The analysis of multiple response variables using the multivariate linear model

BIO708

March 25th, 2019

Spoiler alert!

- Analyzing multiple (continuous) response variables often uses very similar frameworks to general(ized) linear models.
- So much of what you have learned applies!
- We just need to learn a few new ideas to generalize it even further.

By the end of class you will be able to:

- Recognize when multiple response variables are used in a linear model framework.

- Use some simple multivariate effect sizes (Euclidian distance, Mahalanobis D).

- Recognize multivariate test statistics as a generalization of univariate test statistics (like T^2 vs t).

How would you…

- If I asked you to design an experiment (and consider the model) to assess differences in male and female heights, how would you do it?

How would you…

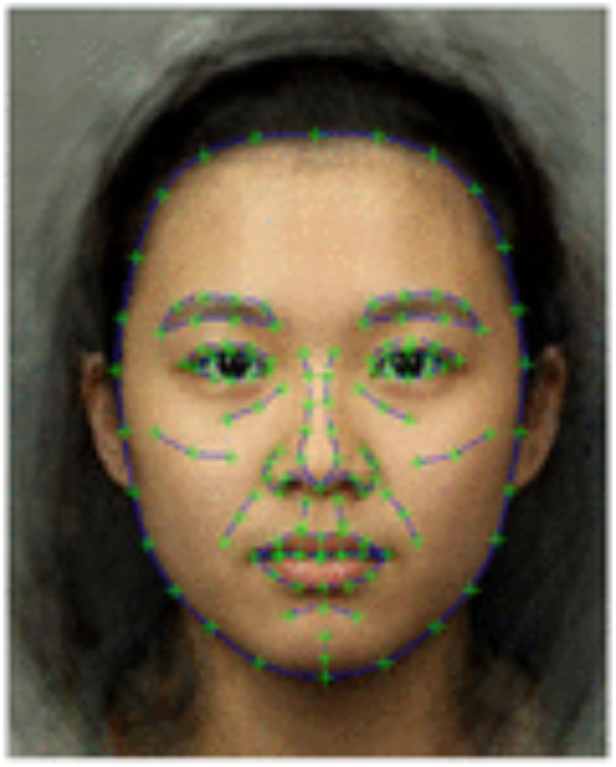
- If I asked you to design an experiment (and consider the model) to assess differences in male and female heights in this class, how would you do it?
- Measure total height (cm) of each individual.
- Use a t-test or a simple linear model to compare mean heights between F and M.

How about…

If I asked you to assess sexual differences in sexual facial shape?

What would you measure?

Multiple measurements



There are many different features we may wish to capture to help us understand differences in facial shape.

- Linear measurements
- Angles
- Landmarks

How do we choose?

https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01883/full

Do we expect all of the measures to be independent?

Do we expect the measure of inter-ocular distance and nose width to be independent of one another?

Why or why not?

https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01883/full

Do we expect all of the measures to be independent?

We expect that many of these “traits” are correlated with each other.

This may due in part to their relationship to overall body size.

But also to more subtle relationships due to (in this case) cranio-facial development.

https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01883/full

Multivariate linear models

- We want to develop approaches that allow us to consider how to model mean differences with multiple, (partially) correlated response traits, which individually may be having subtle effects.
- i.e. So we can ask questions like ”What is the mean difference between facial shape of group X and Y?”

Multivariate Linear Models

- For all of the types of linear models you have been introduced to, there are ***multivariate generalizations***.
- t-statistic $\Rightarrow$ Hotellings T^2
- Linear regression $\Rightarrow$ multivariate linear regression
- Analysis of variance (ANOVA) $\Rightarrow$ multivariate analysis of variance (MANOVA)
- Analysis of covariance (ANCOVA) $\Rightarrow$ multivariate analysis of covariance (MANCOVA)

Multivariate Linear Models

- Indeed all of those (regression, t-test, ANOVA, ANCOVA) are just special cases of the general linear model.

- All of the univariate versions are just special cases of the multivariate linear model.

Multivariate Linear Models

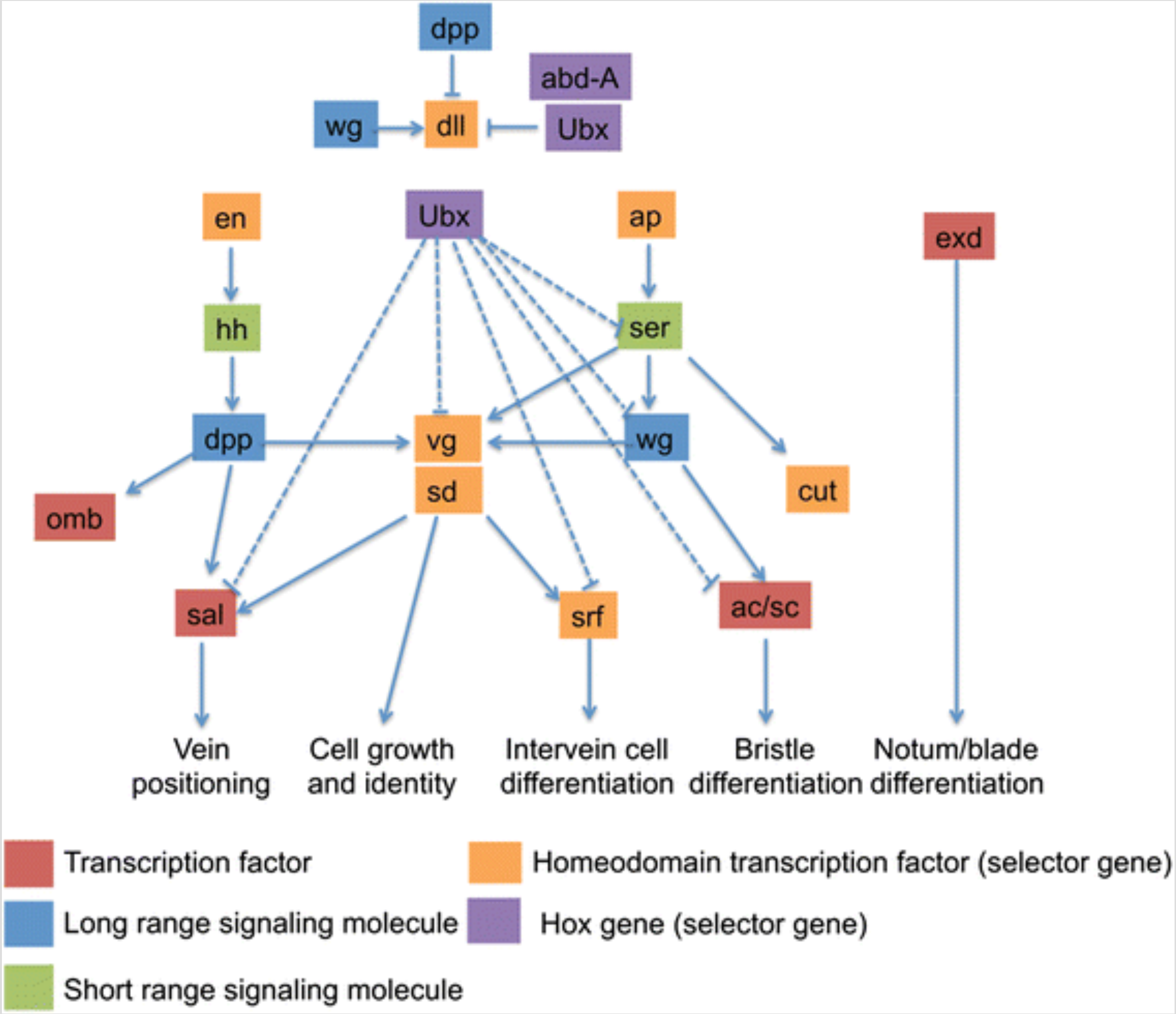
- The general linear model is (as we will see) a special case of the multivariate general linear model.

