MANOVA and Linear Discriminant Analysis

DS705

Multivariate Data

- usually observe more than one variable

```
head(iris) # data built into R
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

- each row is called a case

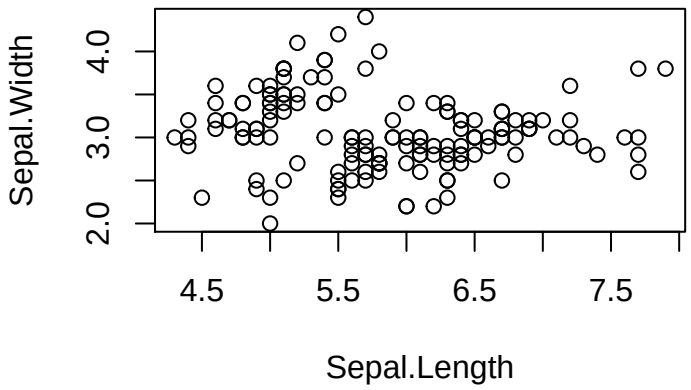
Multivariate Data Matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix}$$

- n is number of units
- each observation is a vector of q measurements on a unit, this vector is one row in the matrix
- x_{ij} is value of the jth variable for the ith unit

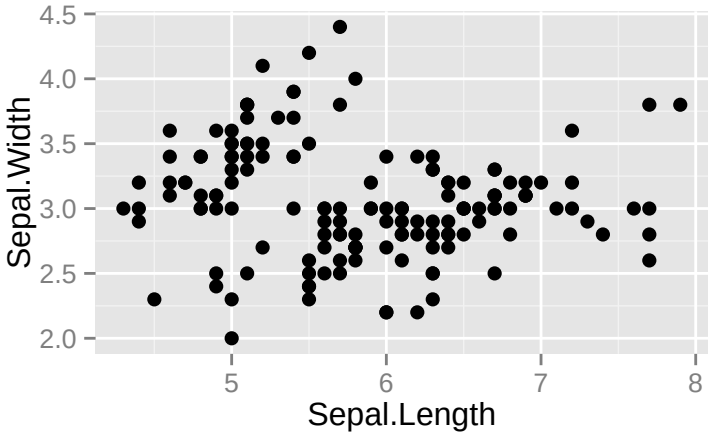
Scatterplots

```
with(iris,plot(Sepal.Length,Sepal.Width))
```



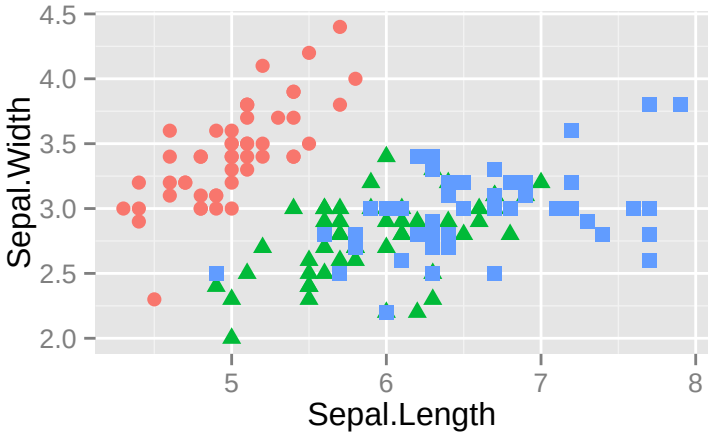
Scatterplots (2)

