Visualization and Learning Structure: Problem Session 7/5/16
Solutions

We're looking at some variables from a real data set about mammographic masses (tumors in breast tissue). The data set has only been altered to remove missing values. **Remember to look at today's code to help you with the below problems.**

M. Elter, R. Schulz-Wendtland, and T. Wittenberg (2007). *The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process*, Medical Physics 34(11), p.4164-4172

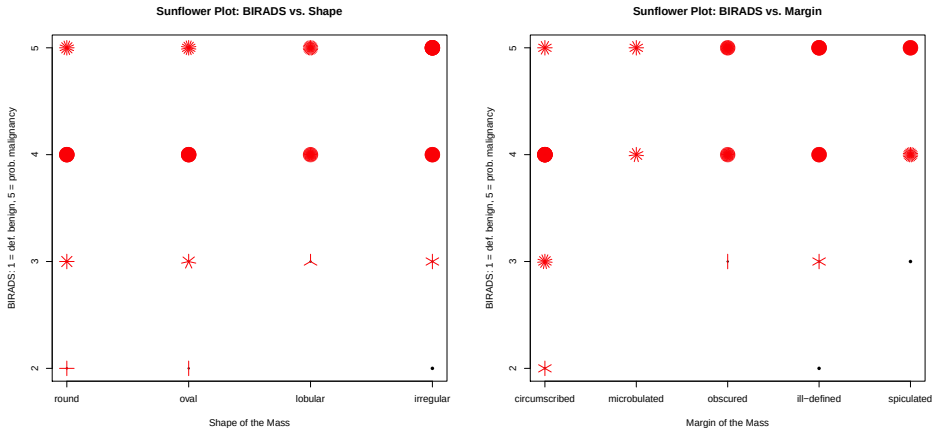
This data set is used to develop classification models for whether or not patients have a malignant mass without invasive biopsies. We have some mammography results for 816 patients whose lesions were later determined to be malignant or benign.

BIRADS is an assessment: 1 = definitely benign; 5 = highly suggestive of malignancy. *Age*: patient's age in years. *Shape*: 1 = round, 2 = oval, 3 = lobular, 4 = irregular. *Margin*: 1 = circumscribed, 2 = microlobulated, 3 = obscured, 4 = ill-defined, 5 = spiculated. *Density*: 1 = high, 2 = iso, 3 = low, 4 = fat-containing.
Severity gives the classification of the mass: 1 = malignant, 0 = benign.

Download *mammogram.txt* from the website (`http://www.stat.cmu.edu/~rnugent/PCMI2016`) and read it into R: `mammogram<-read.table("mammogram.txt")`.

```
data.c<-read.table("mammogram.txt",header=T)
attach(data.c)

names.shape = c("round", "oval", "lobular", "irregular")
names.margin = c("circumscribed", "microbulated", "obscured", "ill-defined","spiculated")
names.density = c("high", "iso", "low", "fat-containing")
names.severity = c("malignant", "benign")
label.BIRADS = "BIRADS: 1 = def. benign, 5 = prob. malignancy"
label.shape = "Shape of the Mass"
label.margin = "Margin of the Mass"
label.density = "Density of the Mass"
label.severity = "Severity: 1 = Malignant, 0 = Benign"
label.age = "Age of the Patient (Years)"
```

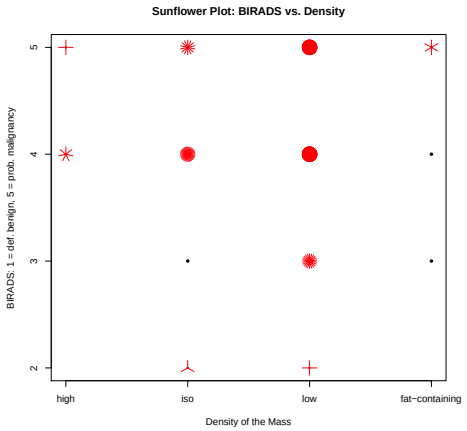
1. *The* Shape*,* Margin*, and* Density *variables are all ordered, as is* BIRADS*.*
   *Use sunflower plots to examine the relationship between the physician's assessment (*BIRADS*) and each of the three lesion descriptions. Describe the trends (if any). Which of the three seem to have the strongest relationship with* BIRADS*?*