Other multivariate response variables that we may wish to consider in this context?

Other multivariate response variables that we may wish to consider in this context?

- Analyzing multiple behavioural events related to mating, aggression, etc..
- Multiple fitness components.
- Different aspects of tests scores (MCQ, short answer etc).
- Gene expression data

Analyzing Gene expression data from a gene network


dpp
abd-A
wg
dll
Ubx
en
Ubx
ap
exd
hh
ser
dpp
vg
wg
sd
cut
omb
sal
srf
ac/sc
Vein positioning
Cell growth and identity
Intervein cell differentiation
Bristle differentiation
Notum/blade differentiation
Transcription factor
Homeodomain transcription factor (selector gene)
Long range signaling molecule
Hox gene (selector gene)
Short range signaling molecule

Connahs et al. 2016 BMC Genomics
https://doi.org/10.1186/s12864-016-2586-5

Back to the (univariate) human height example

- What is the difference in human height between males and females.
- We may start by using a t-statistic (t-test) to start with.
- Anyone remember what goes into a t-test?

t-test review

- For two groups (males and females), we want to estimate the mean difference between them.
- But what else do we need to account for?

$$diff = \bar{x}_F - \bar{x}_M$$

t-test review

- We also need to account for the variation due to sampling (uncertainty).
- How representative would the measure from this class be?
- We capture this using the ***pooled standard error*** of the mean.
- First we need the pooled standard deviation

$$s_p = \sqrt{\frac{(n_F - 1)s_F^2 + (n_M - 1)s_M^2}{n_F + n_M - 2}}$$

t-test review

- With the denominator being the pooled standard error of the mean.

$$t = \frac{\bar{x}_F - \bar{x}_M}{s_p\sqrt{\frac{1}{n_F} + \frac{1}{n_M}}}$$

Pooled standard deviation $s_p = \sqrt{\frac{(n_F - 1)s_F^2 + (n_M - 1)s_M^2}{n_F + n_M - 2}}$

t-test review

- So it is just the difference in mean heights divided by a measure of uncertainty in our estimates of mean heights for both M and F.

How about for a set of measures of facial shape?

- We can use the same basic idea.
- But we have to consider a few additional things.
- First we need to think about covariances between our traits, as well as variances for each trait.

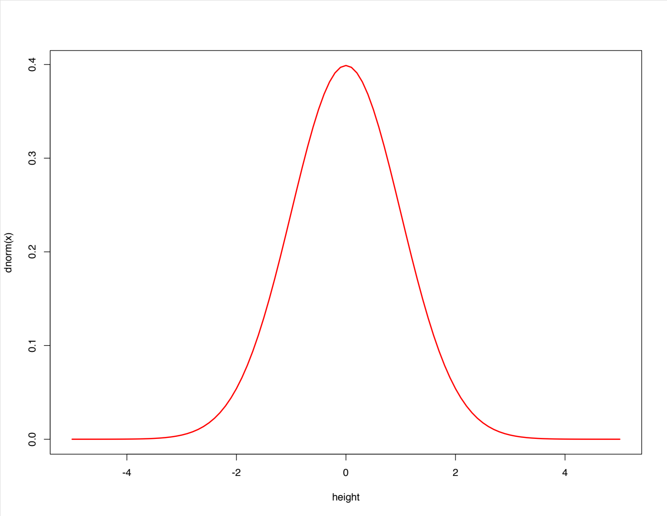
Sample variance and standard deviation

Sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample standard deviation

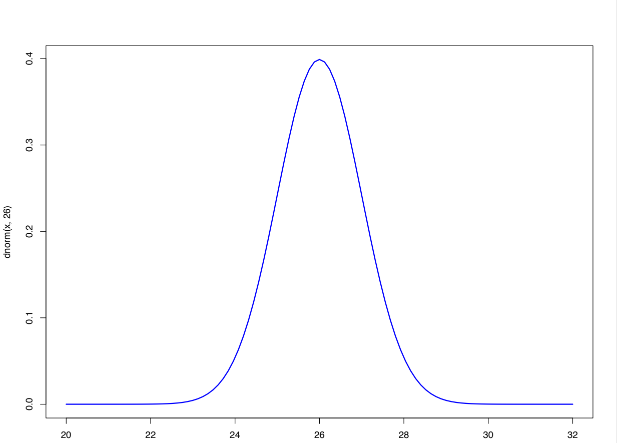
$$s = \sqrt{s^2}$$

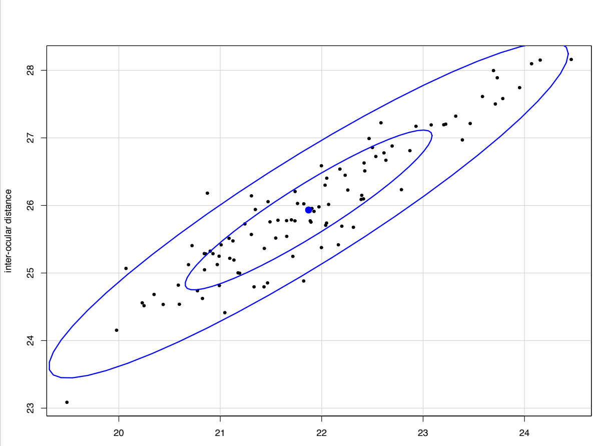


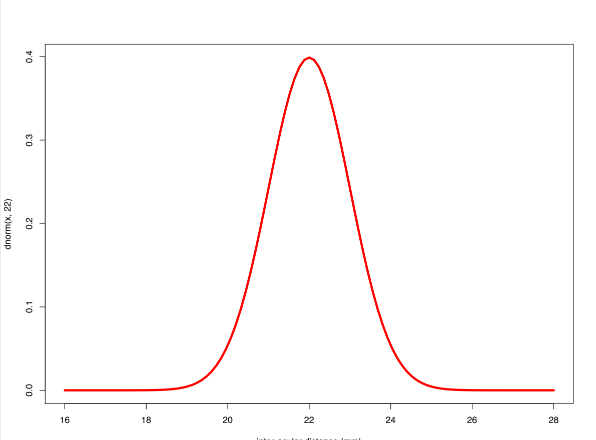
Sample covariance between two variables

Sample covariance:

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$







Variance-covariance matrices are at the heart of the statistical machinery of multivariate methods

- We can summarize all of this in matrix format.

$$\mathbf{S}_{2,2} = \begin{bmatrix} s_x^2 & cov(x,y) \\ cov(x,y) & s_y^2 \end{bmatrix}$$

This is a 2 row by 2 columns (or 2x2) variance covariance matrix.

