# Supervised Classification

A training sample is used to build the rules
to accurately predict a categorical response.


# Example

## Italian olive oils

Number of samples: 572

Number of variables: 10

Super-classes, 3 regions, and 8 classes, areas within region.

Explanatory variables are % fatty acids in the sample: palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic

*How do we distinguish the oils from different regions and areas in Italy based on their combinations of the fatty acids?*

# Visual classification

Code the response using color and symbol, explore a variety of plots of the explanatory variables, to learn how distinctions between classes arise from the explanatory variables.

# Strategy

Work from plots of single variables up to plots of multiple variables.

Work from large groups to small groups.

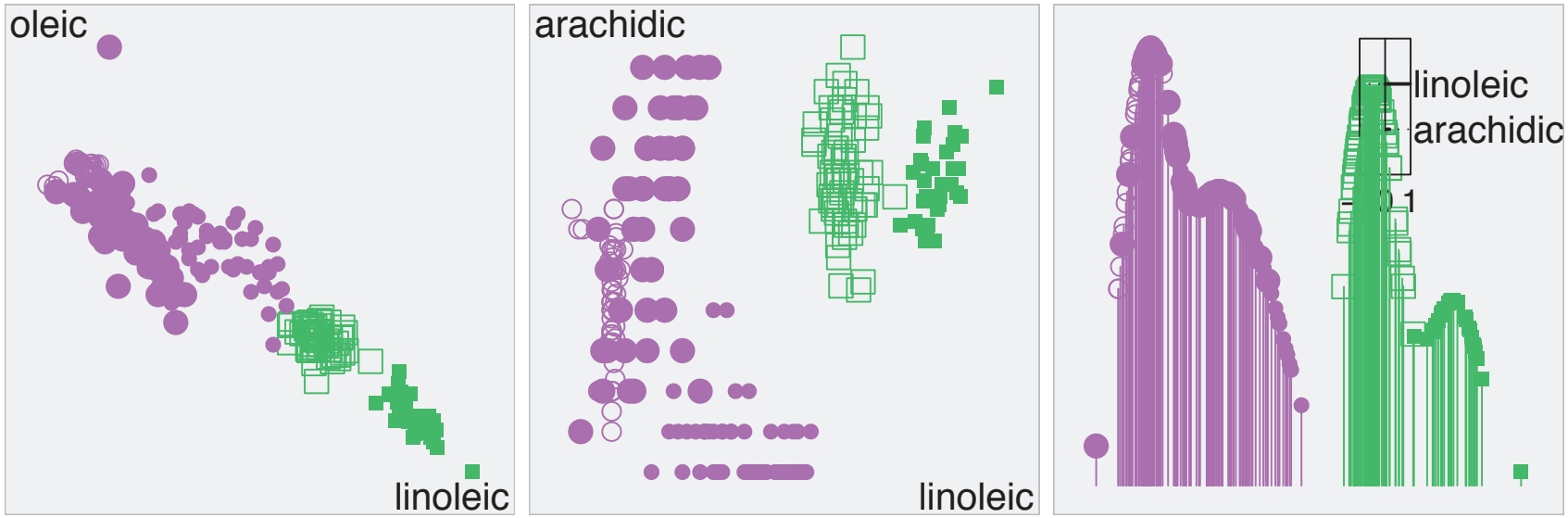
After separating out one group, focus on the remaining.

# One variable


Only oils from southern Italy have detectable amount of eicosenoic acid.

Linoleic acid, and oleic, contribute to distinguishing north from Sardinia.

# Two variables




Oils from north and Sardinia can be distinguished by oleic and linoleic acid, but much better using linoleic and arachidic. A linear separation can be obtained by taking a projection of linoleic and arachidic.

# Your turn

- Subset the data to the oils from the north only.
- Find which fatty acids distinguish the oils from the three areas, Umbria, East and West Liguria.

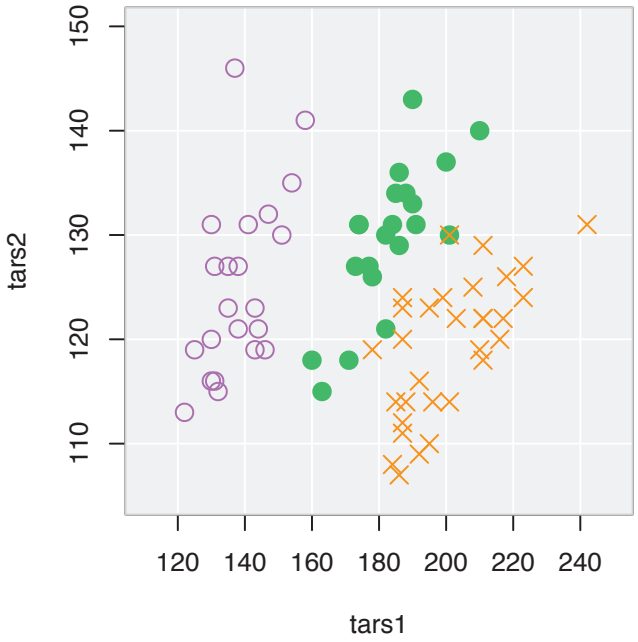
- Note: These are not as neatly separated as the super-classes.

