



## Latent Variables (Variates) and Multicollinearity

**“Good or Bad?”**

*Comparisons between Principal Components Analysis,  
Factor Analysis and Discriminant Analysis*

**Melinda K. Higgins, Ph.D.**

*[Principal Research Scientist – Georgia Tech Research Institute &  
Senior Statistician/Assistant Research Prof – Emory University]*

**7 February 2008**

*Latent Variates and Multicollinearity*



## Outline

- I. **Multicollinearity**
- II. **Latent Variates**
- III. **IRIS Dataset Described and Visualizations**
- IV. **Principal Components Analysis**
- V. **Factor Analysis**
- VI. **Discriminant Analysis**
- VII. **Summary**
- VIII. **Statistical Resources and Help**

*Latent Variates and Multicollinearity*

# I. Multicollinearity

**Multicollinearity exists when there is a strong correlation between two or more variables (predictors) in a regression (and other) model. Perfect collinearity exists when at least one variable is a perfect linear combination of the others, e.g. correlation coefficient = 1.\***

### DISADVANTAGES:

- **POOR INTERPRETATION**
  - For two perfectly correlated variables, their associated regression coefficients ("b") are interchangeable.
- **UNSTABLE PREDICTION MODEL**
  - As the collinearity between two variables increases so does the standard error of the "b"s.
- **REJECTION OF GOOD PREDICTORS**
  - High levels of multicollinearity increase the probability that a good predictor variable will be found non-significant and rejected from the model.

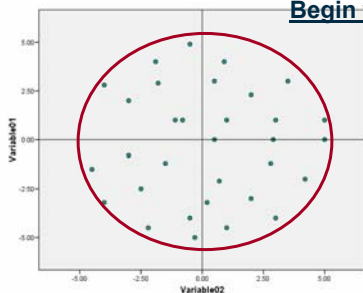
### ADVANTAGES:

- **SIMPLIFY MODEL**
  - When two variables are highly correlated one of them may be deleted from the model without loss of prediction capability.
- **IDENTIFY UNDERLYING CONSTRUCTS**
  - The presence of multicollinearity can indicate that there is an underlying "construct" or "latent" variable which may be important for understanding and interpretability.
- **REDUCE MODEL DIMENSIONALITY**
  - Rather than deleting one of the "redundant" highly correlated variables, they can both (all) be replaced with their underlying "latent" variable (reduce the "dimensionality" of the model).

\*Field, Andy. *Discovering Statistics Using SPSS, 2<sup>nd</sup> Ed.*, SAGE Publications, London, 2005, p. 174.

# I. Multicollinearity

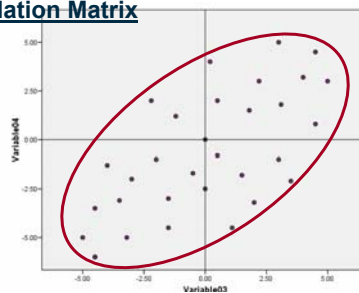
## Begin with Correlation Matrix



Correlations

		Variable01	Variable02
Variable01	Pearson Correlation	1	.016
	Sig. (2-tailed)		.932
	N	32	32
Variable02	Pearson Correlation	.016	1
	Sig. (2-tailed)	.932	
	N	32	32

**Almost 0 Correlation**



Correlations

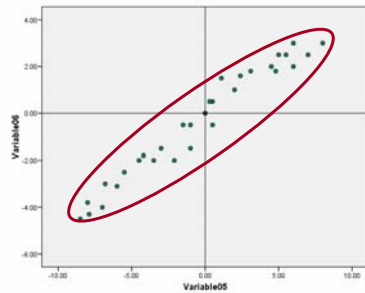
		Variable03	Variable04
Variable03	Pearson Correlation	1	.666**
	Sig. (2-tailed)		.000
	N	32	32
Variable04	Pearson Correlation	.666**	1
	Sig. (2-tailed)	.000	
	N	32	32

\*\* . Correlation is significant at the 0.01 level (2-tailed).

**Medium Correlation**

## Latent Variables and Multicollinearity

## I. Multicollinearity



Now that we have “measured the extent” of the multicollinearity ...

– how do we “extract” the underlying “latent variates?”