This matrix has the overall (phenotypic) variances and covariances.

However, in multivariate methods, we will use this same form for our matrices to express uncertainty as well (like with the standard errors).

Note the boldface

- In math and statistics we use boldface to generally suggest that we are looking at matrices or vectors (an object containing at least one, but potentially more numbers).

Vectors

- Usually lower case like

$$\mathbf{y}_{3,1} = \begin{bmatrix} 31 \\ 26 \\ 18 \end{bmatrix}$$

A (single) column vector with 3 rows (3 rows by one column)

$$\mathbf{y}_{1,3} = \begin{bmatrix} 31 & 26 & 18 \end{bmatrix}$$

A (single) row vector with 3 columns (1 row by 3 columns)

Sometimes this notation is used, but not in biology or stats much

$$\vec{y}$$

Matrices

- Just an extension of vectors.
- More than 1 row and column
- You have seen these before when Ben B. discussed design matrices

$$\mathbf{P}_{3,2} = \begin{bmatrix} 2 & 9 \\ 7 & 11 \\ 3 & 16 \end{bmatrix}$$

3 rows by 2 columns

Vector and Matrix form for linear regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_1 + e_1 \\
y_2 &= \beta_0 + \beta_1 x_2 + e_2 \\
&\;. \\
&\;. \\
y_n &= \beta_0 + \beta_1 x_n + e_n
\end{aligned}$$

$$\begin{array}{cccccc}
\boldsymbol{Y}_{n,1} & & \boldsymbol{X}_{n,2} & \boldsymbol{\beta}_{2,1} & & \boldsymbol{e}_{n,1} \\
\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} & = & \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & . \\ 1 & . \\ 1 & x_n \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} & + & \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix}
\end{array}$$

Distances between vectors.

- In multivariate statistics we generally use distances as a basic measure of magnitude (essentially the squared difference)

$$diff = \begin{bmatrix} \bar{x}_{IO} \\ \bar{x}_{NW} \end{bmatrix}_F - \begin{bmatrix} \bar{x}_{IO} \\ \bar{x}_{NW} \end{bmatrix}_M$$

This is a bit messy

$$diff = \bar{\mathbf{x}}_{\mathrm{F}} - \bar{\mathbf{x}}_{\mathrm{M}}$$

Much easier to read!

The problem is that we still have a vector of differences

Distances between vectors (for the male, female example)

$$\bar{\mathbf{x}}_{\text{diff}} = \bar{\mathbf{x}}_{\text{F}} - \bar{\mathbf{x}}_{\text{M}}$$

$$d^2 = \sum_{i=1}^{k} \left(\bar{x}_{i,F} - \bar{x}_{i,M}\right)^2$$

For k different traits

$$d^2 = (\bar{\mathbf{x}}_{\text{F}} - \bar{\mathbf{x}}_{\text{M}})' \cdot (\bar{\mathbf{x}}_{\text{F}} - \bar{\mathbf{x}}_{\text{M}})$$

Same thing, but using matrix algebra

$$d = \sqrt{d^2}$$

$$\|d\| = \sqrt{\bar{\mathbf{x}}_{\text{diff}}' \cdot \bar{\mathbf{x}}_{\text{diff}}}$$

Again, same thing. Starting with differences

This is the Euclidean distance between the mean vectors. sometimes called the magnitude of the difference vector (or the L2 norm).

The pooled variance-covariance matrix for males and females

$$\mathbf{S}_{\mathrm{pl}} = \frac{(n_F - 1)\mathbf{S}_{\mathrm{F}} + (n_M - 1)\mathbf{S}_{\mathrm{M}}}{n_F + n_M - 2}$$

Exactly the same form as we saw for the pooled standard deviation for a single trait.

Now we can construct the multivariate generalization of a t statistic

- For the t-test, the multivariate generalization is called Hotelling’s T^2.

- ”squared” because we are using the distance.

Hotelling's T^2

$$T^2 = \frac{n_F n_M}{n_F + n_M}(\bar{\mathbf{x}}_F - \bar{\mathbf{x}}_M)'\mathbf{S}_{\mathrm{pl}}^{-1}(\bar{\mathbf{x}}_F - \bar{\mathbf{x}}_M)$$

Let's take this apart…
Ignoring the sample size parts (n_M, n_F) for the moment.

Matrix algebra notation means we need to write this in a particular way but all this says is compute the distance

and multiply by the inverse of the pooled covariance matrix (kind of like dividing by the standard error for a regular t-test).

Most of the ideas with multivariate linear models extend from this (with all the same complications)

- One important “complication” to be aware of is that there are multiple multivariate generalizations for the univariate test statistics.

- Where as a univariate glm may use an F statistic for an ANOVA, there are multiple generalizations of this, with slight differences.

How about effect sizes

- We will continue to use distances.
  - We can make an effect size analogous to Cohen’s D (univariate) quite easily as well.