```
require(ggplot2)
ggplot(iris) + geom_point(aes(x=Sepal.Length,y=Sepal.Width),size=2.5)
```



Scatterplots (3)

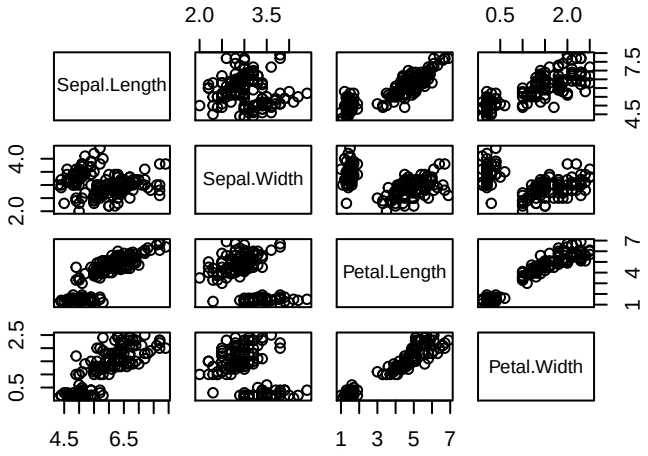
```
ggplot(iris) + theme(legend.position='none') + geom_point(aes(x=
  Sepal.Length,y=Sepal.Width,color=Species,shape=Species),size=2.5)
```



Scatterplot Matrix - code

```
pairs(~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width,data=iris)
```

Scatterplot Matrix - the plot



Summarizing Multivariate Data

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa     :50
##  versicolor:50
##  virginica :50
```

Column Means

```
apply( iris[,-5], 2, mean)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##     5.843333     3.057333     3.758000     1.199333
```

```
colMeans( iris[,-5] )
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##     5.843333     3.057333     3.758000     1.199333
```

Column Variances

```
apply( iris[,-5], 2, var)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.6856935    0.1899794    3.1162779    0.5810063
```

```
var( iris[,-5])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
## Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
## Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
## Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

Population Covariance Matrix

$$\sigma_{ij} = \mathrm{Cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)]$$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1} & \sigma_{q2} & \cdots & \sigma_{qq} \end{pmatrix}$$

Sample Covariance Matrix

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)$$

x_{ik} = the kth observation of variable x_i
x_{jk} = the kth observation of variable x_j

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1q} \\ s_{21} & s_{22} & \cdots & s_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ s_{q1} & s_{q2} & \cdots & s_{qq} \end{pmatrix}$$

Example Sample Covariance Matrix

```
x <- c(9,11,13,18,19)
y <- c(19,17,13,4,7)
sum( (x - mean(x))^2 )/(5-1)
```

```
## [1] 19
```

```
sum( (y - mean(y))^2 )/(5-1)
```

```
## [1] 41
```

```
sum( (x - mean(x))*
       (y - mean(y)) )/(5-1)
```

```
## [1] -27
```

```
cov(cbind(x,y))
```

```
##      x   y
## x   19 -27
## y -27  41
```

Sample Correlation Matrix

$$\text{cor}(x_i, x_j) = r_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} \frac{(x_{ik} - \overline{x}_i)}{s_i} \frac{(x_{jk} - \overline{x}_j)}{s_j}$$

x_{ik} = the kth observation of variable x_i
x_{jk} = the kth observation of variable x_j

Note: $r_{ii} = 1$

Example Sample Correlation Matrix

```
x <- c(9,11,13,18,19)
y <- c(19,17,13,4,7)
sx <- sd(x); mx <- mean(x)
sy <- sd(y); my <- mean(y)
1/(5-1)*sum((x-mx)^2)/(sx*sx)
```

```
## [1] 1
```

```
1/(5-1)*sum((y-my)^2)/(sy*sy)
```

```
## [1] 1
```

```
1/(5-1)*sum(((x-mx)/sx)*
              ((y-my)/sy))
```

```
## [1] -0.9673754
```

```
cor(cbind(x,y))
```

```
##            x          y
## x  1.0000000 -0.9673754
## y -0.9673754  1.0000000
```

Compare Covariance and Correlation