BIRADS and Shape do appear to be related. In particular, for BIRADS = 2, 3, 4, round and oval shapes are very common. However, for BIRADS = 5, the highest physician's assessment score, round and oval shapes do not occur as frequently, while irregular and lobular shapes occur more frequently. It seems that for BIRADS scores that are "highly suggestive of malignancy", the mass is more likely to have non-round, more irregular shapes. While there does appear to be a relationship, it is not very strong, since the conditional distribution of shape given BIRADS is only different for BIRADS = 5.

BIRADS and Margin also appear to be related. Similarly to BIRADS vs. Shape, the conditional distribution of Margin given BIRADS is roughly equal for BIRADS = 2, 3, and 4 and differs greatly for BIRADS = 5. In particular, for BIRADS = 2, 3, and 4, circumscribed and ill-defined margins occur more frequently, while spiculated margins occur less frequently. For BIRADS = 5, spiculated margins occur much more frequently. For BIRADS = 5, there appears to be an "increasing" trend in the margin - that is, the margin "increases" from circumscribed to spiculated. Again, there is a not very strong relationship; the relative frequencies of each Margin category are roughly the same for BIRADS = 1-4, and only different for BIRADS = 5.

BIRADS and Density do not appear to be strongly related. The relative frequencies of each Density category are roughly the same for all values of BIRADS. It is difficult to determine if there are any definite trends in these two variables. Of the three lesion descriptions, it appears that Shape and Margin have the strongest relationships with BIRADS, while Density does not.



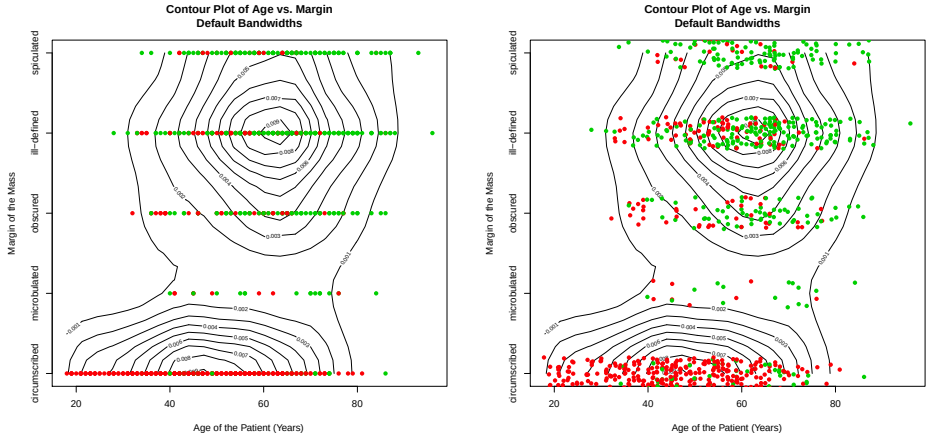

```
library(graphics)
?sunflowerplot

sunflowerplot(Shape, BIRADS, main = "Sunflower Plot: BIRADS vs. Shape",
ylab = label.BIRADS,xlab = label.shape, xaxt = 'n', yaxt = 'n')
axis(1, at = 1:4, labels = names.shape)
axis(2, at = 2:5, labels = 2:5)

sunflowerplot(Margin, BIRADS, main = "Sunflower Plot: BIRADS vs. Margin",
ylab = label.BIRADS,xlab = label.margin, xaxt = 'n', yaxt = 'n')
axis(1, at = 1:5, labels = names.margin)
axis(2, at = 2:5, labels = 2:5)

sunflowerplot(Density, BIRADS, main = "Sunflower Plot: BIRADS vs. Density",
ylab = label.BIRADS, xlab = label.density, xaxt = 'n', yaxt = 'n')
axis(1, at = 1:4, labels = names.density)
axis(2, at = 2:5, labels = 2:5)
```

2. Create a contour plot (default *nlevels*) of a two-dimensional kernel density estimate (default bandwidth) for *Age* and *Margin*. (`library(MASS); kde2d`) Add the observations to the graph using `points() (pch = 16)`, color-coded by the severity of the masses: malignant = green; benign = red.