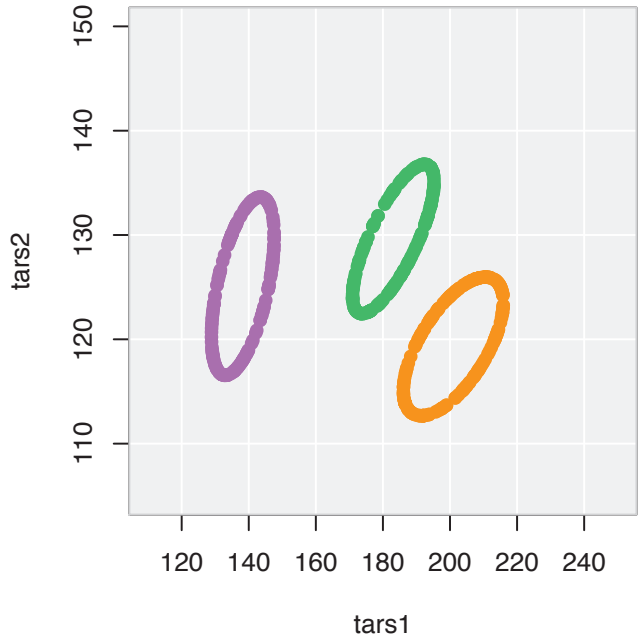
# Numerical methods

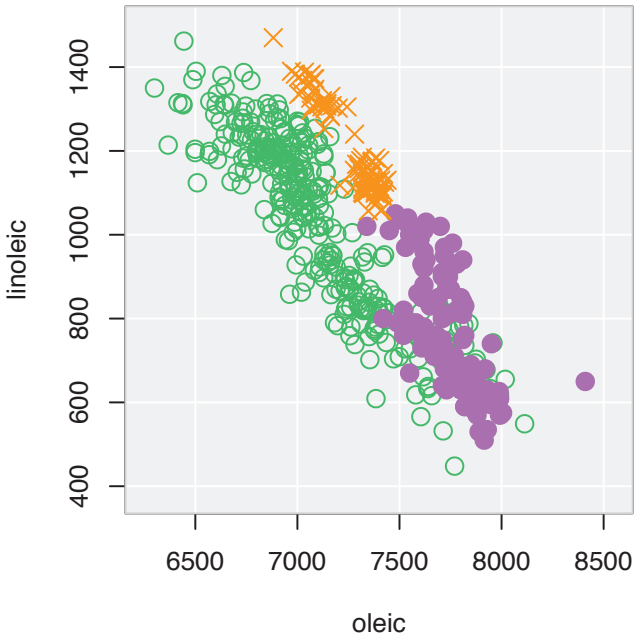
- Classical (Statistical): More parametric/ explicit assumptions, some guarantees if assumptions true. e.g. linear discriminant analysis
- Algorithmic (Data mining): More heuristic, implicit assumptions. e.g. trees, random forests

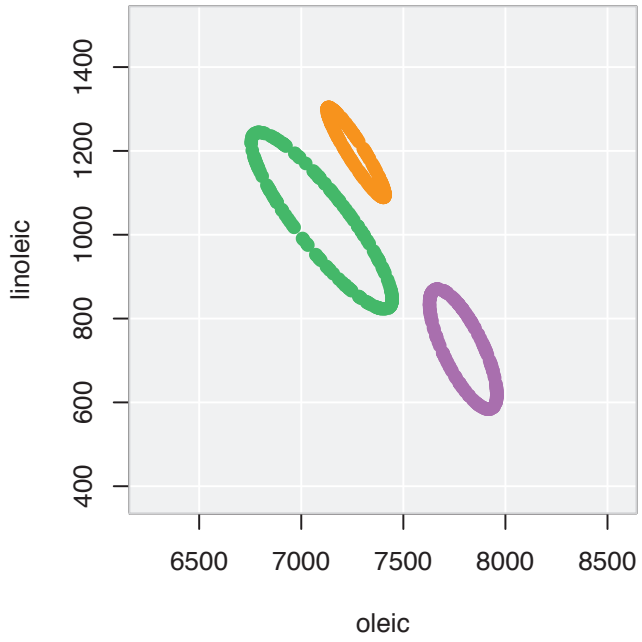
# How can graphics help?

- Classical: Check assumptions such as whether the samples are consistent with a multivariate normal distribution.
- Algorithmic: Open the black box, to learn if the algorithm matches the class structure.
- Both: Assess the predictions, and accuracy of the rules.



