation using the eigenvector and original variables (remember that the variables need to be standardized - the scale() function is a nice option). Show that they are the same.*

```
[1] 0
```

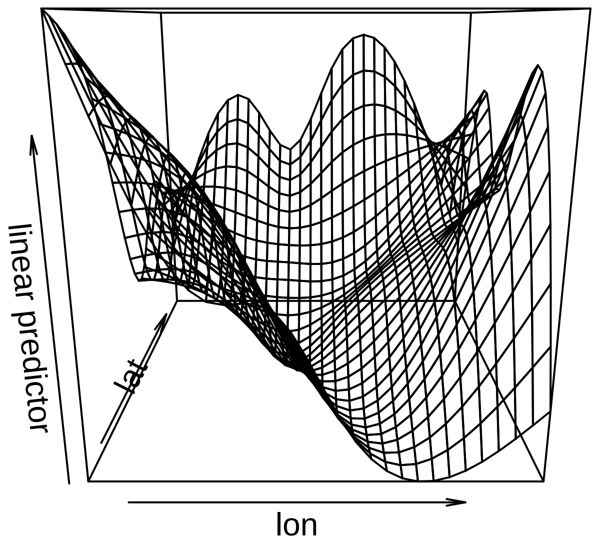
5) Now use your interpretation of PC4 to define a set of coefficients that should involve a reduced set of coefficients that are "different" from 0 to calculate the PC 4 scores. Make a plot of the real scores using all coefficients and based on this subset and compare the results.

6) For the moment, we will focus on just January minimum temperatures (something they used as an explanatory variable in their predictive model). The following fits a bivariate tensor-product penalized regression spline as function of the latitude and longitude of the observations and generates an estimated surface for the mean temperature as a deviation from the mean. Does location seem to matter for the temperatures? (I am not expecting you to know anything about the GAM I am using - it is just an estimate of the mean temperature surface.)

```r
require(mgcv)
gm1<-gam(tmin1~te(lon,lat),data=tc1)
 plot(gm1)
```




```r
 vis.gam(gm1)
```



7) Perform a Mantel test for a Euclidean distance matrix between the tmin1's vs a Euclidean distance matrix between the spatial locations defined by the lat and lon variables. Report the null hypothesis for the test specific to the situation. And report what you can conclude based on the result. [Note: this may take a while to run on your computer and might! cause you to run out of RAM. You are welcome to work with other students to obtain a computer with sufficient resources to complete the permutations.] Does this result agree or disagree with your previous result.

```
Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = spat.dists, ydis = tmin1.dists)

Mantel statistic r: 0.206
      Significance: 0.001

Upper quantiles of permutations (null model):
   90%    95%  97.5%    99%
0.0109 0.0141 0.0170 0.0194
Permutation: free
Number of permutations: 999
```

R Code Appendix:

Problem 2:

```r
# PCA
pcs <- prcomp(tc1_r, scale=T, center=T)

# Biplot
biplot(pcs, col =  c("#00000040", "#ff0000c0"))

# Scree plot
plot(pcs, type="lines", main = "Eigenvalues")
abline(h=c(1, 0.75), lty=2:3)
```

Problem 3:

```
require(pander)

pander(pcs$rotation[,c(1,4)])
```

Problem 4:

```
predict.scores <- predict(pcs)[,4]
ev.4 <- pcs$rotation[,4]
ev.scores <- as.matrix(scale(tc1_r)) %*% ev.4

round(sum(predict.scores - ev.scores), 10)
```

Problem 5:

Problem 7: # `{r a7, ref.label='p7_a', eval=F}` #