

Statistics 503 Homework 2

Due: in class on Friday, Feb 16 2007

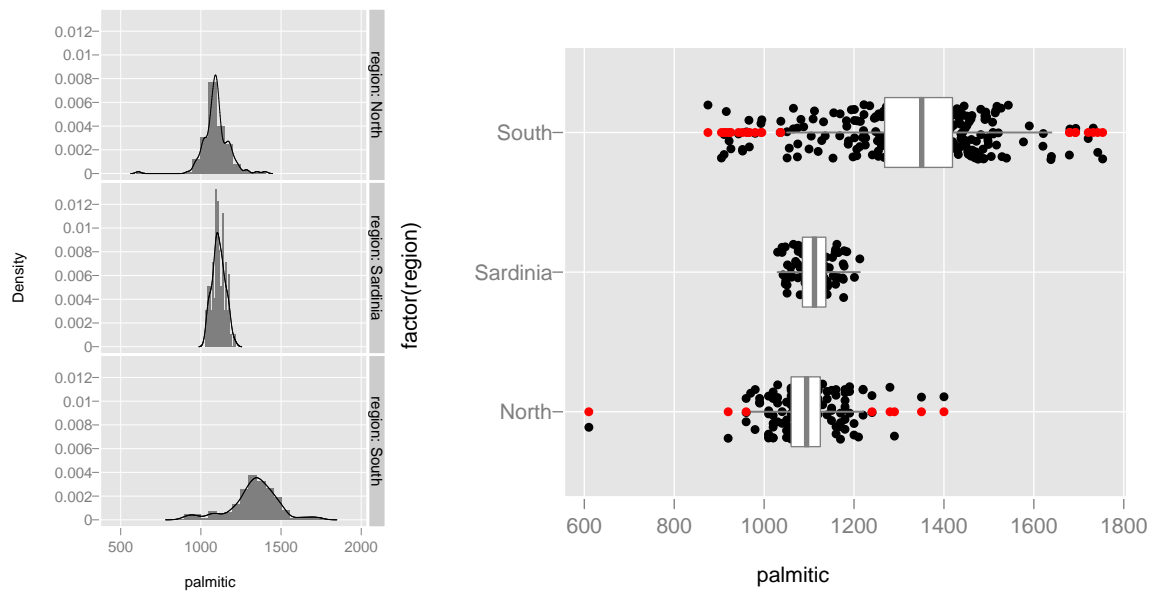
Background: The graphics in the olive oil case study are not suitable for high quality publication. We would like to produce better quality plots that elucidate the class/cluster structure in this data using R. One part of this is to use the ggplot package for low-dimensional plots, and the other part is to reproduce the four plots revealing higher dimensional interactions from ggobi and enhance them.

Purpose: To produce useful presentation graphics for classification data.

Exercises:

1. This code using the ggplot package produces the following histograms with density plots overlaid for the olive oil data:

```
library(ggplot)
qplot(palmitic, data=d.olive, facet=region~., type=c("histogram","density"))
qplot(palmitic,factor(region), data=d.olive, type=c("jitter","boxplot"))
```



These plots help to decide which variables are important to separate the three regions.

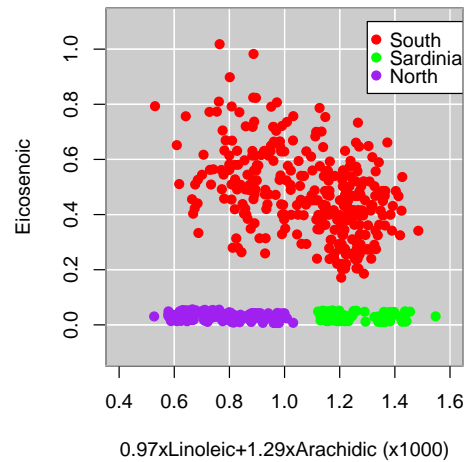
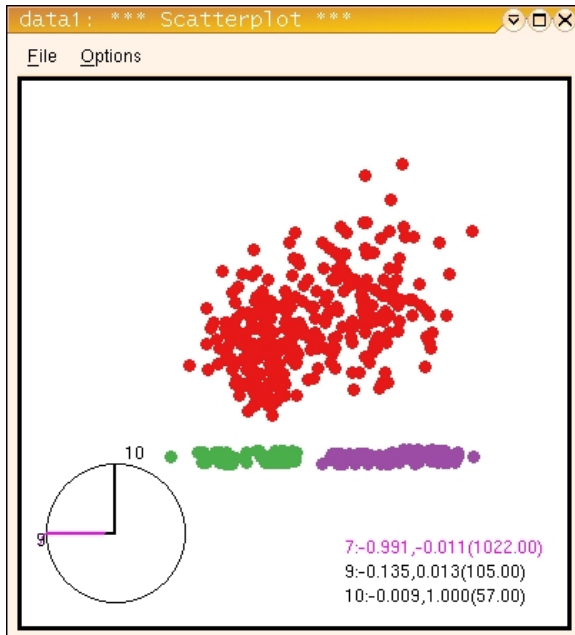
Your job is to do generate summary plots like these for all of the fatty acids for the three areas of the northern oils. Hand in your plots and your R code.

2. Using ggobi, we found a linear combinations of variables that separates the oils from the three regions. The ggobi screendump is below. We reconstructed the projected data using R and enhanced the plot with the following code:

```

coefs<-matrix(c(0,0,0,0,0,0,0,0,0,0.991/1022,-0.011/1022,0,0,
  0.135/105,0.013/105,0.009/57,1.000/57),ncol=2,byrow=T)
proj<-as.matrix(d.olive[,3:10])%*%coefs
par(pty="s")
plot(proj[,1],proj[,2],xlab="0.97xLinoleic+1.29xArachidic (x1000)",
  ylab="Eicosenoic",type="n",xlim=c(0.4,1.6),ylim=c(-0.1,1.1))
rect(0.3,-1,2,1.2,col="gray80")
abline(v=c(1.6,1.4,1.2,1.0,0.8,0.6),h=c(0.0,0.2,0.4,0.6,0.8,1.0),
  col="white",lty=1)
points(proj[1:323,1],proj[1:323,2],pch=16,col="red")
points(proj[324:421,1],proj[324:421,2],pch=16,col="green")
points(proj[422:572,1],proj[422:572,2],pch=16,col="purple")
legend(1.3,1.1,pch=16,col=c("red","green","purple"),
  legend=c("South","Sardinia","North"))
box(col="gray70")

```



Your job is to do something similar for the three areas of the northern oils. Find a linear combination of the fatty acids where the three areas (Umbria, East and West Liguria) are reasonably separated. Take a screendump of the plot window. Reconstruct this projection in R and enhance it similarly to the plot above. (You may be able to use ggplot instead of all of these lines of code.) Hand in the screendump, the R plot and your R code.