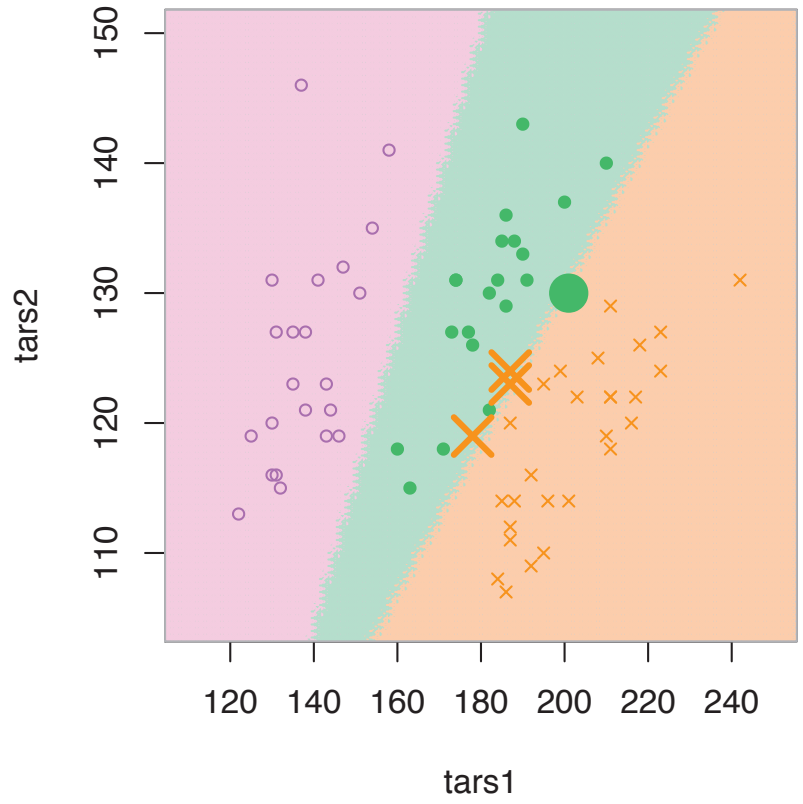



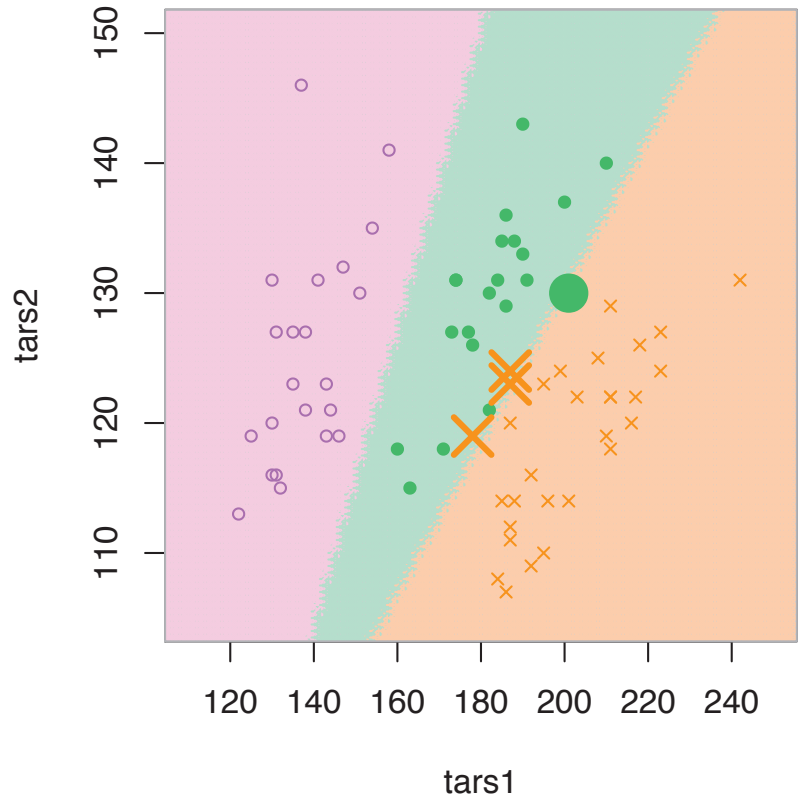

# Classical

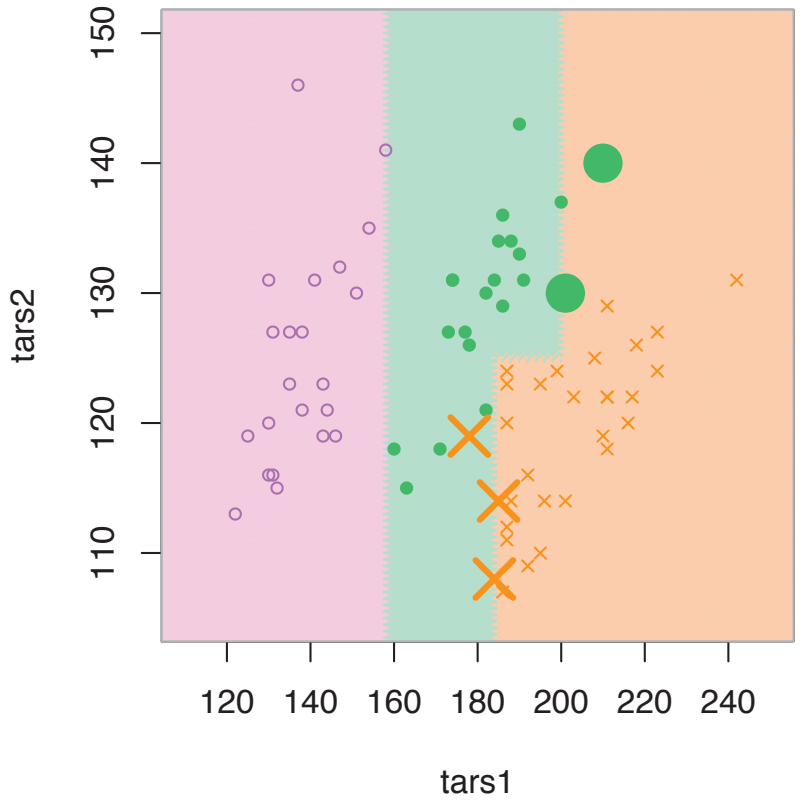
Normal assumption, elliptical variance-covariance, equal for each class.