   You can create a vector of colors by using ifelse():
   `col.vec<-ifelse(Severity==1,"green","red")`; then use `col=col.vec`

   (a) How many modes does the density estimate have? Given the integer-valued variables, why do we still see "circular/oval" modes in the density estimate?

   Can you characterize this relationship by whether the mass was malignant?



For both graphs, the density estimate is calculated for the original Age and Margin values (note that Margin is an unordered variable). However, the left graph overlays the original values on the top of the density estimate; the right graph overlays jittered values in order to better visualize the frequency of values and their subsequent effect on the modes of the density estimate.

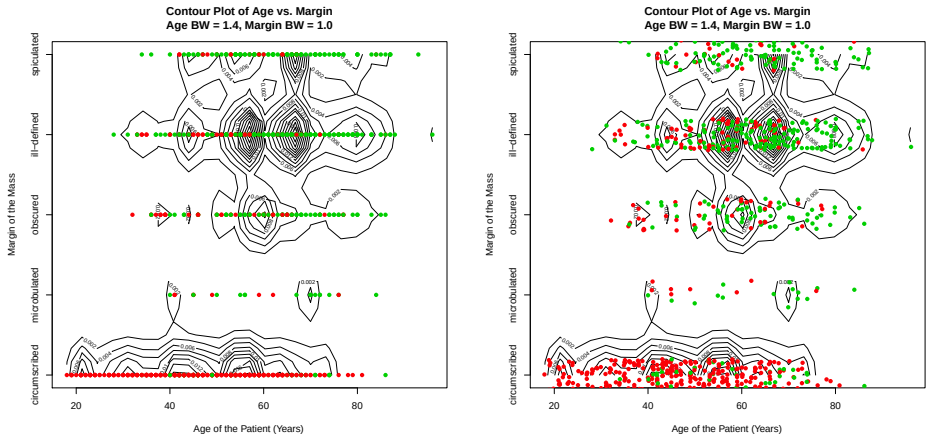
This density estimate of Age and Margin has two modes. The first is centered at about 60 years and at an "ill-defined" margin, the second at about 45 years and a circumscribed margin. We still see circular / oval modes in this "striped" data because the density estimate uses a Gaussian kernel, which puts a normal bell curve on top of each point. Because the Gaussian kernel has infinite support, the density estimate smoothes out non-zero mass in the areas between the modes. Recall the concentric contour lines represent the "height" or magnitude of the density estimate at that location. Sets of small concentric circles/contours indicate the presence of multiple modes. Larger contours indicate the level set/cross-section at lower heights. The number of points in each "stripe" will influence the height of any related modes. If the bandwidth were very small, smoothing would be less apparent and the valley(s) between modes would be much deeper. Larger bandwidth = more smoothing.

Malignant masses seem to be more common in margins that are microbulated, obscured, ill-defined, or spiculated, while benign masses are most common in circumscribed masses. Also, patients with benign tumors may be younger than those with malignant tumors. If a mass is benign, it is likely circumscribed, and its patient is likely younger than 70 years old. If a mass is malignant, it is difficult to determine the margin (other than being able to say that it is likely not circumscribed), and its patient is likely older than 40 years old.

```
library(MASS)
?contour
?kde2d
d4a = kde2d(Age, Margin)
contour(d4a, main = "Contour Plot of Age vs. Margin\nDefault Bandwidths",
xlab = label.age, ylab = label.margin, yaxt = 'n')
points(Age, Margin, col = Severity+2, pch = 16)
#repeat graph with jitter(Age), jitter(Margin)
axis(2, at = 1:5, labels = names.margin)
```

(b) Experiment with the bandwidths to find a density estimate that better reflects the "striped" nature of the data (include your graph). What do you choose?

Which dimension dictates whether or not there are density estimate valleys?



Again, both graphs calculate the density estimate with the original values. The right graph has jittered observations for visualization purposes.