$$Cohen's\ d = \frac{\bar{x}_F - \bar{x}_M}{s_{pooled}}$$

The difference just scaled by the pooled standard deviation

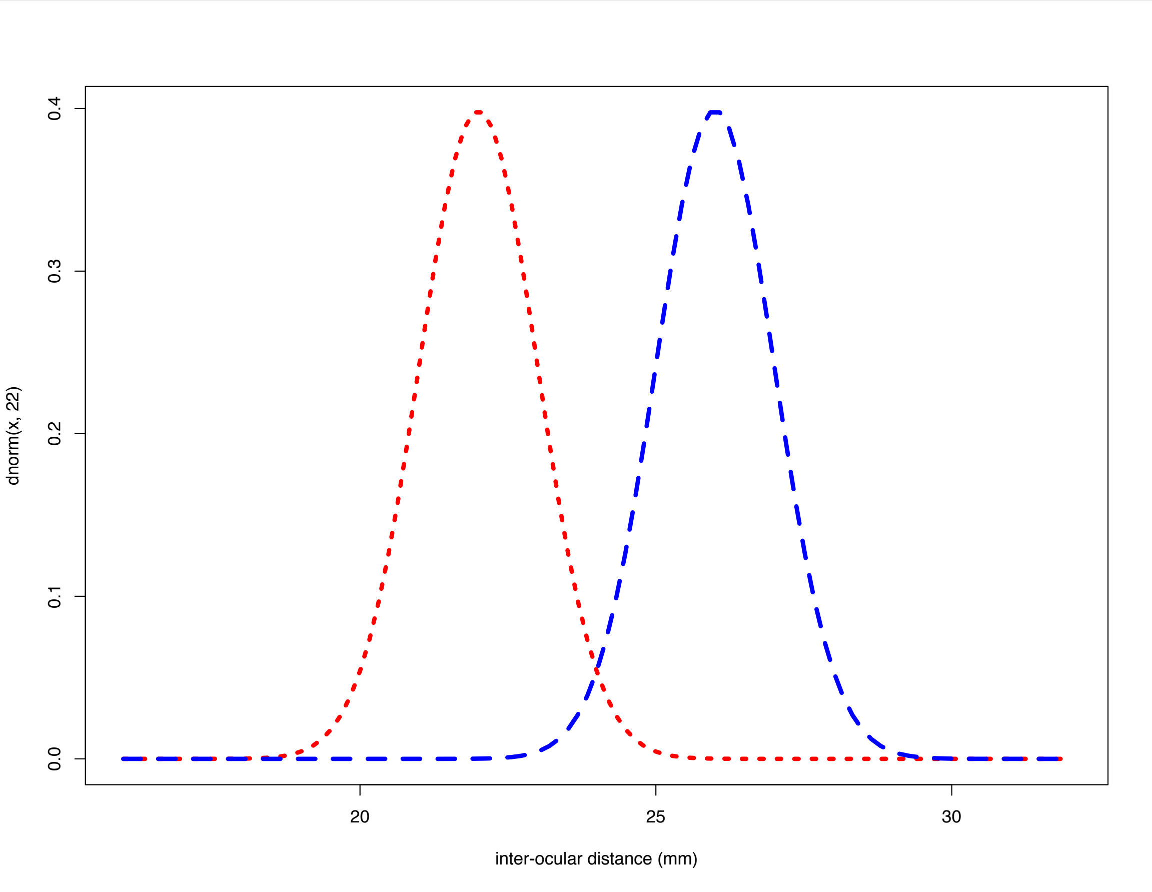
Mahalanobis Distance

- Multivariate analog to Cohen's D
- The distance is scaled by the covariance matrix.

$$D^2 = (\bar{\mathbf{x}}_F - \bar{\mathbf{x}}_M)'\mathbf{S}_{\text{pl}}^{-1}(\bar{\mathbf{x}}_F - \bar{\mathbf{x}}_M)$$

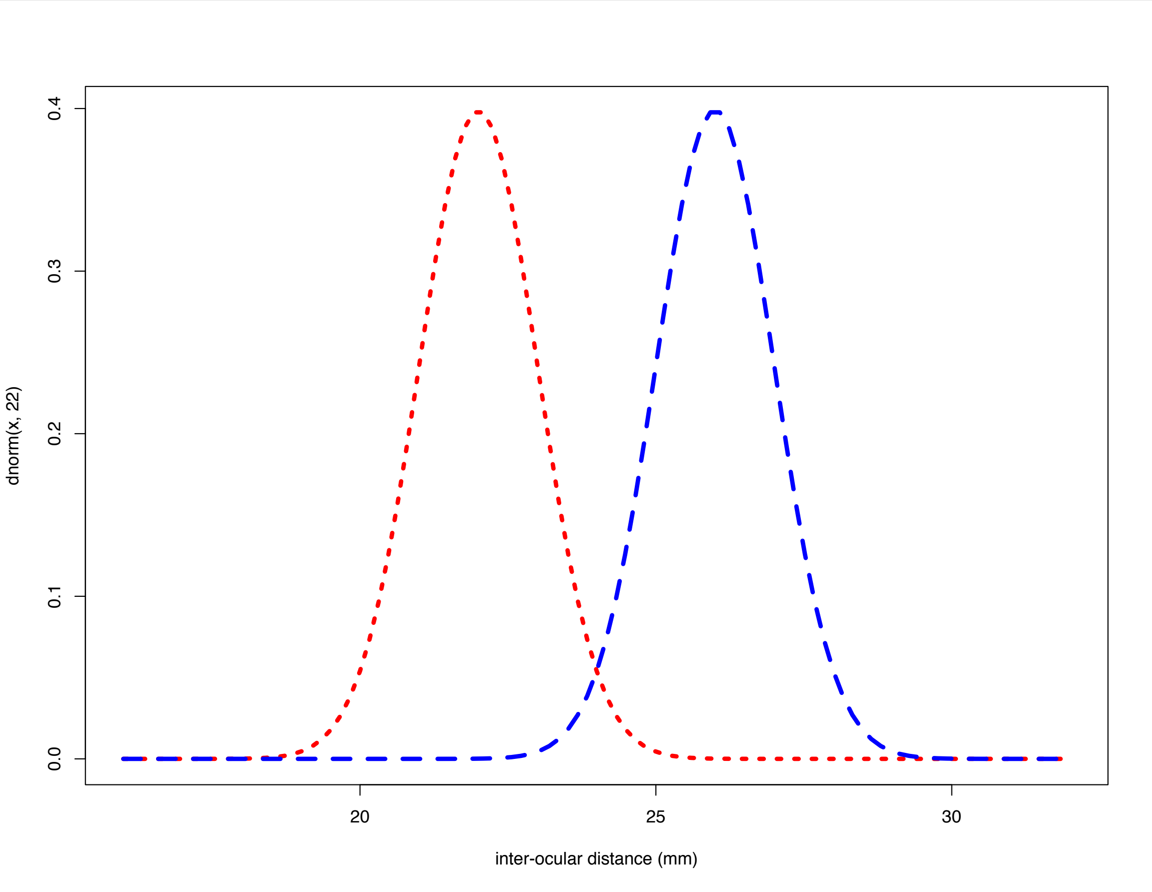
In R the basic function is mahalanobis()

Why do I keep going on about "pooled"



What would happen to the estimate of the variance/sd if I just combined all of the data together, ignoring the mean?