Olive oils data doesn't follow this model.




Classical linear discriminant analysis: boundaries are consistent with the shape and orientation of the class clusters. Errors occur along the boundary between the two regions.




Classical linear discriminant analysis: boundaries are consistent with the shape and orientation of the class clusters. Errors occur along the boundary between the two regions.

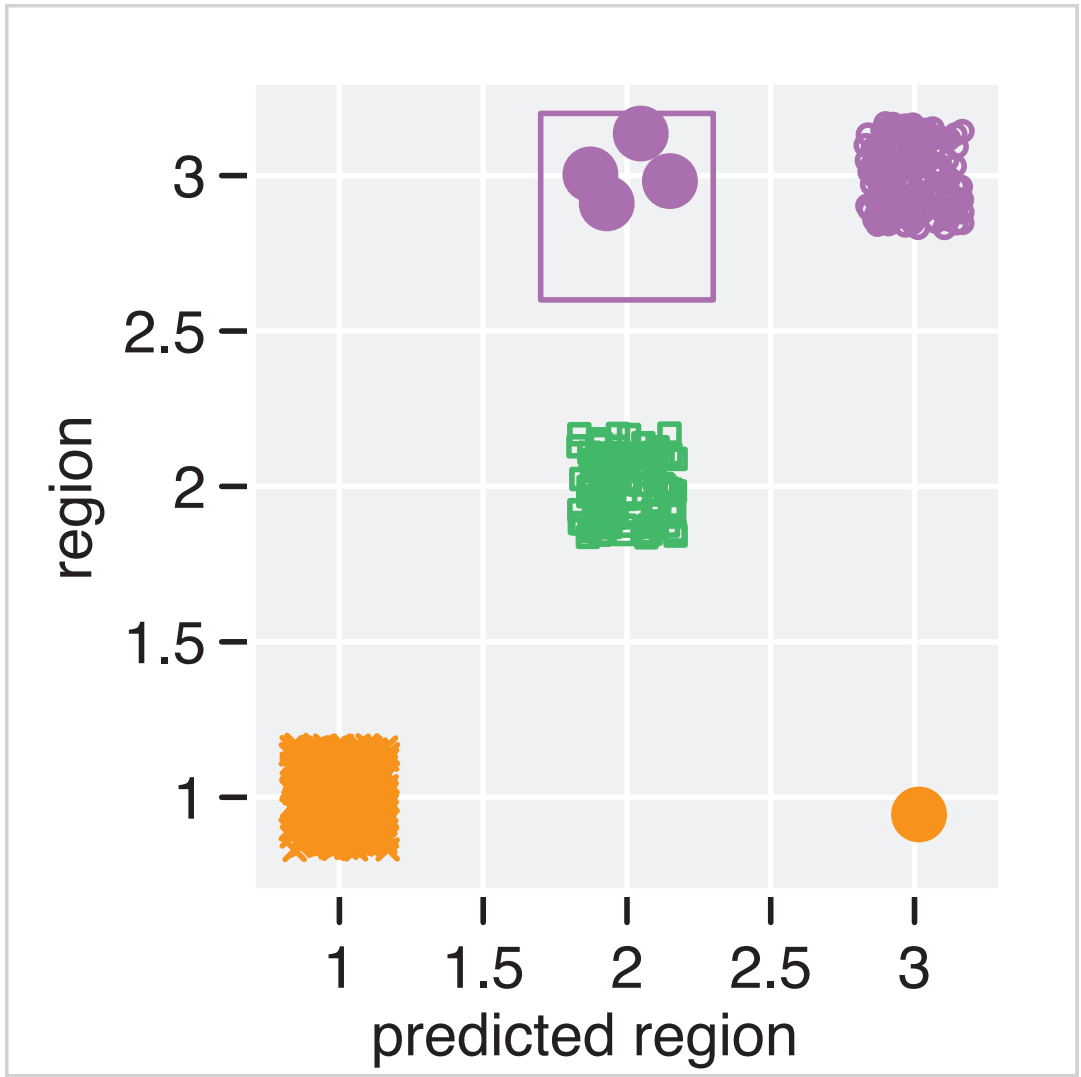
Tree: boundaries are in horizontal or vertical direction. Errors occur because cluster shape is ignored.

# Linear Discriminant Analysis

```
> library(MASS)
> library(rggobi)
> d.olive <- read.csv("olive.csv", row.names=1)
> d.olive.sub <- subset(d.olive,
   select=c(region,palmitic:eicosenoic))
> olive.lda <- lda(region~., d.olive.sub)
> pregion <- predict(olive.lda, d.olive.sub)$class
> table(d.olive.sub[,1], pregion)
   pregion
> plot(predict(olive.lda, d.olive.sub)$x)
> gd <- ggobi(cbind(d.olive, pregion))[1]
> glyph_color(gd) <- c(rep(6,323), rep(5,98), rep
(1,151))
```

| | | Predicted region | | | Error |
|---|---|---|---|---|---|
| | | South | Sardinia | North | |
| region | South | 322 | 0 | **1** | 0.003 |
| | Sardinia | 0 | 98 | 0 | 0.000 |
| | North | 0 | **4** | 147 | 0.026 |
| | | | | | 0.009 |

# LDA: Olive oils




Classes do not have equal, or elliptical shape.