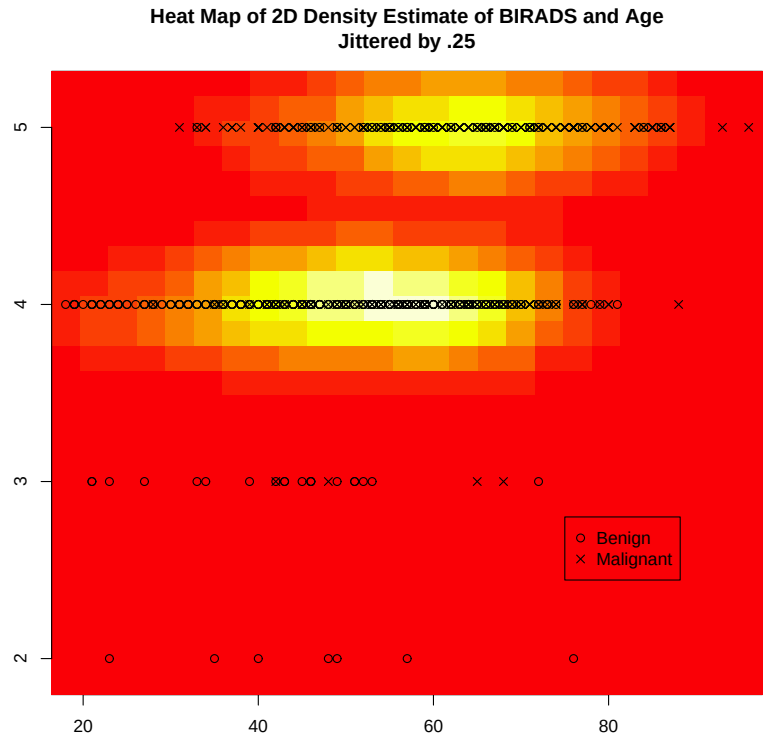
Here we chose the bandwidth combination of 1.4 for Age and 1.0 for Margin. Margin's 1.0 band- width shows the striped nature of the categorical data, while still smoothing out some of the higher- density areas between margin values in the density estimate. Other bandwidths are also suitable here. Anything less than about 1.5 for the Margin bandwidth shows the striped nature of the data well.

The categorical Margin dimension dictates whether or not there are valleys in the density estimate, because Margin is a discrete variable. Because Age is continuous, we would only see valleys - or low-density areas - in the density estimate if there was some definite group separation in the data. Note that if we reduced the bandwidth enough, we would get contours that completely isolate the margin stripes.

```
d4b = kde2d(Age, Margin, h = c(1.4, 1.0))
contour(d4b, main = "Contour Plot of Age vs. Margin\nAge BW = 1.4, Margin BW = 1.0",
xlab = label.age, ylab = label.margin, yaxt = 'n')
points(Age, Margin, col = Severity+2, pch = 16)
#repeat graph with jitter(Age), jitter(Margin)
axis(2, at = 1:5, labels = names.margin)
```

3. We're interested in the relationship between *Age* and *BIRADS* with respect to *Severity* but are concerned that the ordinal *BIRADS* variable will make it difficult to see any possible trend. Jitter both *Age*, *BIRADS* by 0.25 (`help(jitter)`), and create a density estimate of the jittered variables using the default parameters.

   Create a heat map of this density estimate. Add the observations to the heat map using `points()` (`col=1`) where the patients are pch-coded according to severity of the mass: malignant = x (`pch=4`); benign = o (`pch=1`). Can use `pch.vec<-ifelse(Severity==1,4,1)`, then `pch=pch.vec`.




```
d5a = kde2d(jitter(Age, amount = .25), jitter(BIRADS, amount = .25))
image(d5a, main = "Heat Map .....", yaxt='n')
axis(2, at = 2:5, labels = 2:5); pch.vec<-ifelse(Severity==1,4,1)
points(Age, BIRADS, col = 1, pch = pch.vec)
legend(75, 2.8, c("Benign", "Malignant"), pch = c(1,4))
```

(a) Describe any high frequency areas.
Do they correspond more commonly to malignant or benign masses?

There are two high-frequency areas. The first is at BIRADS = 4, age between about 30 to about 70. The second is at BIRADS = 5, age between about 40 and about 75. These are marked on the heat map by yellow / white areas; red and orange areas are low-density areas. It appears that the high-density areas correspond more commonly to malignant masses than benign masses.

