

Week 12: Multivariate statistics (Part 1)

ANTH 674: Research Design & Analysis in Anthropology

Professor Andrew Du

Andrew.Du2@colostate.edu

Office Hours: Thursdays, 9:00am–12:00pm

In person: GSB 312

Virtual: <https://tinyurl.com/F22ANTH674>

1

Statistics vignette

Are correlations always transitive?

- E.g., if X & Y are correlated and Y & Z are correlated, are X & Z correlated?
- Will prove the answer geometrically, so need to go over the geometric interpretation for correlation

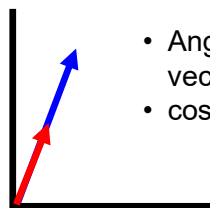
2

Cosine similarity

= Pearson correlation on centered variables

X	Y
1	2
2	4
3	6
4	8
5	10
⋮	⋮

Looking at first two observations only
(otherwise, need *many* axes)



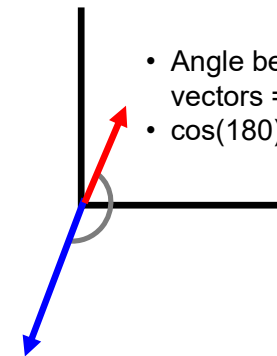
- Angle between vectors = 0
- $\cos(0) = 1$

3

Another example

X	Y
1	-2
2	-4
3	-6
4	-8
5	-10
⋮	⋮

Looking at first two observations only



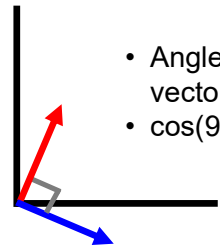
- Angle between vectors = 180
- $\cos(180) = -1$

4

Another example

X	Y
1	2
2	-1
⋮	⋮

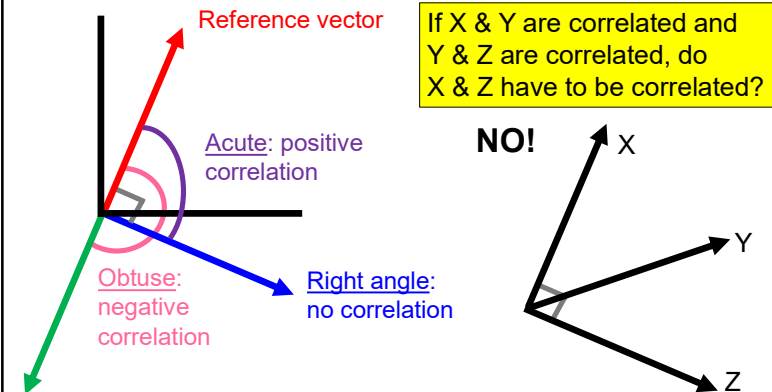
Looking at first two observations only



- Angle between vectors = 90
- $\cos(90) = 0$

5

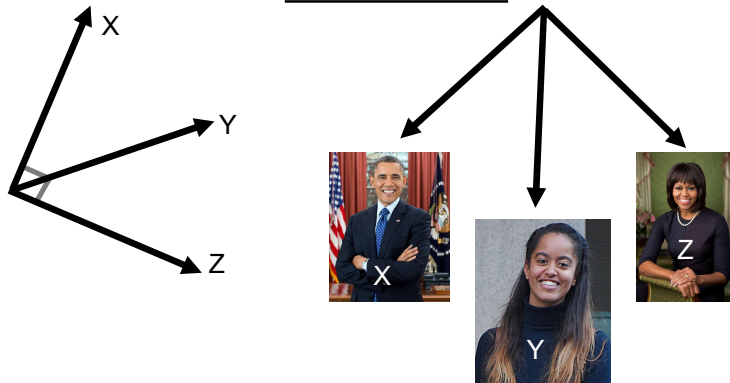
Putting it all together



6

How to think about it

"Blood relation"