This leads to unfortunate misclassifications.

There shouldn't be any error, based on what we learned from the initial graphical classification.

# Trees

```
> library(rpart)
> olive.rp <- rpart(region~., d.olive.sub,
method="class")
> olive.rp
```

Rule:

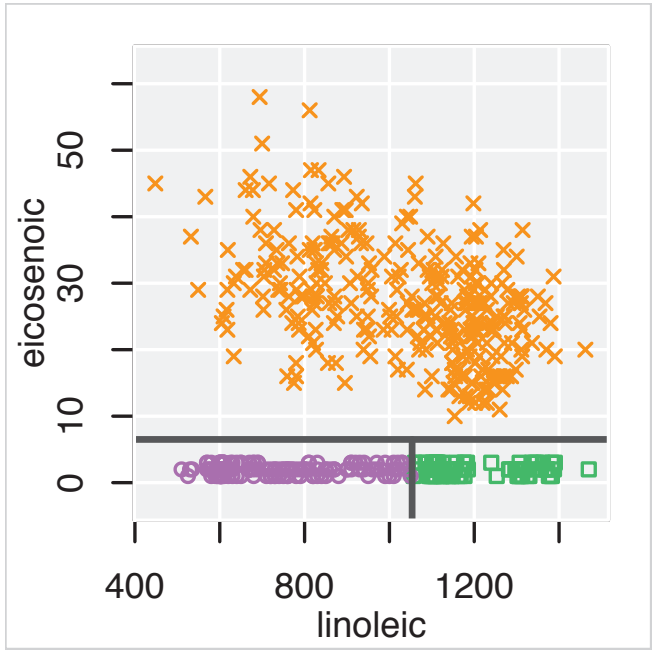
*if eicosenoic >= 6.5 assign the sample to South*
*else*
*if linoleic >= 1053.5 assign the sample to Sardinia*
*else assign the sample to North*

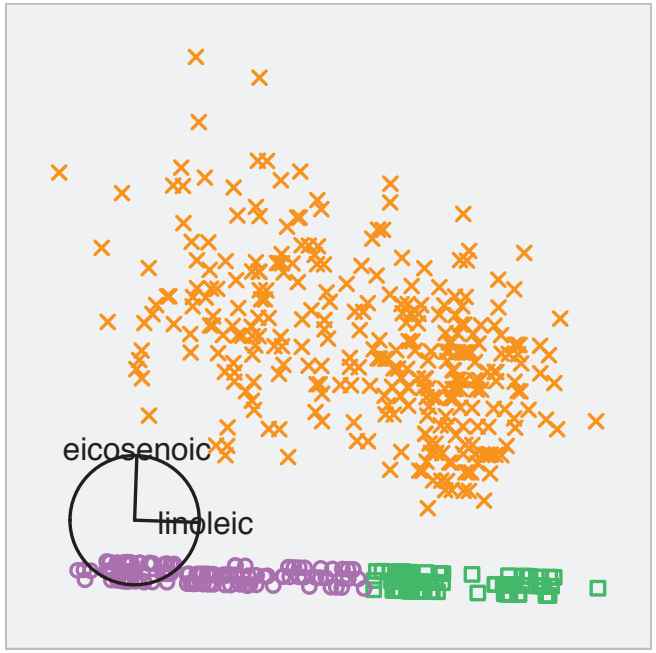
Error: 0

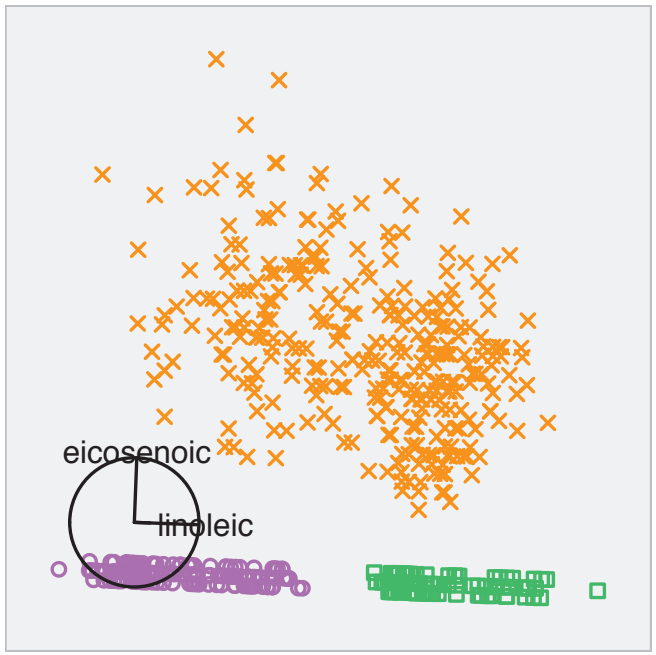
Looks good!