```
cov(cbind(x,y))
```

```
##     x   y
## x  19 -27
## y -27  41
```

```
cor(cbind(x,y))
```

```
##            x          y
## x  1.0000000 -0.9673754
## y -0.9673754  1.0000000
```

- Covariance matrix is unstandardized correlation matrix
- Divide covariance matrix rows and columns by each standard deviation

Multivariate Normal Distribution

$$\mathbf{x} \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

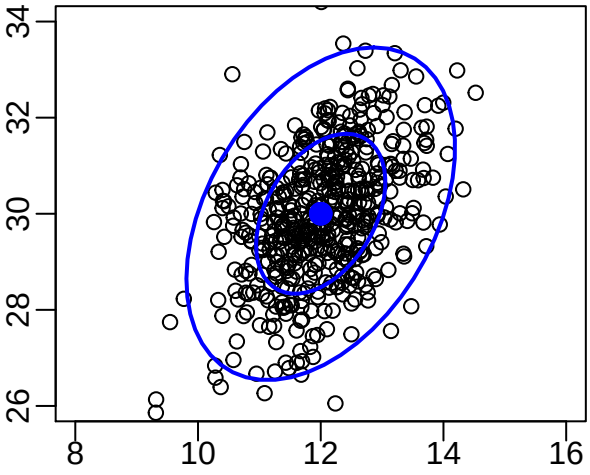
- $\mathbf{x}$ is a vector of q numbers
- $\boldsymbol{\mu}$ is the population mean vector of length q
- $\boldsymbol{\Sigma}$ is the $q \times q$ population covariance matrix

Example

```
require(MASS)
mu <- c(12,30); Sigma <- rbind( c(.8,.5),c(.5,2) )
x <- mvrnorm(1000,mu,Sigma)
plot(x[,1],x[,2],xlim=c(8,16),ylim=c(26,34))
ellipse(mu,Sigma,sqrt(qchisq(.5,2)),col='blue')
ellipse(mu,Sigma,sqrt(qchisq(.95,2)),col='blue')
```

- Plot on next page.

Example Plot



Mahalanobis Distance

- Multivariate version of "how many standard deviations from the mean?"
- Idea: "divide" by the covariance matrix

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

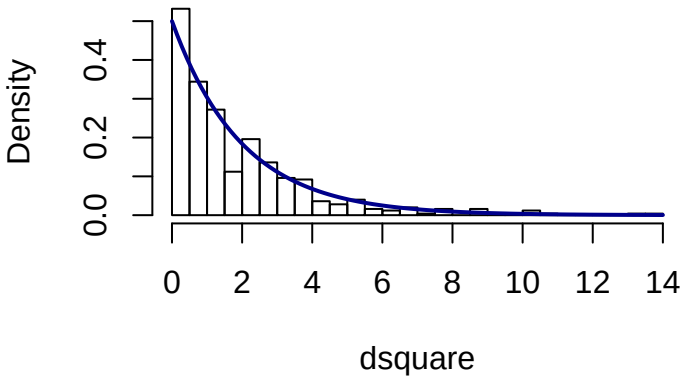
$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \overline{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \overline{\mathbf{x}})}$$

Mahalanobis Example

```
mu <- c(12,30); Sigma <- rbind( c(.8,.5),c(.5,2) )
x <- mvrnorm(500,mu,Sigma)
dsquare <- mahalanobis(x,mu,Sigma)
hist(dsquare,prob=TRUE,breaks=30)
curve(dchisq(x, df=2),
          col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

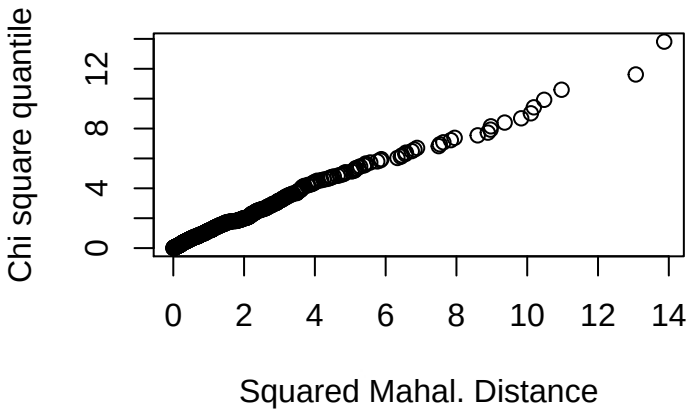
Plot on next slide.