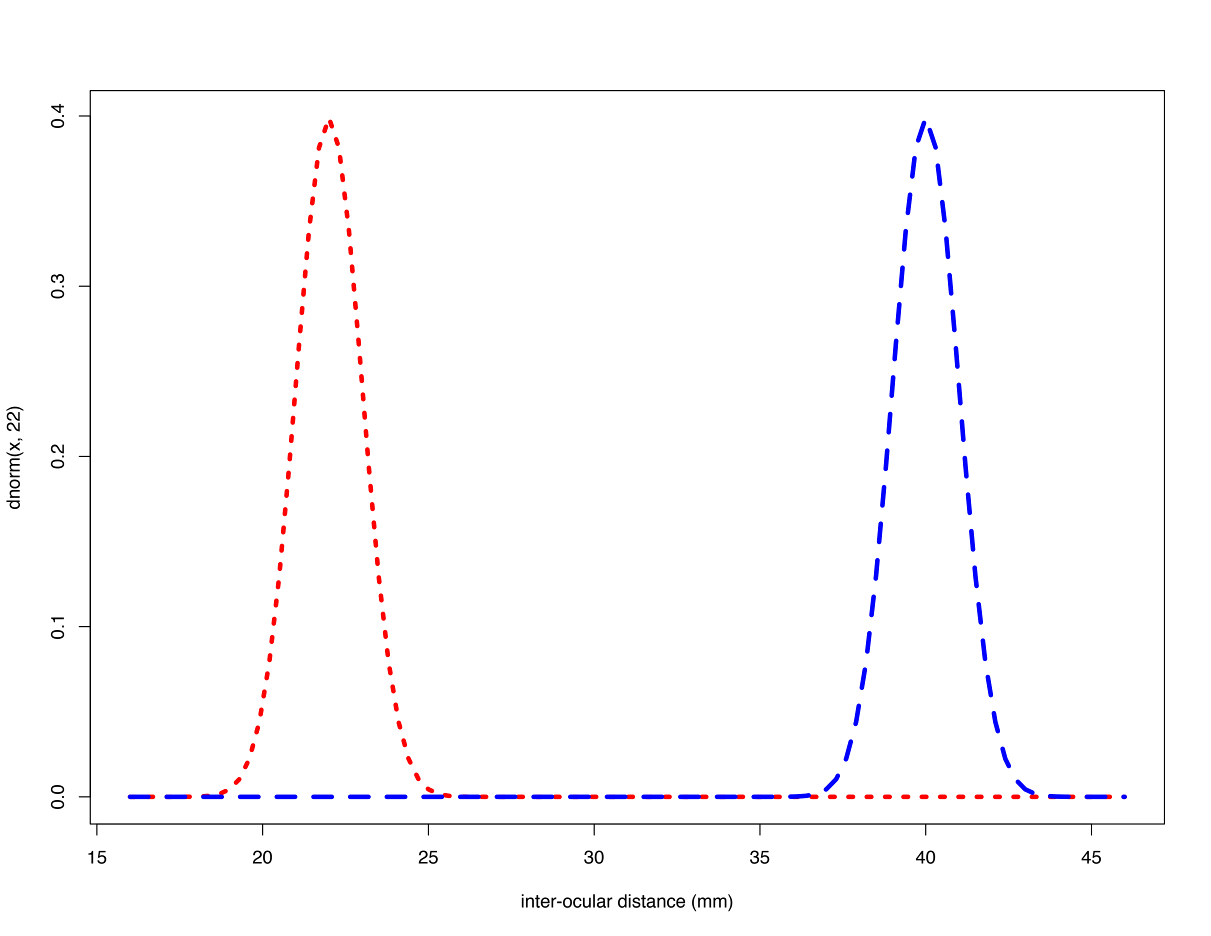
Why do I keep going on about "pooled"



Each group has a sd =1,

If I just lump them together… they have a ~ sd of 2.23

Even further apart



Each group has a sd =1,

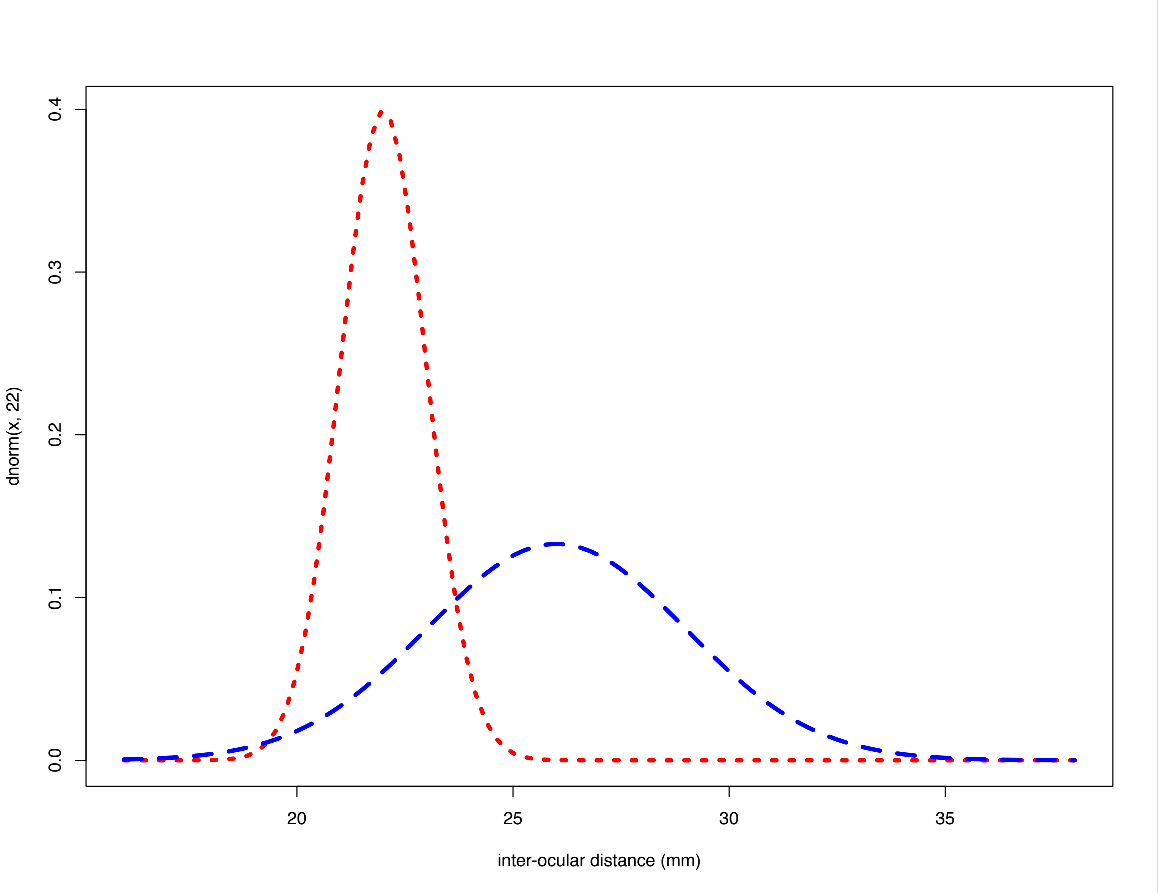
If I just lump them together… they now have a sd of ~9