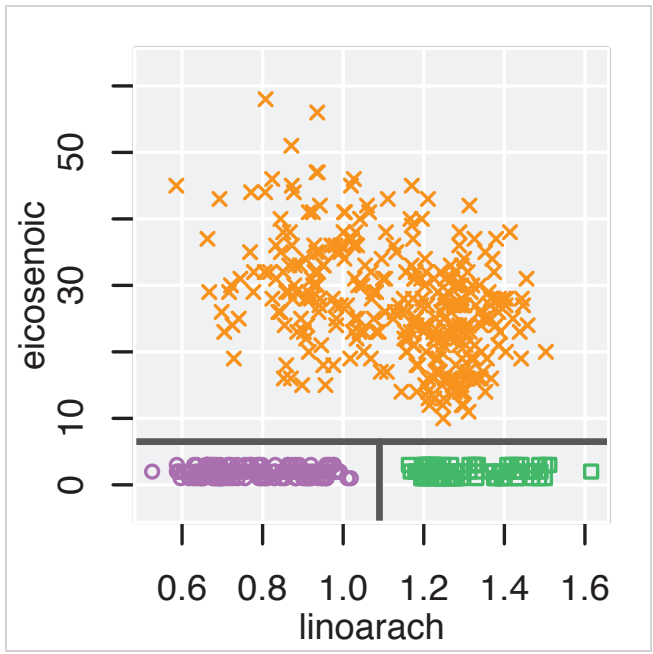
Boundary between north and Sardinia is very tight.

A bigger gap obtained by creating new variable using linoleic and arachidic acids.