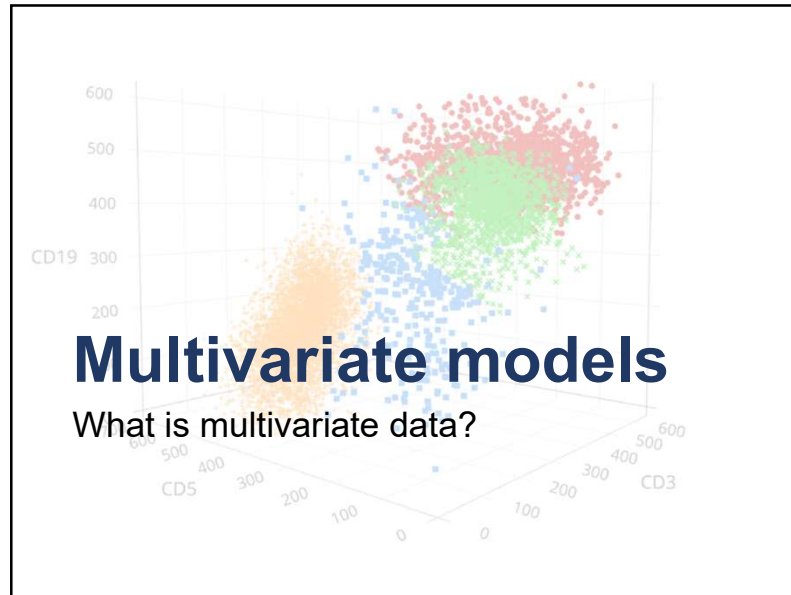


7

Lecture outline

- Multivariate models
 - What is multivariate data?
 - MV models w/ one categorical IV
 1. Hotelling's T^2
 2. One-way MANOVA
- Introduction to ordination
 - Principal components analysis (PCA)
 - The general mechanics of PCA

8



9

What is multivariate data?

- Thus far, we have modeled only univariate DVs ~ one or more IVs
 - E.g., `mtcars$qsec ~ mtcars$hp`
- But many times, we want to simultaneously analyze ≥ 2 DVs ~ ≥ 1 IVs
 - E.g., `cbind(iris$Petal.Length, iris$Petal.Width) ~ iris$Species`
 - Looks at overall petal morphology & size ~ species
- Each DV variable can be continuous or categorical (we'll focus on continuous only)

10

Independent DVs?



- Because each DV is measured on the same individual, DVs could be non-independent
- E.g., `cbind(iris$Petal.Length, iris$Petal.Width) ~ iris$Species`

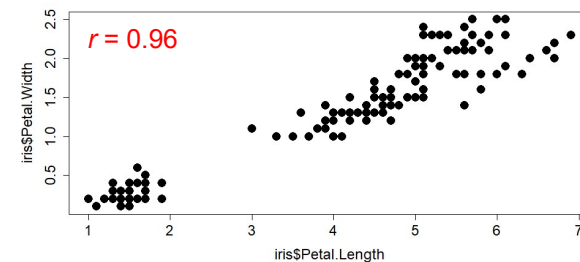
Are these two DVs independent?

	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

11

Independent DVs?

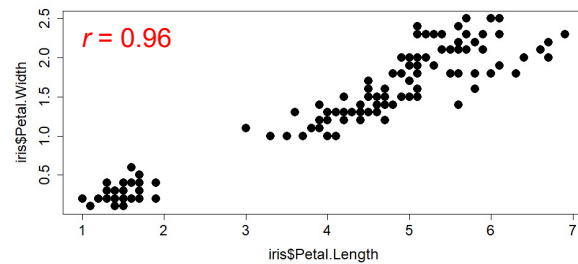
- **NO!**
- Individuals w/ larger petal lengths are more likely to have larger petal widths (i.e., size-based)



12

Multivariate models

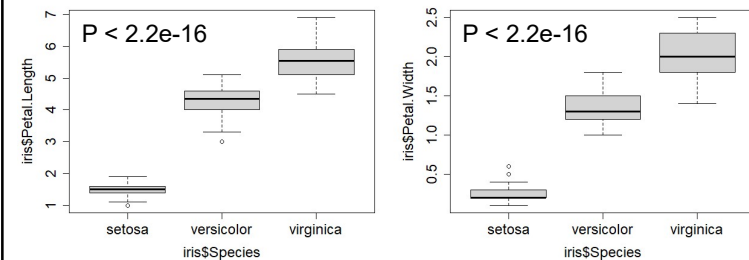
- Takes into account non-independence (correlation) between DVs