What would happen to our estimate of Cohen's D if we did not do proper pooling?

$$Cohen's\ d = \frac{\bar{x}_F - \bar{x}_M}{s_{total}}$$

…It would cause us to underestimate the magnitude of the difference

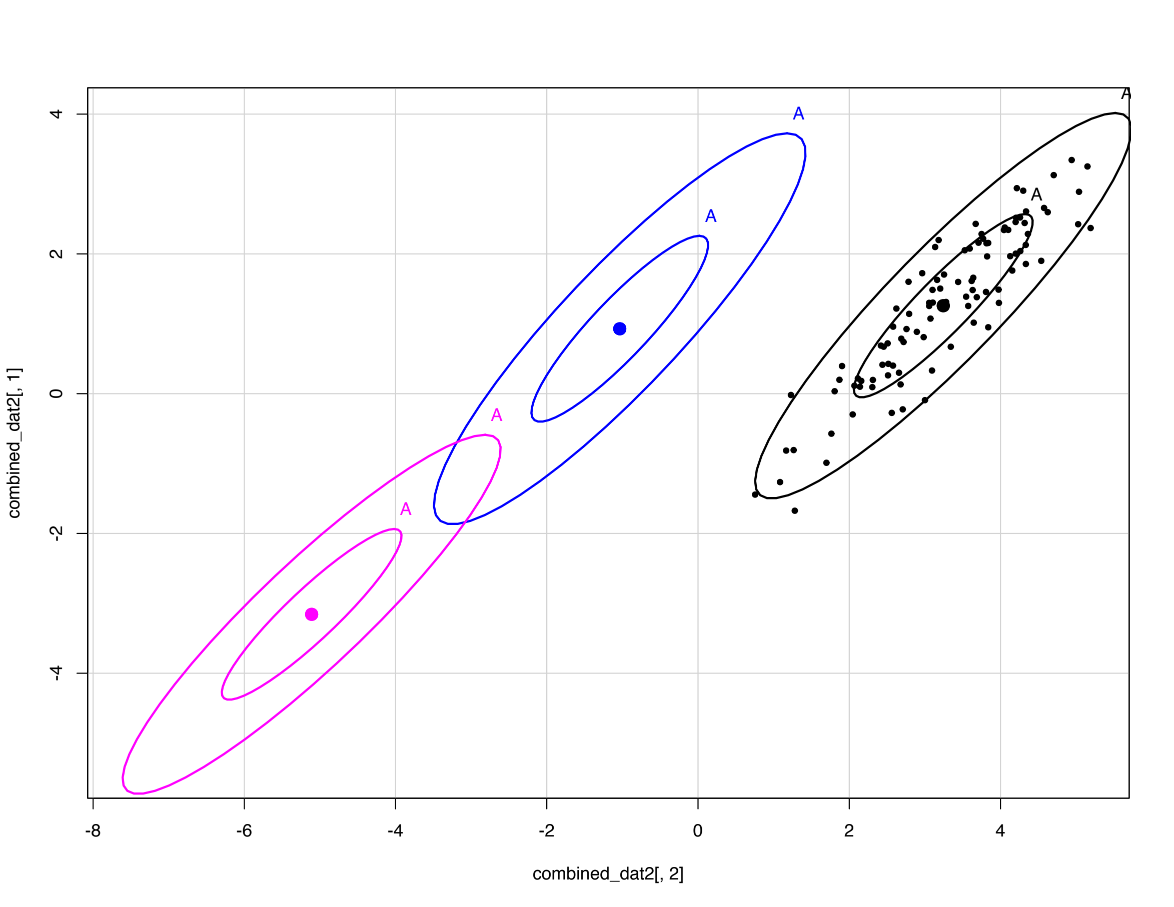
Also this….



- Variances may not be equal in each group….

The pooled covariance matrix

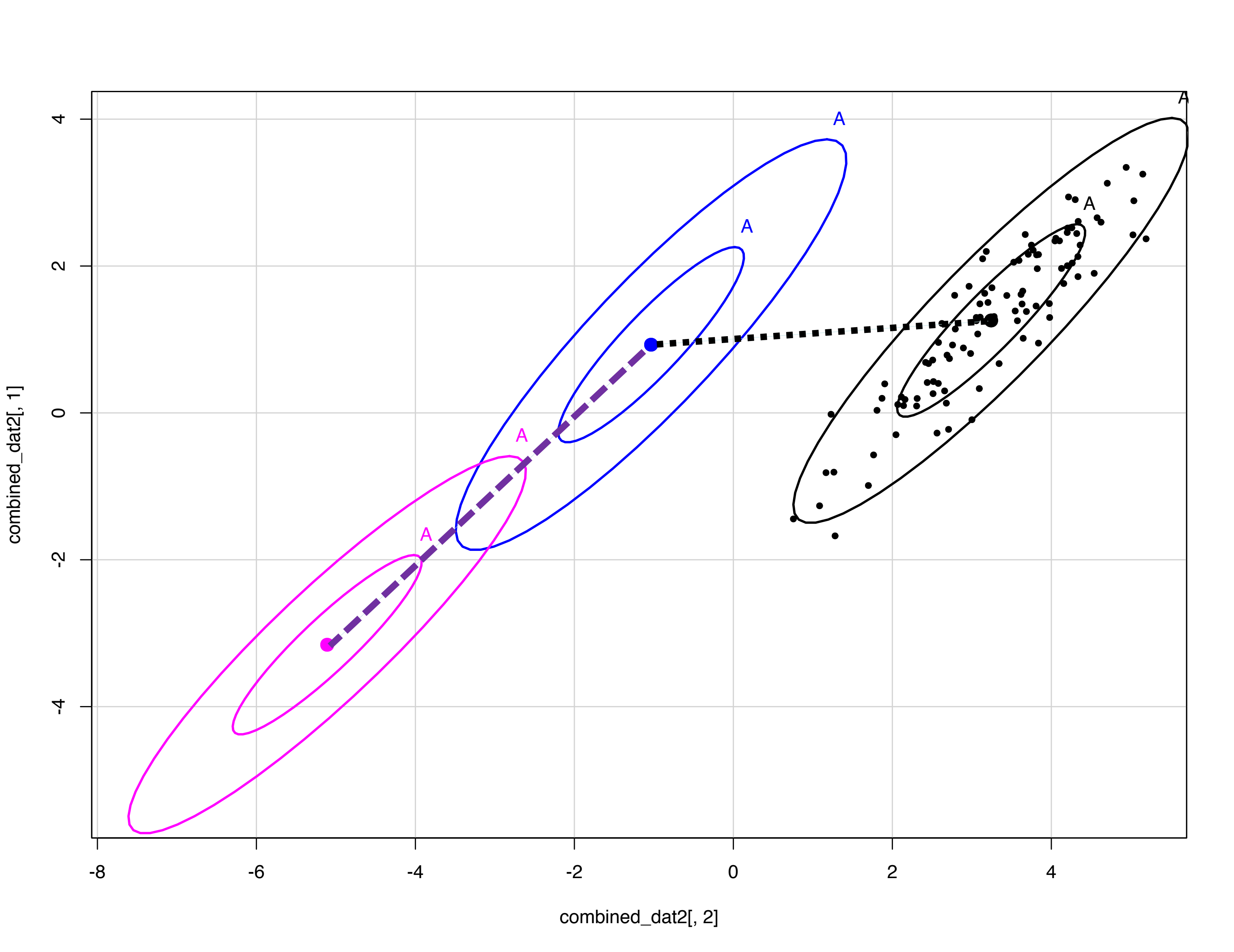
For multiple traits, not only can the means and variances differ (as with the univariate situation), but patterns of covariance as well.