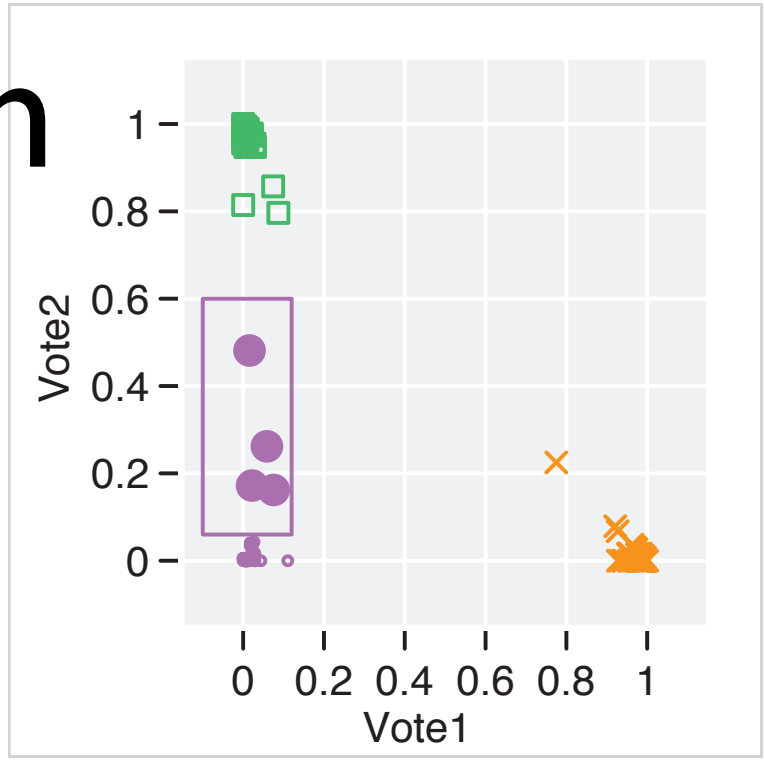


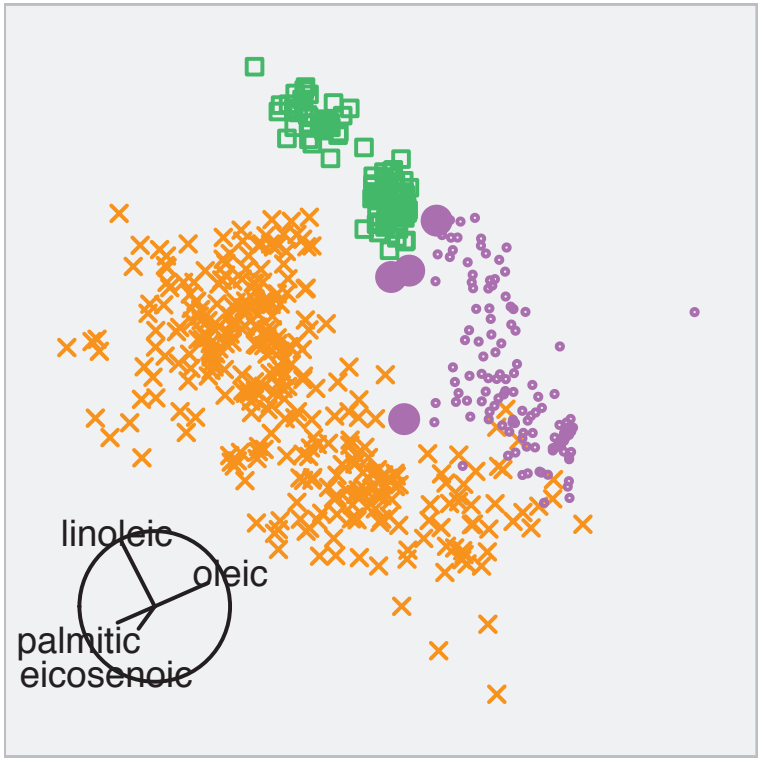



# Random forests

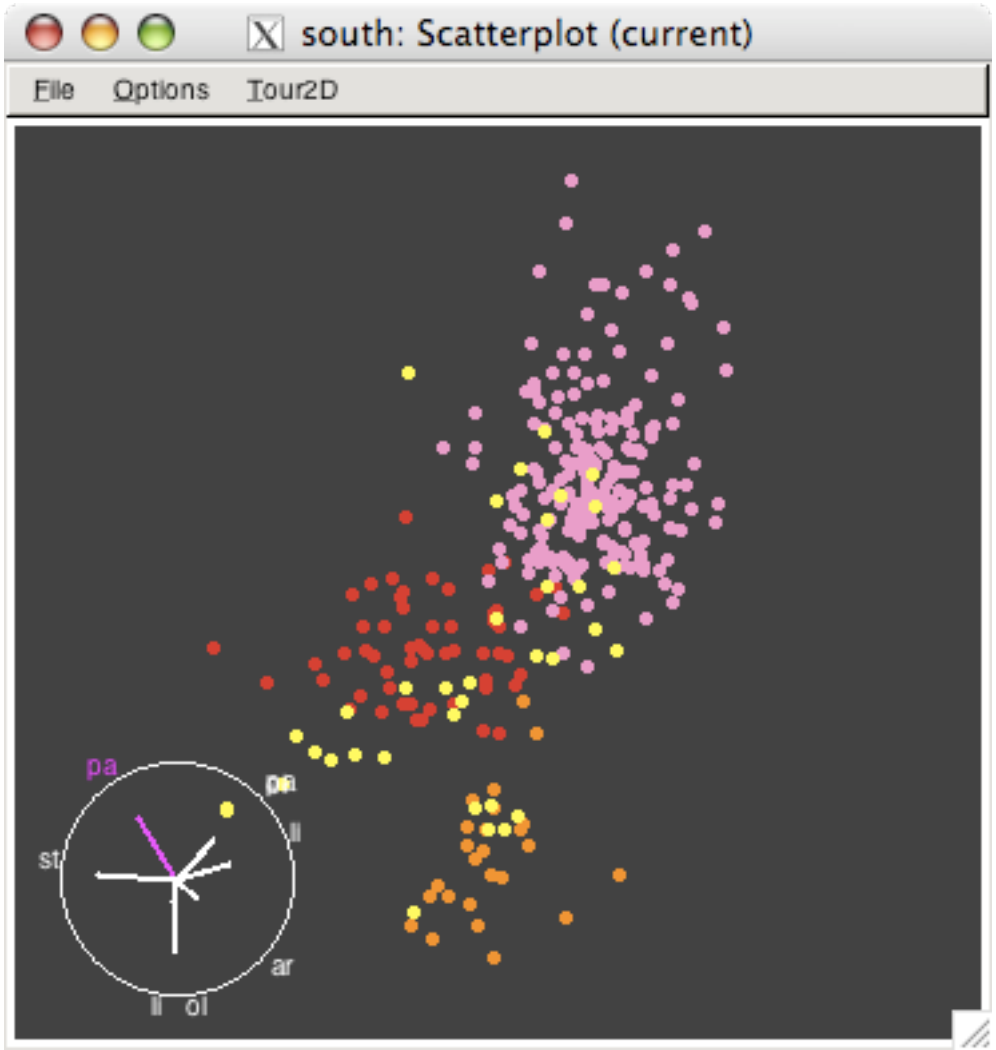



Random forests are made out of trees - so they have the same problems.

But they open the black box a little with diagnostics including importance and votes