Correlations

		Variable05	Variable06
Variable05	Pearson Correlation	1	.979**
	Sig. (2-tailed)		.000
	N	32	32
Variable06	Pearson Correlation	.979**	1
	Sig. (2-tailed)	.000	
	N	32	32

\*\* . Correlation is significant at the 0.01 level (2-tailed).

**Almost Perfect (1) Correlation**

Latent Variates and Multicollinearity

## II. Latent Variates

**What are they? – Latent Variates are things/constructs which are present but which can not be measured directly.**

For example, psychologists might want to measure “burnout” (person working long periods of time on a project suddenly finds themselves devoid of motivation and inspiration), but this can not be measured directly. Instead, stress levels, depression, and creativity (number of ideas during some period of time) might be measured instead.

### USES:

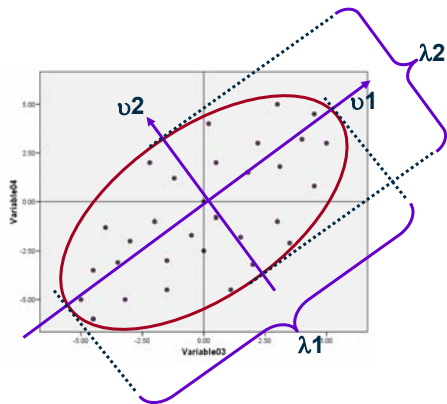
- Understand the “structure” of the variables
  - (e.g. Spearman and Thurston used Factor Analysis to understand the latent variate “intelligence”)
- Construct a questionnaire to measure “latent variate”
  - (e.g. develop a questionnaire to measure “burnout”)
- Reduce the dataset to a more manageable size
  - (e.g. replace stress, depression and creativity variables with one “burnout” variate/measure”)

Latent Variates and Multicollinearity

## II. Latent Variates

How do we “measure the extent of” and then “extract” the Latent Variates?

Find the Eigenvalues and Eigenvectors of the Correlation\* Matrix



What are Eigenvectors and Eigenvalues?

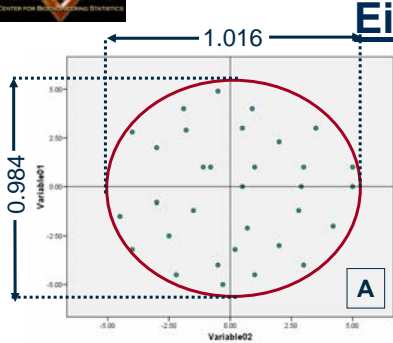
**Eigenvectors ( $v$ )** are the major and minor axes (2D) of the “ellipsoid” encompassing the variability in the data

**Eigenvalues ( $\lambda$ )** are the lengths of these axes across the “ellipsoid”

\* NOTE: The Covariance and SSCP Matrices are also sometimes used

Latent Variates and Multicollinearity

## Eigenvalues



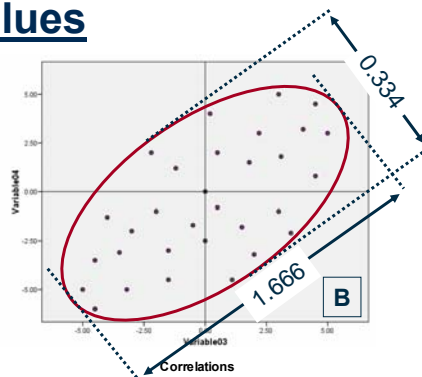
Correlations

Variable01	Pearson Correlation	Variable01	Variable02
		1	.016
	Sig. (2-tailed)		.932
	N	32	32
Variable02	Pearson Correlation	.016	1
	Sig. (2-tailed)	.932	
	N	32	32

Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	1.016	50.789	50.789
2	.984	49.211	100.000

Extraction Method: Principal Component Analysis.



Correlations

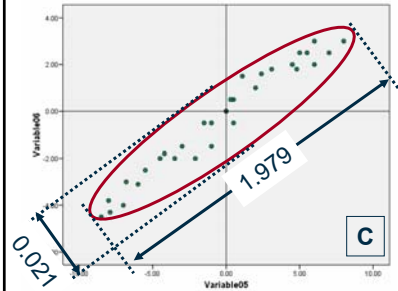
Variable03	Pearson Correlation	Variable03	Variable04
		1	.666
	Sig. (2-tailed)		.000
	N	32	32
Variable04	Pearson Correlation	.666	1
	Sig. (2-tailed)	.000	
	N	32	32

Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	1.666	83.296	83.296
2	.334	16.704	100.000

Extraction Method: Principal Component Analysis.