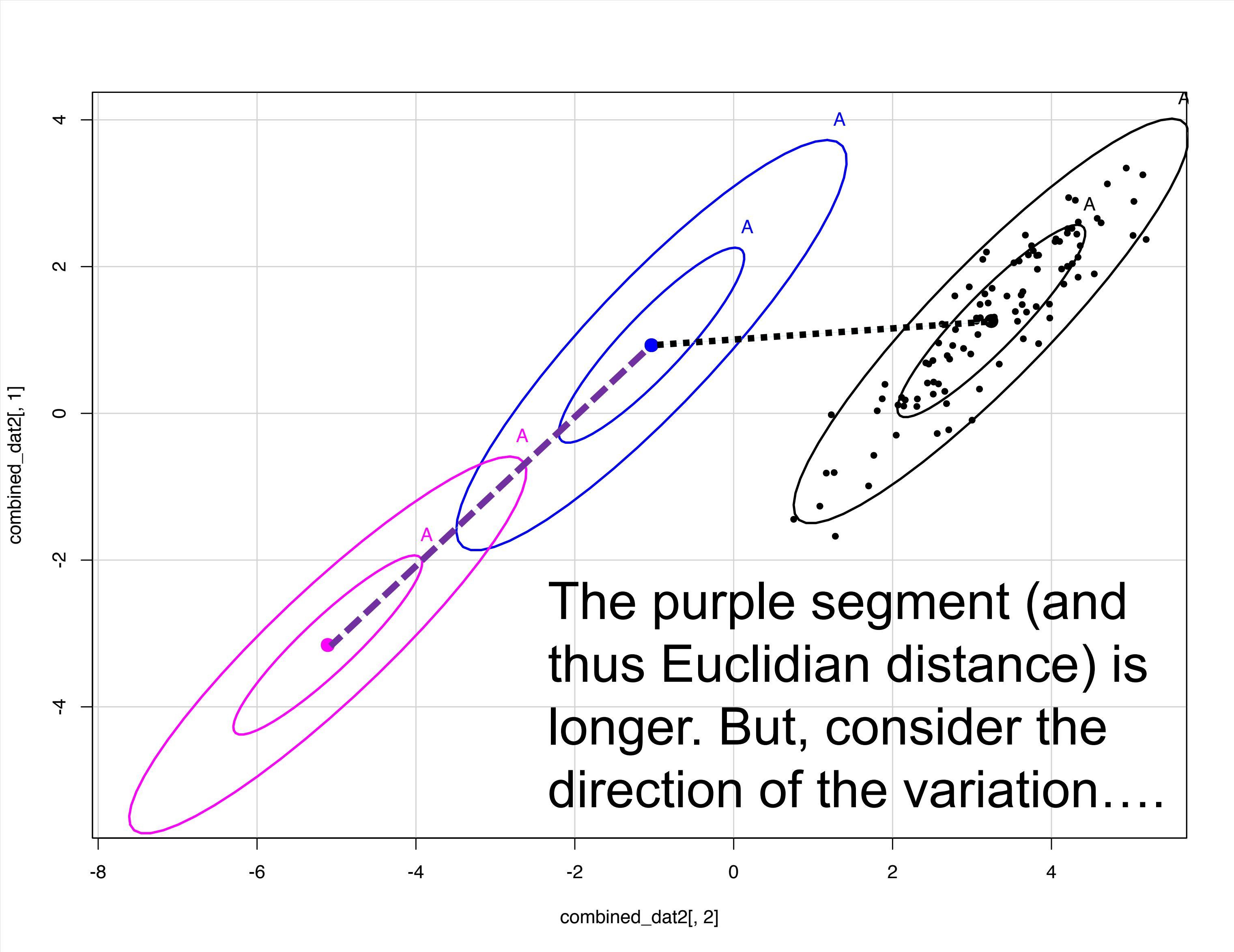
Consider the mean differences between these three groups.

What is more different from “blue”?

The pooled covariance matrix



The pooled covariance matrix



The purple segment (and thus Euclidian distance) is longer. But, consider the direction of the variation….

The pooled variance-covariance matrix for males and females

$$\mathbf{S}_{\mathrm{pl}} = \frac{(n_F - 1)\mathbf{S}_{\mathrm{F}} + (n_M - 1)\mathbf{S}_{\mathrm{M}}}{n_F + \ n_M - 2}$$

Exactly the same form as we saw for the pooled standard deviation for a single trait.

The effect of multiplying by $\boldsymbol{S}^{-1}$

- It has the effect of “squishing the data”, changing the cigar shaped ellipse to that of a circle.

- It will also change the distances.
  - In effect the Mahalanobis Distance between pink and blue would decrease because their difference is in the same direction as much of the variation.

Now we can proceed very much as we do with other linear models.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

Now we have a matrix, $\mathbf{Y}$ of response variables (with each row corresponding to an individual sample).

The design matrix, $\mathbf{X}$ is as before.

We are estimating a vector of parameters for each predictor variable instead of a just a single parameter. So $\mathbf{B}$ will be much larger.