13

Univariate models

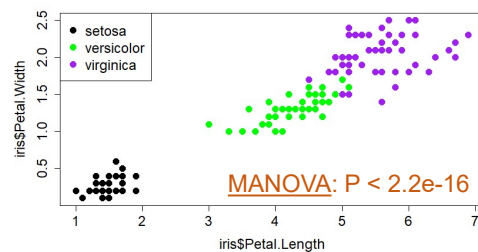
- Can do ANOVAs on DVs separately
- Ignores correlation between **Petal.Length** and **Petal.Width**



14

Multivariate models

- Takes into account correlation between **Petal.Length** and **Petal.Width**
- Are means of **Petal.Length** & **Petal.Width** point clouds different across species? (different question!)

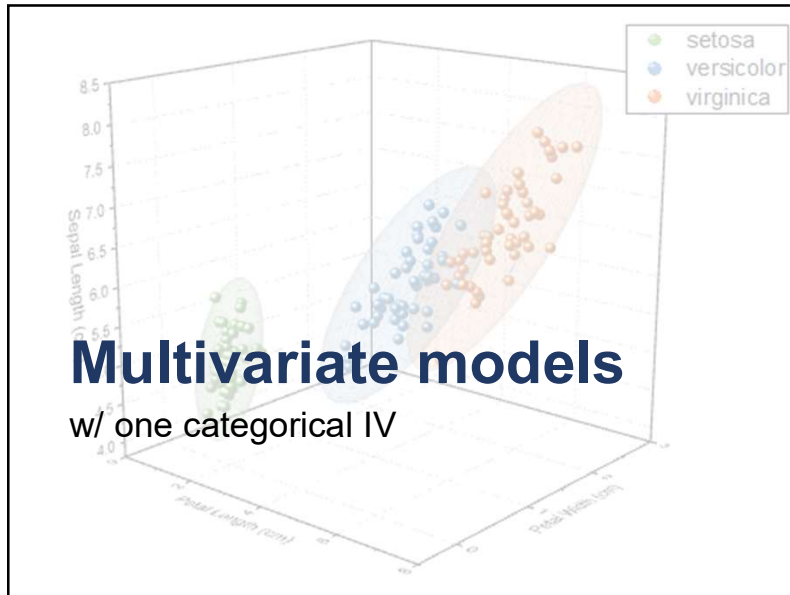


15

Questions?



16



17

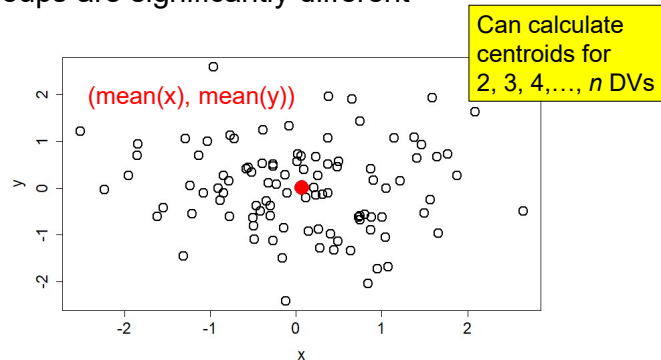
Multivariate models

- Models two or more DVs ~ one or more IVs
- We will look at models w/ one categorical IV only
- Which test? Univariate DV ~ one binomial IV?
 - t-test → Hotelling's T^2 (multivariate)
- Which test? Univariate DV ~ one multinomial IV?
 - One-way ANOVA → One-way MANOVA (multivariate)