## Eigenvalues (cont'd)



Other important indicators of multicollinearity (stuff you'll see in statistics software)

- **Condition Index =  $\lambda_{max} / \lambda_{min}$**   
Increases with increasing multicollinearity

- (A) Cond Index = 1.0325
- (B) Cond Index = 5.0006
- (C) Cond Index = 94.2381**

Correlations

		Variable05	Variable06
Variable05	Pearson Correlation	1	.979**
	Sig. (2-tailed)		.000
	N	32	32
Variable06	Pearson Correlation	.979**	1
	Sig. (2-tailed)	.000	
	N	32	32

- **Determinant (of Correlation Matrix)**

Decreases with Condition Index. As Determinant  $\rightarrow 0$ , the Correlation Matrix can not be inverted and error messages will be displayed.

- (A) Determinant = 0.9997
- (B) Determinant = 0.5555
- (C) Determinant = 0.0416**

Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	1.979	98.945	98.945
2	.021	1.055	100.000

Extraction Method: Principal Component Analysis.

## Eigenvectors

[Previously] Eigenvectors ( $v$ ) are the major and minor axes (2D) of the "ellipsoid" encompassing the variability in the data

The eigenvectors are also the new "latent variates" which are linear combinations of the original variables (var1, var2, ...)

$$\text{Latentvar1} = v_{11}(\text{Var1}) + v_{12}(\text{Var2})$$

$$\text{Latentvar2} = v_{21}(\text{Var1}) + v_{22}(\text{Var2})$$

Component Score Coefficient Matrix

	Component	
	1	2
Variable03	.548	-1.223
Variable04	.548	1.223

Extraction Method: Principal Component Analysis.

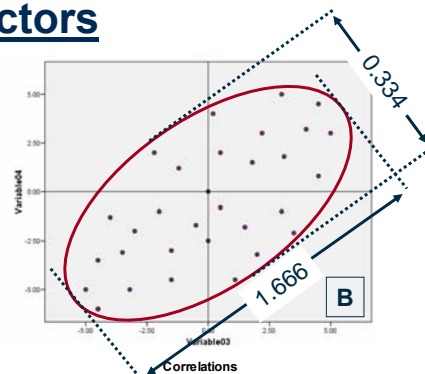
Component Scores.

$$\text{Latentvar1} = 0.548(\text{Var03}) + 0.548(\text{Var04})$$

$$\text{Latentvar2} = -1.223(\text{Var03}) + 1.223(\text{Var04})$$

Scores

Loadings (Coefficients or Weights)



Correlations

		Variable03	Variable04
Variable03	Pearson Correlation	1	.666**
	Sig. (2-tailed)		.000
	N	32	32
Variable04	Pearson Correlation	.666**	1
	Sig. (2-tailed)	.000	
	N	32	32

Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	1.666	83.296	83.296
2	.334	16.704	100.000

Extraction Method: Principal Component Analysis.

### III. IRIS Dataset

Fisher, R.A. "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 1936, 7, 179-188.



Latent Variates and Multicollinearity

### III. IRIS Dataset

- Iris Dataset
- 3 Species (50 samples each)
  - Iris Setosa
  - Iris Versicolor
  - Iris Virginica
- 4 Features
  - Sepal Length
  - Sepal Width
  - Petal Length
  - Petal Width

Sample	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
1	50	33	14	2	1
2	44	30	13	2	1
3	49	31	15	2	1
4	52	41	15	1	1
5	50	34	16	4	1
...	...	...	...	...	...
51	60	22	40	10	2
52	69	31	49	15	2
53	61	30	46	14	2
54	57	28	41	13	2
55	61	28	47	12	2
...	...	...	...	...	...
101	77	26	69	23	3
102	58	28	51	24	3
103	71	30	59	21	3
104	63	29	56	18	3
105	72	30	58	16	3