Solution for the parameter estimates in the general linear models

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$$

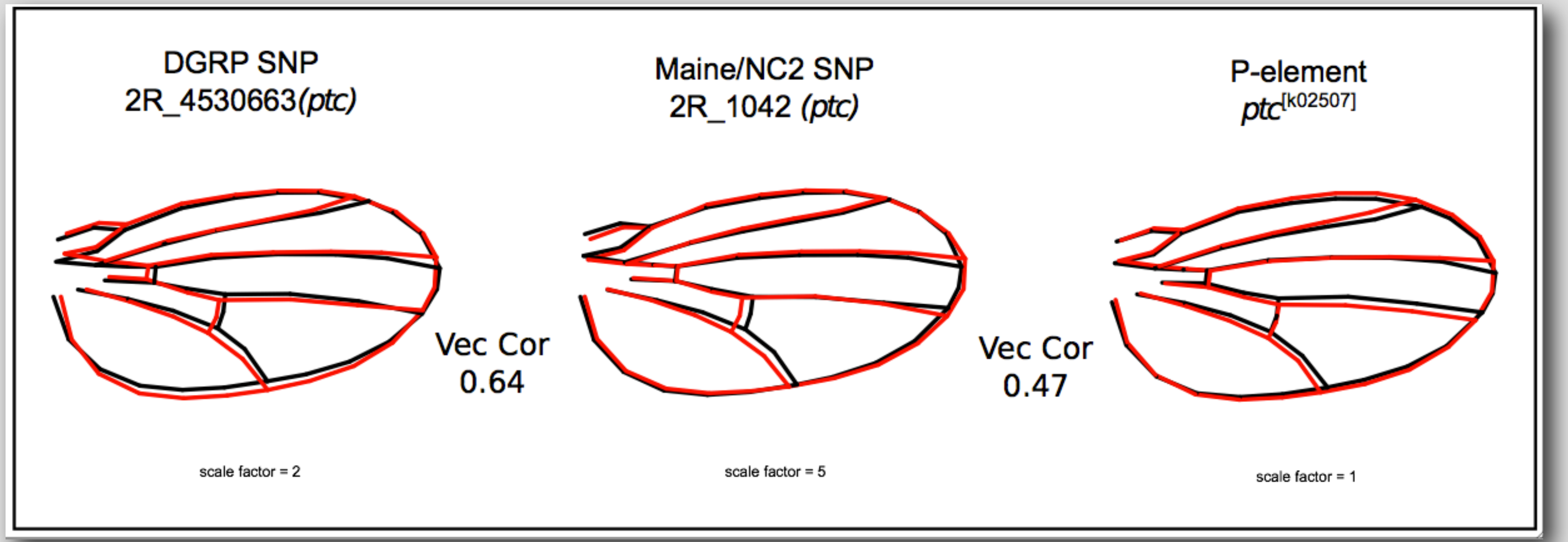
$$\widehat{\boldsymbol{B}} = (\mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$$

Our solutions are the same for univariate and multivariate.

Benefits of multivariate approaches

- In general, the approaches are more “powerful” (in the traditional sense of statistical power).
- When working with correlated response variable it has been demonstrated that such approaches are better than fitting a single model for each response variable.
- Effect sizes allow you to examine both magnitude and direction of effects.

Correlations in the direction of multivariate effects (the angle between vectors is a transformation of this)



So we can address not only how big an effect is, but how similar it is as well

Weber et al. 2005, Dworkin & Gibson 2006, Haber & Dworkin 2016, Pitchers et al. 2019

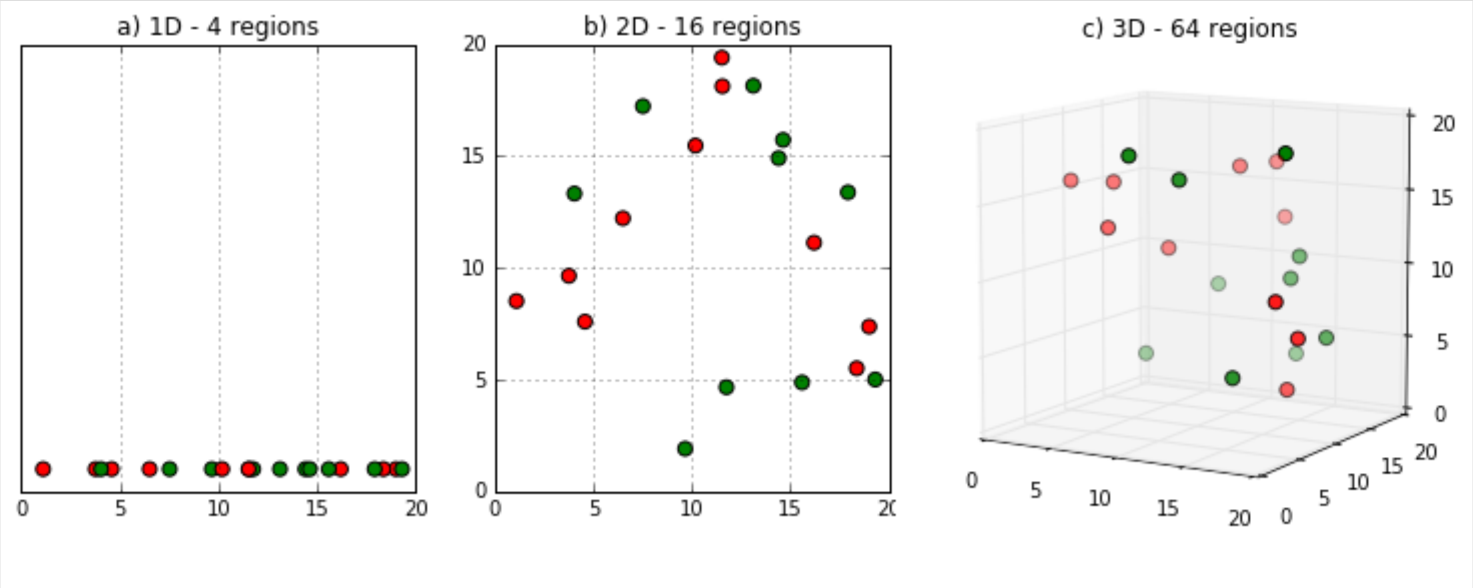
What are the downsides

- Data hungry.
- for k different response variables, there are $k(k+1)/2$ free parameters (variances and covariances) to be estimated.
- These need to be estimated for each predictor (or random effects), and that is potentially asking to estimate a lot of parameters

For two traits, this is only 3 parameters for the simplest model. For the 58 dimensional representation of shape, it is 1711!

Downsides: Curse of dimensionality

Samples can become increasingly sparse in many dimensions.



https://www.kdnuggets.com/2017/04/must-know-curse-dimensionality.html

Downsides: Curse of dimensionality

- This relates in part to the previous issue.

- The data may seem to be clustered when examined in one or two dimensions.

- But when examining dozens of variables, the data could be very sparse (it does not occupy much space), making it difficult to estimate parameters

The curse may not be quite as bad as it is made out to be

- Real biological data tends to be highly clustered and correlated. So the data is not locally “sparse”.

- Still, it is a concern.

Wednesday

- We will go through some examples for multivariate linear models and multivariate linear mixed models.

- I will also introduce one important piece from matrix algebra (eigenvalues) which are used in hypothesis testing for multivariate models.