NORTH
Liguria
Umbria
Apulia
Sardinia
Calabria
Sicily
SOUTH


south: Scatterplot (current)
File
Options
Tour2D
pa
pn
ll
ar
ol
st

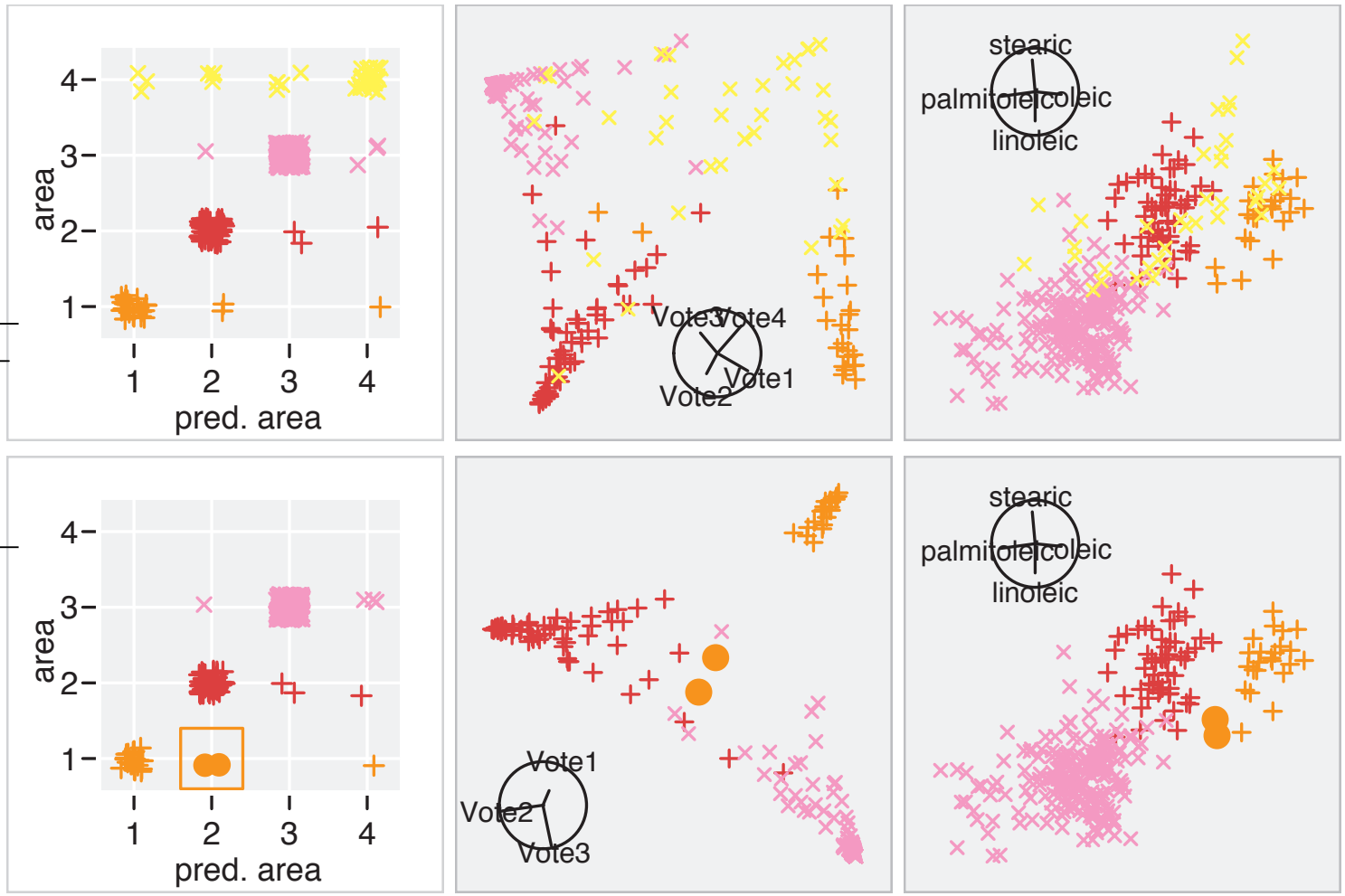
# Random forests: more difficult task of classifying southern oils

```
> d.olive.sth <- subset(d.olive, region==1,
   select=area:eicosenoic)
> olive.rf <- randomForest(as.factor(area)~.,
   data=d.olive.sth, importance=TRUE, proximity=TRUE,
   mtry=2, ntree=1500)
> order(olive.rf$importance[,5], decreasing=T)
[1] 5 2 4 3 1 6 7 8
```

| | | Predicted area | | | | Error |
|---|---|---|---|---|---|---|
| | | North Apulia | Calabria | South Apulia | Sicily | |
| area | North Apulia | 22 | **2** | 0 | **1** | 0.120 |
| | Calabria | 0 | 53 | **2** | **1** | 0.054 |
| | South Apulia | 0 | **1** | 202 | **3** | 0.019 |
| | Sicily | **3** | **4** | **5** | 24 | 0.333 |
| | | | | | | 0.068 |

```
> pred <- as.numeric(olive.rf$predicted)
> margin <- olive.rf$vote
> colnames(margin) <- c("Vote1", "Vote2", "Vote3", "Vote4")
> d.olive.rf <- cbind(pred, margin, d.olive.sth)
> gd <- ggobi(d.olive.rf)[1]
> glyph_color(gd) <- c(6,3,2,9)[d.olive.rf$area]
```

| area | symbol |
|---|---|
| North Apulia | orange + |
| Calabria | red + |
| South Apulia | pink × |
| Sicily | yellow × |



# Data summary

- We can accurately classify the oils from the three broad regions
- Classifying the Southern oils is harder, but once we remove the Sicilian oils it is much easier

# Summary

- Graphics help us understand:
  - Our data: so we can see if a method is doing something stupid, or give it extra information to help it out
  - Our methods: so we can understand what assumptions they make
- Same graphics used for analysis and diagnosis
- Combination of analysis and visualisation
- Start with low-D views, then get more complicated

# Your turn

For the Australian crabs data:

From univariate plots assess whether any individual variables are good classifiers of crabs by species or sex.

From either a scatterplot matrix or pairwise plots, determine which pairs of variables best distinguish the crabs by Species and by sex within species.

Using TourI1D (and perhaps projection pursuit with the LDA index), find a 1D projection that mostly separates the crabs by species. Report the projection coefficients.

Now transform the five measured variables into principal components and run TourI1D on these new variables. Can you find a better separation of the crabs by species?

Fit a random forest to the crabs. Which variables are most important? For which cases are the predictions more uncertain, according to the vote matrix?

# Timeline

| | |
|---|---|
| 20 mins | Toolbox |
| 30 mins | Missing values |
| 45 mins | Supervised Classification |
| Break | |
| 45 mins | Unsupervised Classification |
| 30 mins | Inference |