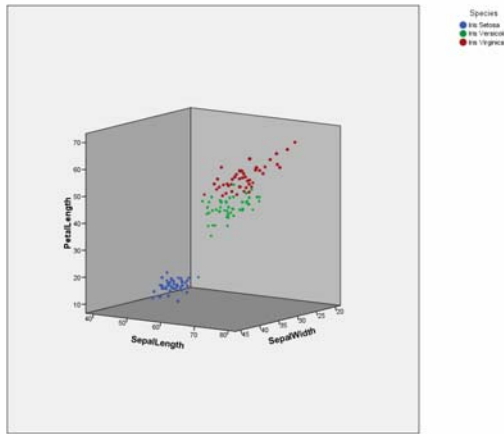
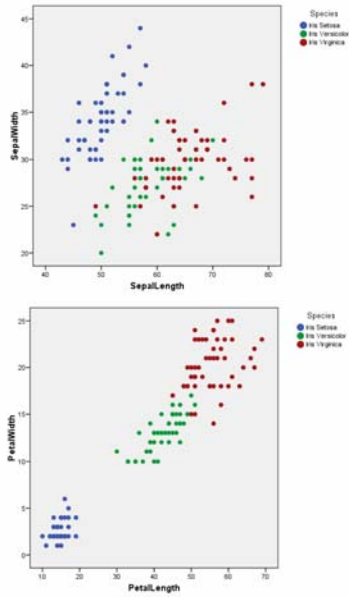
Species	SepalLength	SepalWidth	PetalLength	PetalWidth
1 Iris Setosa	50.06	34.28	14.62	2.46
2 Iris Versicolor	59.36	27.70	42.60	13.26
3 Iris Virginica	65.88	29.74	55.52	20.26
Total	58.43	30.57	37.58	11.99

- Notice that Petal Length and Petal Width changes (increases) quite a bit from Species 1 < 2 < 3
- Notice also that as the flowers get bigger, the Sepal Width gets a little smaller (Species 2 < 3 < 1)

Data Table (150 x 4)

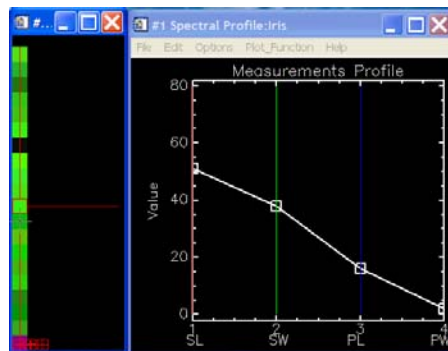
Latent Variates and Multicollinearity

### Make Plots – Visualize the Data



Latent Variates and Multicollinearity

Samples Shown as “Pixels”



Profile Plots of SL, SW, PL, PW

View All Variables at once – Look for Patterns

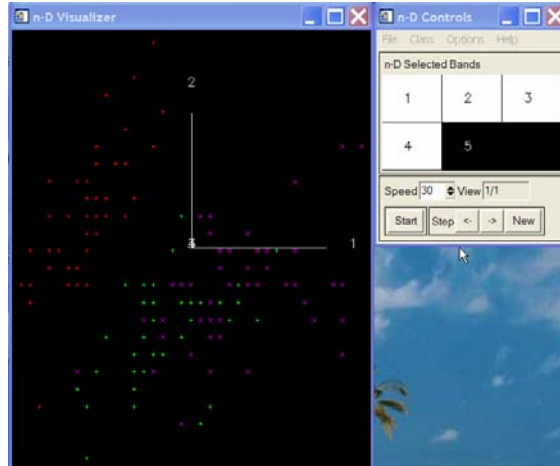
Note changes in “pattern/profile” from “green/species 1” to “purple/species 2”

Latent Variates and Multicollinearity

## How to View Patterns in Multiple Dimensions (e.g. > 3 variables)

Red = Iris Setosa  
 Green = Iris Versicolor  
 Purple = Iris Virginica

Axis 1 = Sepal Length  
 Axis 2 = Sepal Width  
 Axis 3 = Petal Length  
 Axis 4 = Petal Width