18

MV models w/ 1 categorical IV

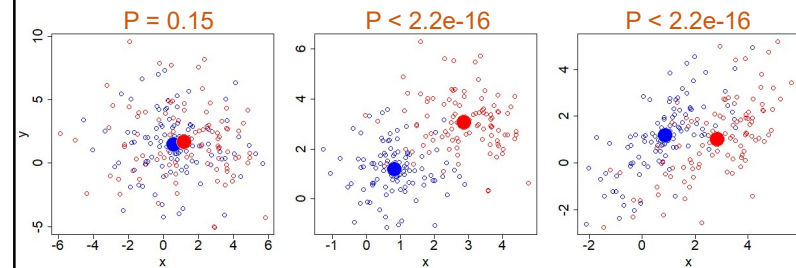
- Asks if **centroids** of point clouds between groups are significantly different



19

Hotelling's T^2

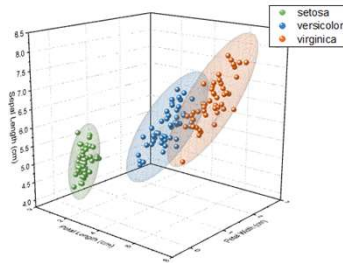
- Tests if two groups' centroids are significantly different, given how much DVs vary & covary
- E.g., simulated example w/ 2 DVs



20

One-way MANOVA

- Multivariate ANOVA → MANOVA
- Tests if ≥ 3 groups' centroids are significantly different, given how much DVs vary & covary

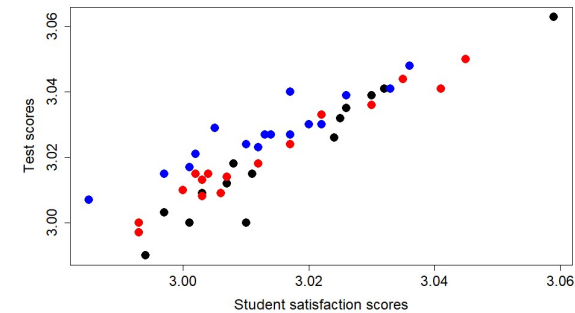


21

E.g., teaching scores

<https://statisticsbyjim.com/anova/multivariate-anova-manova-benefits-use/>

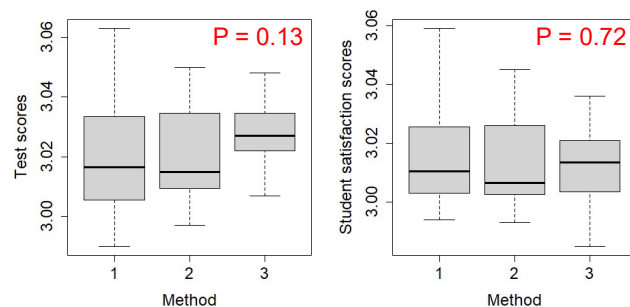
- We have two DVs (student satisfaction & test scores) ~ three different teaching methods



22

Univariate approach

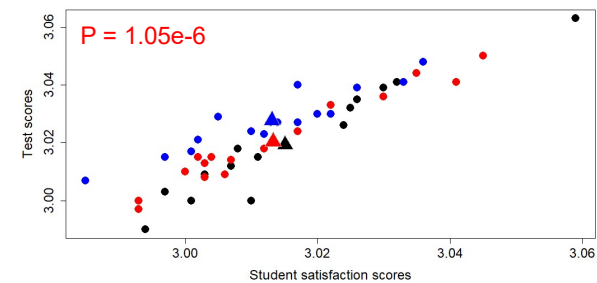
- Two one-way ANOVAs (one per DV)
- Ignores correlation btw test & satisfaction scores



23

Multivariate approach

- Takes into account correlation between DVs
- Overlap between groups in X- or Y-dimension, but not when looking at X & Y together



24

MV model advantages



- Greater statistical power
 - When DVs are correlated, MV models can detect smaller effects than w/ ANOVAs
- Detect IV affecting relationship between DVs
- Limit the number of tests run (e.g., multiple ANOVAs)

<https://statisticsbyjim.com/anova/multivariate-anova-manova-benefits-use/>

25

Assumptions

1. Observations are independent and randomly sampled from population
2. W/in each group, DVs are multivariate normally distributed (cf., each DV is normally distributed w/in each group)
 - ANOSIM relaxes this assumption (nonparametric)
3. Variances of DVs and how they're correlated w/ each other are the same for each group (cf. homoscedasticity)