Mahalanobis Example Plot



Chi square quantile plot

```
plot(sort(dsquare),qchisq(ppoints(500),df=2),
     xlab='Squared Mahal. Distance', ylab='Chi square quantile')
```



Assessing Multivariate Normality

- Chi square quantile plot $\rightarrow$ want a straight line
- no *best* hypothesis test
- MVN paackage
  - Henze-Zinkler - `hzTest()`
  - Royston - `roystonTest()`
  - Mardia - `mardiaTest()`
- try all three
  - good agreement $\Rightarrow$ stop
  - marginal significance or inconsistent results $\Rightarrow$ look harder
- beware of small samples

Assessing MVN example

```r
require(MVN) # install if needed
setosa <- as.matrix(iris[iris$Species=="setosa",1:4])
hzTest(setosa)
```

```
##   Henze-Zirkler's Multivariate Normality Test
## ---------------------------------------------
##   data : setosa
##   HZ      : 0.9488453
##   p-value : 0.04995356
##   Result  : Data are not multivariate normal.
```

Assessing MVN example (2)

```
mardiaTest(setosa)
```

```
##     p.value.skew    : 0.1771859
##     p.value.kurt    : 0.1953229
##     p.value.small   : 0.1127617
##
##     Result          : Data are multivariate normal.
```

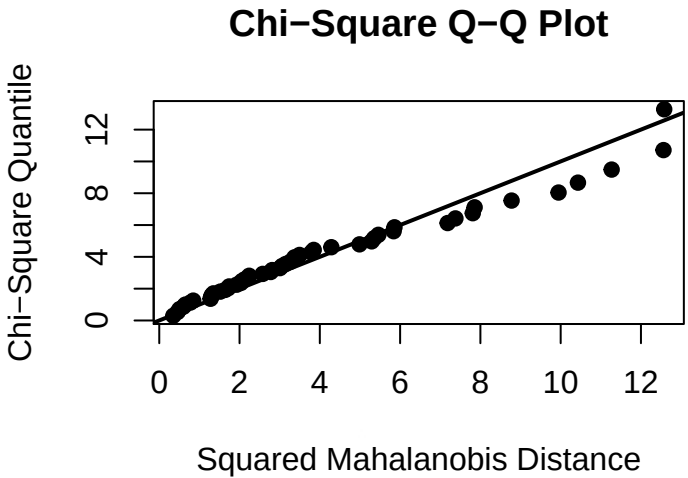
Assessing MVN example (3)

```
roystonTest(setosa)
```

```
##   Royston's Multivariate Normality Test
## ---------------------------------------------
##   data : setosa
##
##   H       : 31.51803
##   p-value : 2.187653e-06
##
##   Result  : Data are not multivariate normal.
```

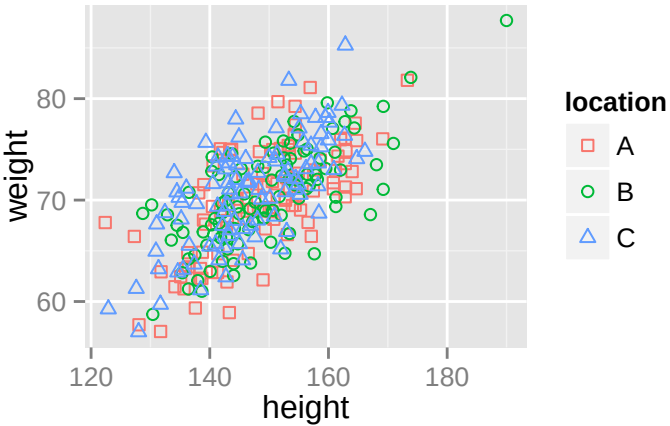
Assessing MVN example (4)