*Latent Variates and Multicollinearity*

## III. IRIS Dataset

Open IrisData.sav @  
[S:\Shared\Statistics\\_MKHiggins\ResearchRoundtables\Lecture\\_12Oct2007](S:\Shared\Statistics_MKHiggins\ResearchRoundtables\Lecture_12Oct2007)

SampleNum	SepalLength	SepalWidth	PetalLength	PetaWidth	Species
1	1.00	50	33	14	2
2	2.00	44	30	13	2
3	3.00	49	31	15	2
4	4.00	52	41	15	1
5	5.00	50	34	16	4
6	6.00	51	35	14	2
7	7.00	43	30	11	1
8	8.00	50	32	12	2
9	9.00	47	32	16	2
10	10.00	51	37	15	4
11	11.00	47	32	13	2
12	12.00	57	36	17	3
13	13.00	48	30	14	1
14	14.00	44	29	14	2
15	15.00	55	42	14	2
16	16.00	54	34	15	4
17	17.00	50	36	14	2
18	18.00	49	30	14	2

*Latent Variates and Multicollinearity*

## Run Bivariate Correlation

Correlations

		SepaLength	SepaWidth	PetaLength	PetaWidth
SepaLength	Pearson Correlation	1	-.118	.872**	.818**
	Sig. (2-tailed)		.152	.000	.000
	N	150	150	150	150
SepaWidth	Pearson Correlation	-.118	1	-.428**	-.366**
	Sig. (2-tailed)	.152		.000	.000
	N	150	150	150	150
PetaLength	Pearson Correlation	.872**	-.428**	1	.963**
	Sig. (2-tailed)	.000	.000		.000
	N	150	150	150	150
PetaWidth	Pearson Correlation	.818**	-.366**	.963**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	150	150	150	150

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Notice that Petal Length and Petal Width are highly correlated (>0.8) with Sepal Length as well as each other and significant

Notice that Petal Length and Petal Width are significantly negatively correlated with Sepal Width

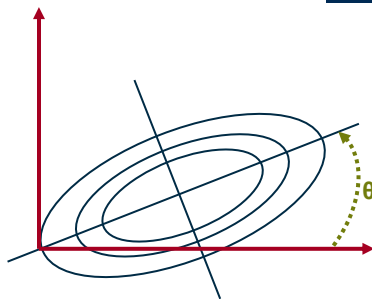
Notice that Sepal Length is also negatively correlated with Sepal Width, not significant

### Latent Variates and Multicollinearity

## IV. Principal Components Analysis

Extraction of the Latent Variates – simple orthogonal (perpendicular, 90°) rotation of original variables

### Axis Rotation



“Given a set of  $p$ -variate observations, determine a rotated axis such that the variable (*latent variate*) thereby defined as a linear combination of the original  $p$  variables has maximum variance.”\*

*i.e. find the eigenvectors and eigenvalues of the correlation matrix*

\*Tatsuoka, M. *Multivariate Analysis, 2<sup>nd</sup> Ed.*, MacMillan Publishing Co., New York, 1971, p. 128.

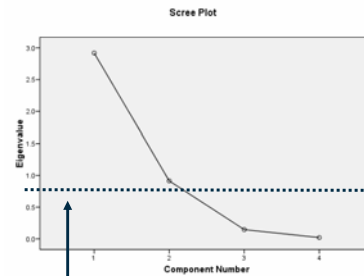
### Latent Variates and Multicollinearity

## IV. Principal Components Analysis

Correlation Matrix<sup>a</sup>

Correlation	SepaLength	SepaWidth	PetaLength	PetaWidth
	1.000	-.118	.872	.818
		1.000	-.428	-.366
			1.000	.963
				1.000
Sig. (1-tailed)	SepaLength		.000	.000
	SepaWidth	.076	.000	.000
	PetaLength	.000	.000	.000
	PetaWidth	.000	.000	.000

a. Determinant = .008 ← Also indicates high multicollinearity



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.918	72.962	72.962	2.918	72.962	72.962
2	.914	22.851	95.813	.914	22.851	95.813
3	.147	3.669	99.482	.147	3.669	99.482
4	.021	.518	100.000			

Extraction Method: Principal Component Analysis.