26

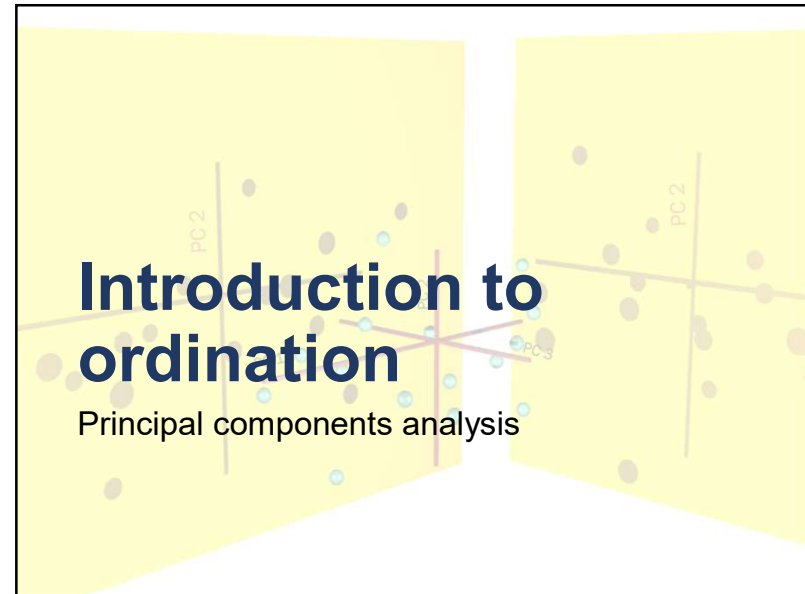
Questions?



27

Introduction to ordination

Principal components analysis



28

What is ordination?

- A method for ordering multivariate data along newly constructed variables (hence the name, **ordination**)
- Uses correlations among variables to collapse them into fewer composite variables that still explain a lot of variation in the dataset (i.e., it's a data reduction technique)
- E.g., collapse highly correlated `iris$Petal.Length` and `iris$Petal.Width` into one composite variable

29

Why do ordination?

- Used primarily to explore MV data, but can also be used for hypothesis testing and prediction
 - Great way to visualize multidimensional MV data on a few axes (e.g., two or three)
- Can distill multiple correlated variables into one
 - can use as IV or DV in plots & linear models
 - E.g., collapse collinear IVs into one composite IV
- Composite variables produced by ordination are uncorrelated → can be used as IVs in multiple regression (i.e., no collinearity)

30

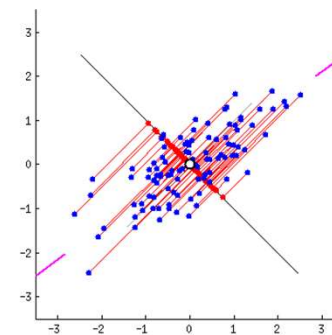
Principal components analysis

- PCA is ordination done on continuous variables
- How it works conceptually:
 1. Center variables & fit line through axis of greatest variation in variables

31

1. Fit line through axis of greatest variation

- Done by minimizing errors in X **AND** Y (i.e., shortest distance from each point to line)



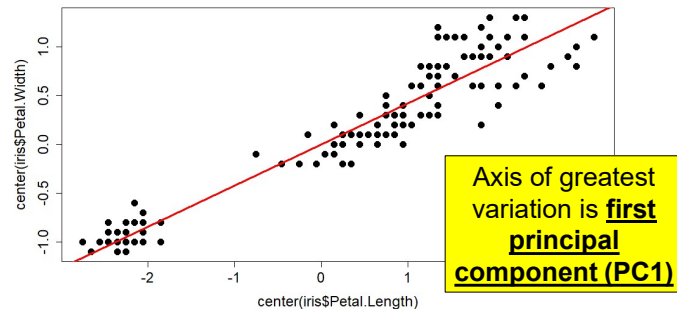
- Must pass through centroid
- *Exactly* the same as major axis regression