(b) Is there a relationship between *Age* and *BIRADS*? Is it dependent on *Severity*?

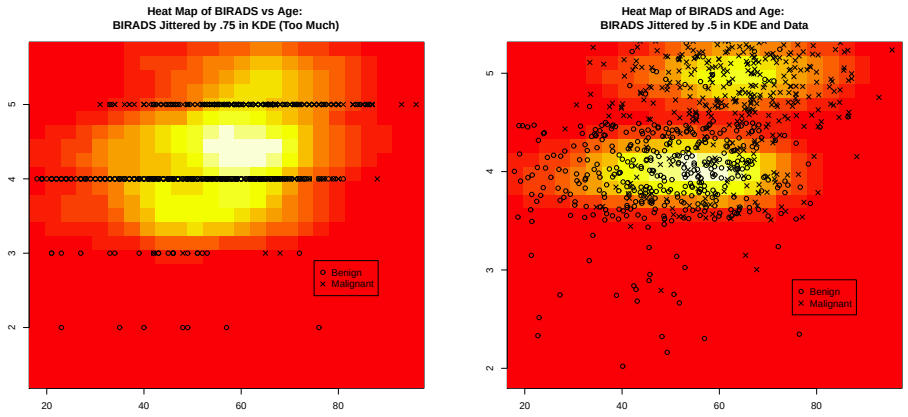
It appears that as age increases, BIRADS also increases although the bandwidth was a bit too large to show more of the local features of the 2D density estimate. It also appears that malignant masses are associated more with high BIRADS scores and patients older than about 50, while benign masses are associated with lower BIRADS scores and patients younger than about 50. We know this because for BIRADS = 5, we see mostly Xs above the high density area and very few Os. The Xs correspond to malignant tumors. The data points for benign tumors or for patients younger than about 40 are almost all marked with Os, indicating that these tumors are usually benign.

(c) What is the maximum amount that we could jitter each variable without changing the structure of the data?

Jittering Age by increasing amounts does not unduly change the structure of the data. It will widen the range of the data and smooth out the density estimate, but it will still be a continuous variable that ranges from about 20 to about 80, plus or minus the jitter amount. However, we would not want to jitter the age values enough that the order of the values is completely disrupted.

Jittering BIRADS by anything less than .5 will keep the categorical/striped structure of the data. While the variable won't appear entirely categorical, the stripes will remain, and each datapoint will be close to its original category.

However, jittering BIRADS by .5 or more will merge/overlap the integer-valued categories and create a more continuous-looking BIRADS variable. Data points may be jittered enough as to make them appear to belong to other categories. As a result, BIRADS will no longer appear to be a discrete variable or a variable with distinct groups - it will become more continuous with less group structure centered around the original integer values. See examples below.




```
d5c = kde2d(jitter(Age, amount = .25), jitter(BIRADS, amount = .75))
image(d5c, main = "Heat Map of BIRADS vs Age: \nBIRADS Jittered by .75
in KDE (Too Much)", yaxt='n')
axis(2, at = 2:5, labels = 2:5)
points(Age, BIRADS, col = 1, pch = Severity*3+1)
legend(75, 2.9, c("Benign", "Malignant"), pch = c(1,4))

image(d5a, main = "Heat Map of BIRADS and Age: \nBIRADS Jittered by .5
in KDE and Data", yaxt='n')
axis(2, at = 2:5, labels = 2:5)
points(jitter(Age, amount = .5), jitter(BIRADS, amount = .5), col = 1,
pch = Severity*3+1)
legend(75, 2.9, c("Benign", "Malignant"), pch = c(1,4))
```

4. Experiment with using `image()` (for example) to visualize matrices you're creating in your wavelet work. In particular, experiment with visualizing "cross-sections" or "level sets" of your matrix by using something like the following:

```
yourmatrix<-read.csv("yourmatrix.csv")
##or read.table("yourmatrix.txt") depending on how you stored it
image(yourmatrix)
levelset<-ifelse(yourmatrix>threshold,1,0)
##you pick the threshold number
image(levelset)
```