Keep 2 "components" – accounts for 95% of variance

### Latent Variates and Multicollinearity

## IV. PCA – Plotting the Data on the Latent Variates ("Score Plots")

Component Score Coefficient Matrix

	Component	
	1	2
SepaLength	.305	.395
SepaWidth	-.158	.966
PetaLength	.340	.026
PetaWidth	.331	.070

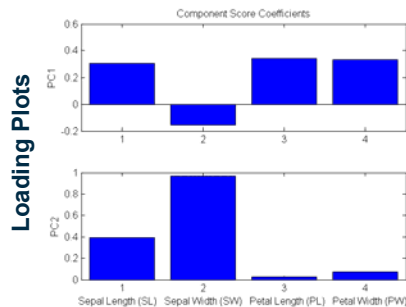
Extraction Method: Principal Component Analysis.

Component Scores.

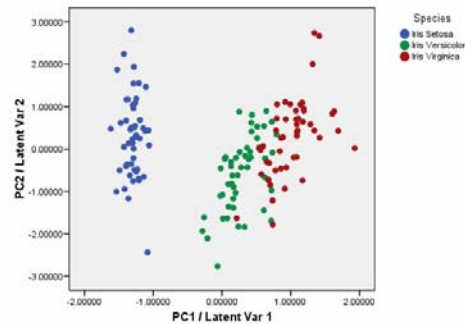
Going from 4 dimensions (SL,SW,PL,PW) to only 2 (PC1,2) !! Latent Variates called PC's

$$PC\ 1 = 0.305(SL) - 0.158(SW) + 0.340(PL) + 0.331(PW)$$

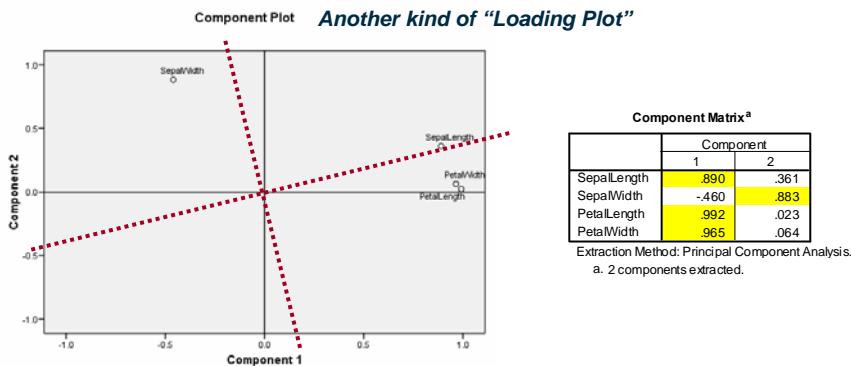
$$PC\ 2 = 0.395(SL) + 0.966(SW) + 0.026(PL) + 0.070(PW)$$



PCA - Score Plot



## IV. PCA – Understanding the Latent Variates

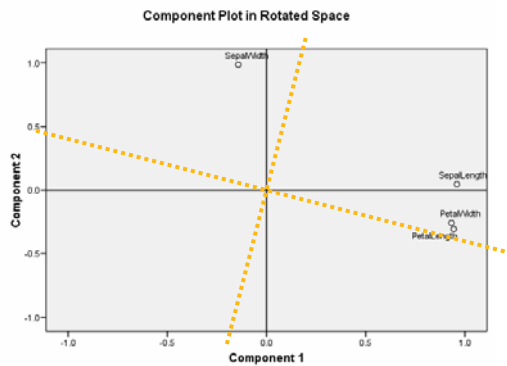
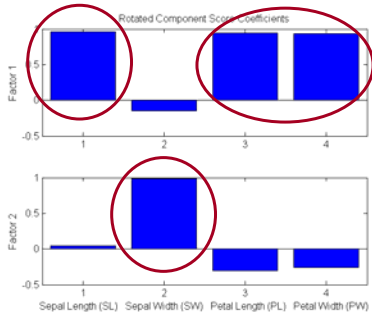
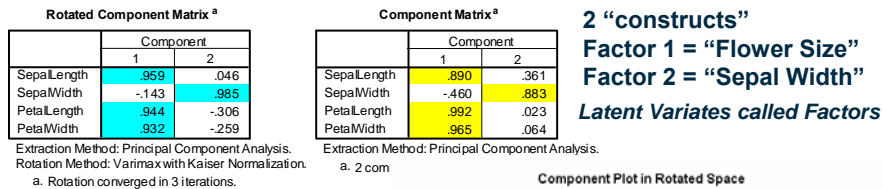


What if we "rotated" these Principal Components to better separate out the original variables and align them more distinctly with the 2 Latent Variates (e.g. hopefully making them more interpretable).