<https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues/2700>

32

1. Fit line through axis of greatest variation

- E.g., `iris$Petal.Width` & `iris$Petal.Length`



33

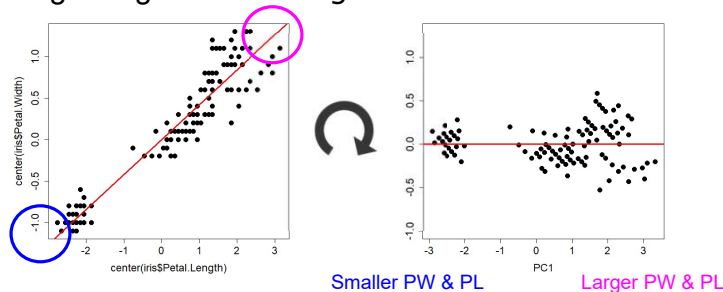
Principal components analysis

- PCA is ordination done on continuous variables
- How it works conceptually:
 1. Center variables & fit line through axis of greatest variation in variables
 2. Rotate plot, so PC1 is now on x-axis

34

2. Rotate plot

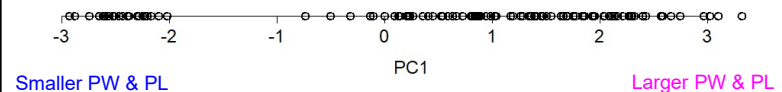
- Larger PC1 scores → larger `Petal.Length` and `Petal.Width`, and vice versa
- Points closer in PCA space are more similar, regarding `Petal.Length` and `Petal.Width`



35

Data reduction

- Distilled two highly correlated variables into one (PC1)!
- Can now represent two variables with one (and thus one axis) that still captures 99% of the variation in the original MV dataset (to be explained later)



36

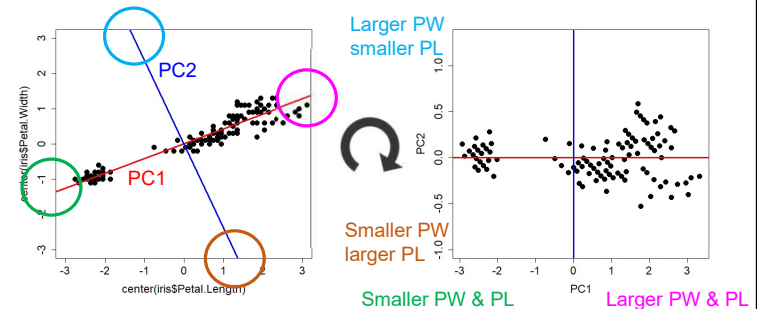
Principal components analysis

- PCA is ordination done on continuous variables
- How it works conceptually:
 1. Center variables & fit line through axis of greatest variation in variables
 2. Rotate plot, so PC1 is now on x-axis
 3. Subsequent PCs (e.g., PC2) are perpendicular to previous ones & explain residual (less) variation from previous PCs
 - Only have as many PCs as you do variables

37

3. Subsequent PCs

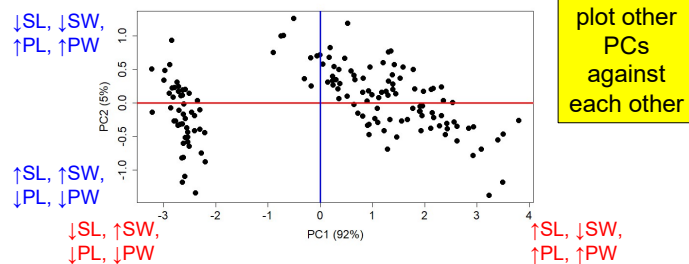
- PC2 is perpendicular to (i.e., independent of) PC1
- More variation along PC1 → more variation in data explained by PC1 (99%)



38

Visualizing many variables

- PCA most useful for visualizing >3 variables, which cannot be plotted (i.e., need >3 axes)
- E.g., Sepal.Length, Sepal.Width, Petal.Length, Petal.Width



39

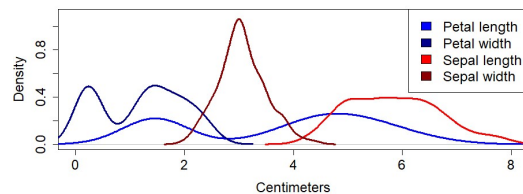
Things to note thus far

- Can only have as many PCs as variables (though the later ones might explain very little variation in variables)
- PCs are perpendicular to and independent of each other (each can be thought of as explaining a different dimension of the data)
- Variation explained by PC1 > variation explained by PC2 > variation explained by PC3, etc.