- tests ambiguous so `hzTest(setosa,qqplot=TRUE)`

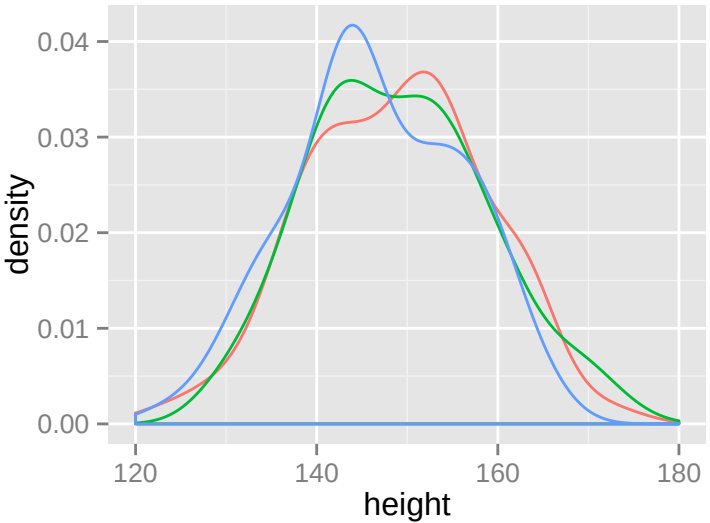


A multivariate problem

Height and weight of apes measured at 3 locations: A, B, C.



Different mean heights?

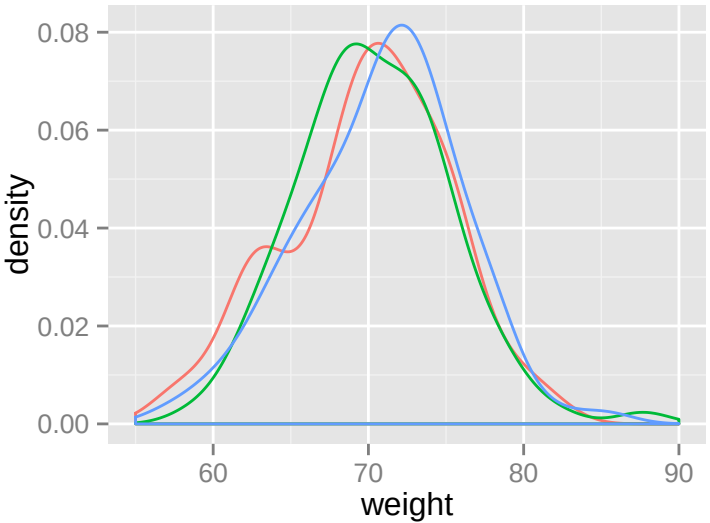


ANOVA on heights

```
aov.model <- aov(height~location,data=apes)
summary(aov.model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## location      2    423  211.74   2.125  0.121
## Residuals   297  29599   99.66
```

Different mean weights?



ANOVA on weights

```
aov.model <- aov(weight~location,data=apes)
summary(aov.model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## location      2     38   19.18   0.738  0.479
## Residuals   297   7719   25.99
```

Why not multiple ANOVA?

- univariate analysis of each variable misses correlations
- multiple tests requires correction to maintain FWER so power is lost

MANOVA

- Multivariate analysis of variance
- do groups have different population mean vectors?

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots \boldsymbol{\mu}_k$$

H_a : at least one mean vector is different

MANOVA is not always appropriate

- if dependendent variables are uncorrelated, then use ANOVA on each variable and correct for multiple tests
- if dependent variables are multicollinear, then should eliminate redundant variables before trying MANOVA

MANOVA requirements

- independent random samples from each population
- data is from multivariate normal distributions
- each distribution has the same covariance matrix

MANOVA Idea

Like ANOVA

$$\text{test stat} \approx \frac{\text{covariance between groups}}{\text{covariance within groups}}$$

but,

- the (co)variances are now matrices
- at least four ways to compute a test statistic

MANOVA Test Statistic

arranged from least likely to make Type I errors to most likely

- Pillai (default in R)
- Wilks Lambda
- Hotelling-Lawley
- Roy

`?summary.manova` for options. None is uniformly most powerful. We will use Pillai.