40

One more thing...

- Because PC1 is fit through axis of greatest variation, PC1 will be dominated by larger variables (which have more variation)
- Thus, it's common practice to scale variables first if they differ in units or orders of magnitude



Doesn't apply
to iris
dataset

41

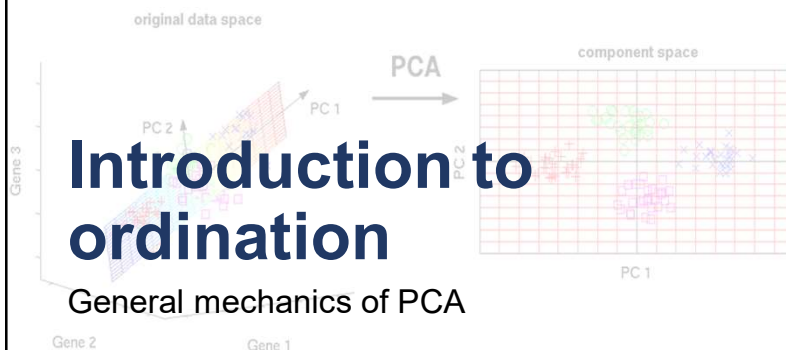
Questions?



42

Introduction to ordination

General mechanics of PCA



43

PCA steps

1. Center variables & fit line through axis of greatest variation in variables
 2. Rotate plot, so PC1 is now on x-axis
 3. Subsequent PCs (e.g., PC2) are perpendicular to previous ones & explain residual (less) variation from previous PCs
- What PCA is *actually* doing is **singular value decomposition (SVD)** of the **variance-covariance matrix** of variables (not important)

44

Eigenvalues & eigenvectors

- SVD produces two main quantities we're interested in:
 - Eigenvalues:** amount of variance in variables explained by each PC
 - Decreases as you go from PC1 to PC2 to PC3, etc.
 - Eigenvectors:** how each PC is related to the original variables

45

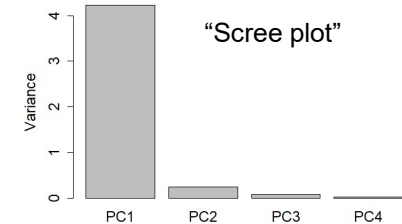
Eigenvalues



- iris dataset has four variables → four PCs → four eigenvalues
- Sum of eigenvalues = summed variance of each variable in original dataset

PC1 = 4.23
PC2 = 0.24
PC3 = 0.08
PC4 = 0.02

4.57



46

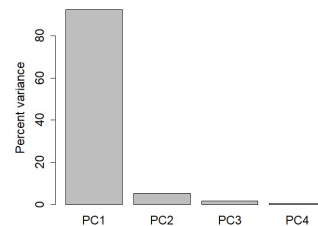
Eigenvalues



- iris dataset has four variables → four PCs → four eigenvalues
- Can rescale to get % variance explained by each PC

PC1 = $100 \times 4.23 / 4.57 = 92\%$
PC2 = $100 \times 0.24 / 4.57 = 5\%$
PC3 = $100 \times 0.08 / 4.57 = 1.7\%$
PC4 = $100 \times 0.02 / 4.57 = 0.5\%$

100%



47

Eigenvectors



- AKA **loadings**: tell us how each original variable contributes to each PC
- Ranges from -1 to 1, where sign indicates direction of relationship btw. variable and PC & magnitude indicates strength of "correlation"

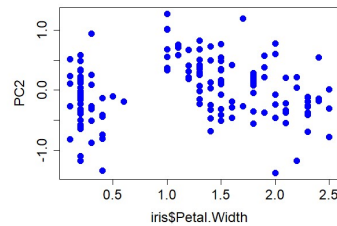
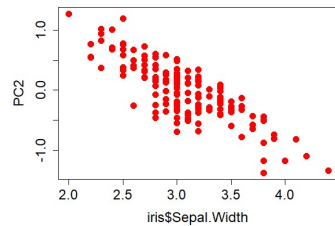
	PC1	PC2	PC3	PC4
Sepal.Length	0.36	-0.66	0.58	0.32
Sepal.width	-0.08	-0.73	-0.60	-0.32
Petal.Length	0.86	0.17	-0.08	-0.48
Petal.width	0.36	0.08	-0.55	0.75

48

Eigenvectors



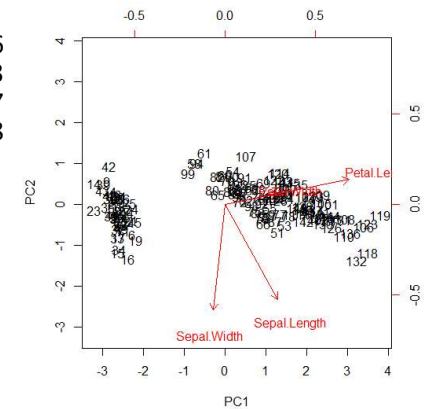
	PC1	PC2	PC3	PC4
Sepal.Length	0.36	-0.66	0.58	0.32
Sepal.Width	-0.08	-0.73	-0.60	-0.32
Petal.Length	0.86	0.17	-0.08	-0.48
Petal.Width	0.36	0.08	-0.55	0.75



49

Plotting of eigenvectors

	PC1	PC2
Sepal.Length	0.36	-0.66
Sepal.Width	-0.08	-0.73
Petal.Length	0.86	0.17
Petal.Width	0.36	0.08



50

Eigenvectors

- Also describes how to transform data from original variable coordinate system to PCA space and back
 - Scores**: coordinates of each data point in PCA space
- E.g., to get PC1 score of Plant #1, multiply plant's measurement for each variable (after centering) by corresponding PC1 loading and then sum everything

51

Eigenvectors



Plant #1 (after centering)

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
-0.7	0.4	-2.36	-1.0

Loadings

	PC1
Sepal.Length	0.36
Sepal.Width	-0.08
Petal.Length	0.86
Petal.Width	0.36

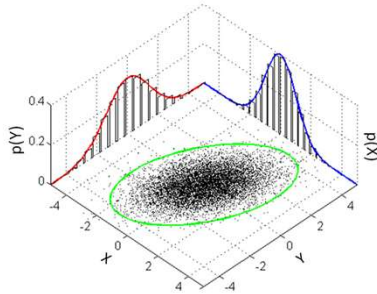
$$\text{Plant \#1 PC1 score} = -0.7 \times 0.36 + 0.4 \times -0.08 + -2.36 \times 0.86 + -1.0 \times 0.36 = -2.68$$

Loadings act as weights!
Quantifies how much each variable linearly contributes to PC score

52

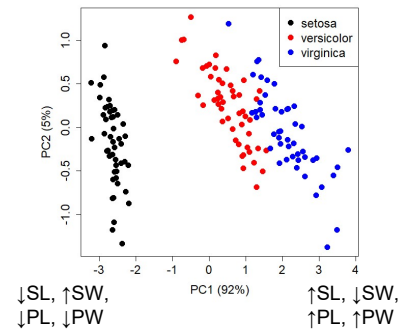
PCA assumptions

- Variables are multivariate normally distributed



53

Example of a hypothesis test



- Does PC1 sig. differ among species?
- ANOVA: $P < 2.2e-16$
- Or how does PC1 vary ~ N level?

54

Questions?



55

Summary

- MV models: multiple DVs ~ one or more IVs
 - Takes into account correlation between DVs
- Ordination distills multiple variables into a few important, independent axes (good for visualization!)
- PCA is ordination for continuous variables
 - Eigenvalues: variance explained by each PC
 - Eigenvectors: loadings tell us how much each variable contributes to each PC

56