MANOVA Example (1)

Are the apes at locations A, B, and C different in terms of mean height and weight? (different mean vectors?)

$$H_0 : \boldsymbol{\mu}_A = \boldsymbol{\mu}_B = \boldsymbol{\mu}_C$$

H_a : at least one mean vector is different

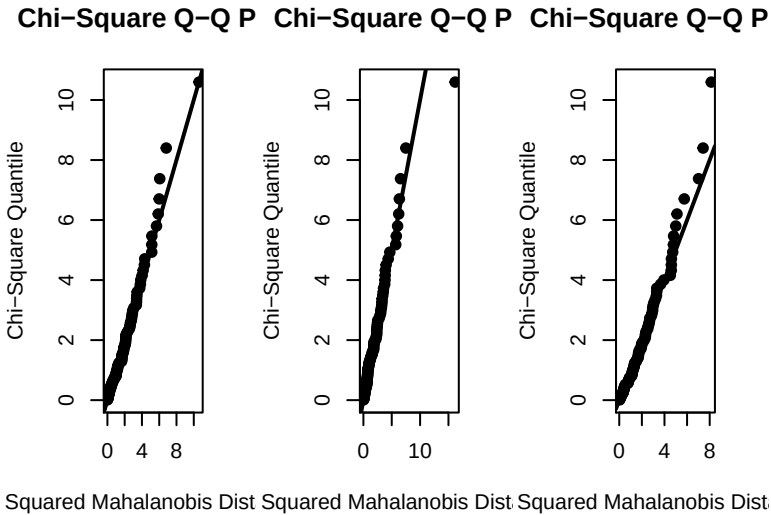
MANOVA Example (2)

Check condition: is data multivariate normal?

```
require(mvoutlier) # install if necessary for chisq plot
old.par <- par()  # save graphics parameters
par(mfrow=c(1,3))
out <- with(apes,hzTest(apes[location=='A',c('height','weight')],qqplo
out <- with(apes,hzTest(apes[location=='B',c('height','weight')],qqplo
out <- with(apes,hzTest(apes[location=='C',c('height','weight')],qqplo
par(old.par) # reset graphics parameters
```

Plots on next slide.

MANOVA Example (3)



MANOVA Example (4)

Equal covariance matrices? Box's M Test can be used to test for equality of covariances.

```
source('BoxMTest.R')
out<-BoxMTest(as.matrix(apes[,1:2]),apes$location)
```

```
## ------------------------------------------------
##  MBox Chi-sqr. df P
## ------------------------------------------------
##     3.8606     3.8231           6       0.7006
## ------------------------------------------------
## Covariance matrices are not significantly different.
```

Do not reject H_0. There is not evidence to show the population covariance matrices are different at locations A, B, and C.

MANOVA Example (5)

```
lmodel <- lm(cbind(height,weight)~location,data=apes)
m.out <- manova(lmodel)
summary(m.out,test="Pillai")
```

```
##             Df   Pillai approx F num Df den Df   Pr(>F)
## location     2 0.050308   3.8318      4    594 0.004384 **
## Residuals 297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reject H_0. There is strong evidence to show apes at the three locations are different in terms of population mean weight and height.

Posthoc Analysis

Often follow up with univariate ANOVAs. Shortcut:

```
summary.aov(m.out)
```

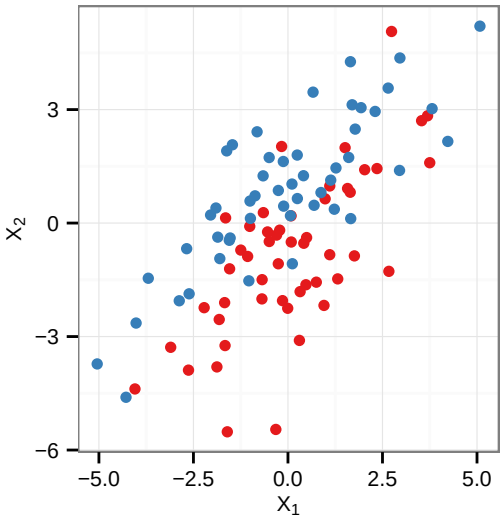
```
##  Response height :
##              Df  Sum Sq Mean Sq F value Pr(>F)
## location      2   423.5 211.742  2.1247 0.1213
## Residuals   297 29598.5  99.658
##
##  Response weight :
##              Df Sum Sq Mean Sq F value Pr(>F)
## location      2   38.4   19.18   0.738 0.4789
## Residuals   297 7719.0   25.99
```

Separating the groups

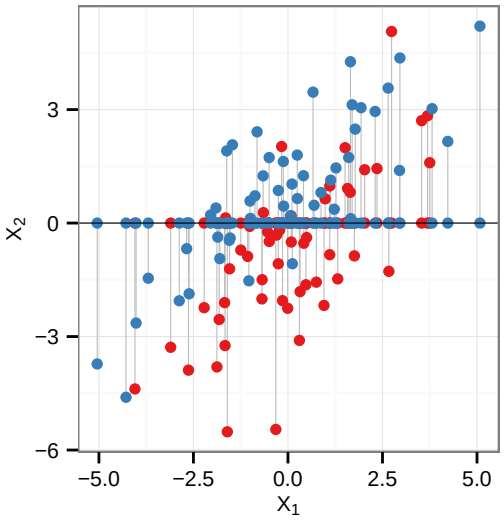
Linear Discriminant Analysis (LDA)

- idea: combine original independent variables to produce new variables
- e.g. x = 0.3 * height + 0.5 * weight
- LDA finds the linear combination(s) that maximizes group separation while minimizing within group variance

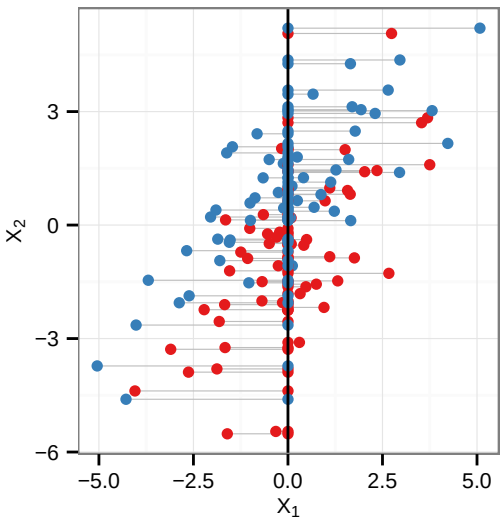
LDA - picture 1



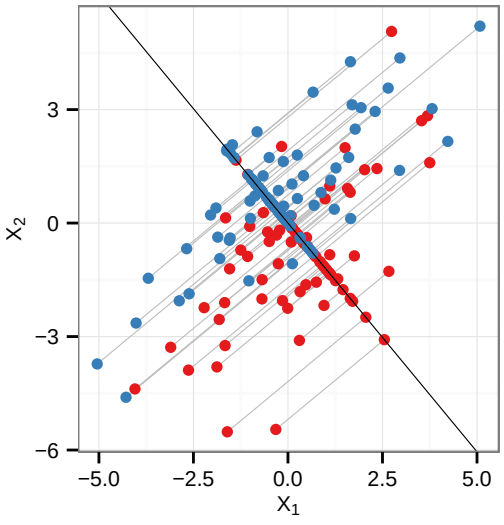
LDA - picture 2



LDA - picture 3



LDA - picture 4



How to separate groups

- use linear discriminant analysis to find combination of variables that maximizes group separation
- apply univariate multiple comparison procedure to new variable

Separating Groups of Apes

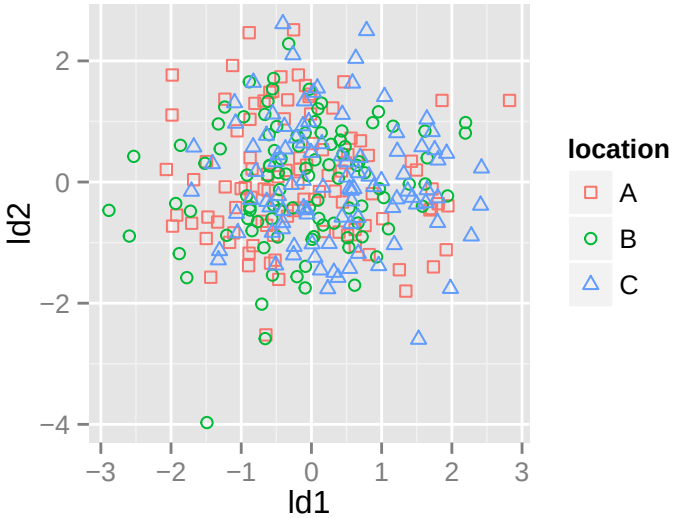
Here is the R code:

```
fit <- lda( location~height+weight,data=apes) # fit model
plda <- predict(object=fit, newdata=apes) # compute combinations
ld1 <- plda$x[,1] # extract most separating combination
ld2 <- plda$x[,2] # second most separating combination
```

New variables scatterplot - code

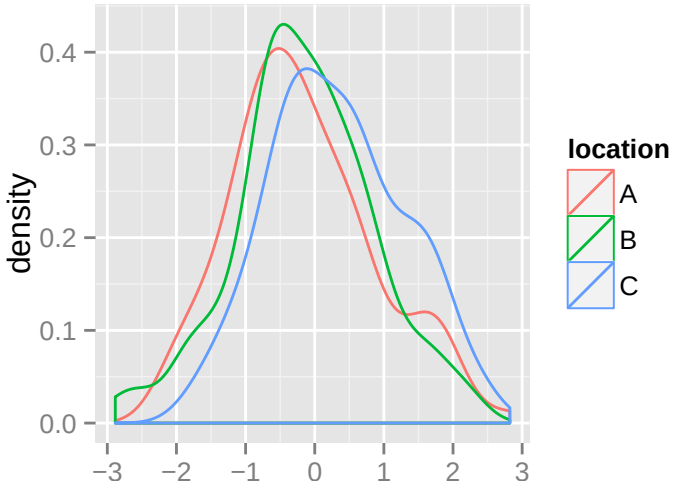
```
apes <- data.frame(apes,ld1=ld1,ld2=ld2)
ggplot(apes)
  + geom_point( aes( x=ld1, y=ld2, color=location, shape=location) )
  + scale_shape_manual( values=c(0,1,2) )
  + scale_size_manual( values=2*c(1,1,1) )
```

New variables scatterplot - the plot



Density Plots for LD1

```
ggplot(apes, aes(ld1, color=location)) + geom_density(alpha=.3)
```



Apply Tukey to LD1

```
linear.model<-aov(ld1~location,data=apes)
TukeyHSD(linear.model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = ld1 ~ location, data = apes)
##
## $location
##             diff        lwr       upr     p adj
## B-A 0.009560405 -0.3235611 0.3426819 0.9974836
## C-A 0.485246799  0.1521253 0.8183683 0.0019781
## C-B 0.475686394  0.1425649 0.8088079 0.0025002
```

Posthoc Conclusions

At the 5% significance level there is strong evidence that the population mean vectors of height and weight for the apes at locations A and B both differ from those at location C. There is not a significant difference in population mean height and weight for the apes at locations A and B.