### Latent Variates and Multicollinearity

## V. Factor Analysis Improving "Interpretation" through Rotation

Varimax Rotation – orthogonal (rigid) rotation

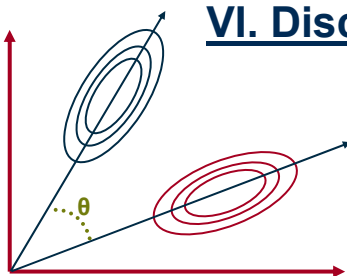


## VI. Discriminant Analysis

- **GOAL** – to extract latent variates (“discriminants”) which maximize the separation of the classification groups (e.g. maximize the between group variance to the within group variance).
- **Advantages**
  - **Extracted model can be used to “classify” new samples/subjects**
  - **Data reduction possible through extracted discriminants**
  - **Extracted latent variates DO NOT have to be orthogonal (90°) to one another – results are usually oblique axes**
- **Disadvantages**
  - **Requires a priori knowledge of “group classifications”**
  - **Requires training set of data to establish model (more samples needed)**

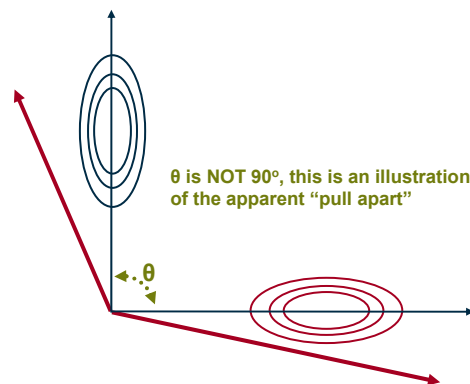
Latent Variates and Multicollinearity

## VI. Discriminant Analysis



Essentially we are trying to find a new set of axes (most likely not orthogonal to each other) which “pull” the groups apart.

Discriminant Analysis attempts to “discriminate” between a set of K groups by maximizing the between groups variance to the within groups variance.



Latent Variates and Multicollinearity

## VI. Discriminant Analysis

**Eigenvalues**

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	32.192 <sup>a</sup>	99.1	99.1	.985
2	.285 <sup>a</sup>	.9	100.0	.471

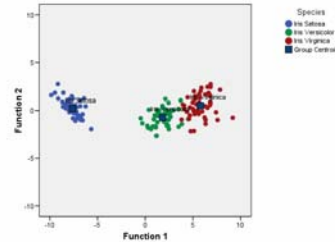
a. First 2 canonical discriminant functions were used in the analysis.

**Canonical Discriminant Function Coefficients**

	Function	
	1	2
SepaLength	-.083	.002
SepaWidth	-.153	.216
PetaLength	.220	-.093
PetaWidth	.281	.284
(Constant)	-2.105	-6.661

Unstandardized coefficients

**Canonical Discriminant Functions**



**Functions at Group Centroids**

Species	Function	
	1	2
1 Iris Setosa	-7.608	.215
2 Iris Versicolor	1.825	-.728
3 Iris Virginica	5.783	.513

Unstandardized canonical discriminant functions evaluated at group means

*Latent Variates and Multicollinearity*

## VII. Summary

- **Visualize your data**
  - **Make bivariate (2D) and trivariate (3D) plots**
  - **Look for patterns in samples/subjects across all variables**
- **Run a bivariate correlation**
  - **look for significant correlations and correlations close to 1.0**
  - **Pay attention to the “signs” (positive, negative)**
- **Principal Components Analysis**
  - **Useful for extracting eigenvalues and eigenvectors**
  - **Initial first-step to understanding how many latent variates there may be in the data**
  - **May not yield interpretable latent variates**

*Latent Variates and Multicollinearity*



## VII. Summary

- **Factor Analysis**
  - Multiple options for rotating initial principal components such that the extracted latent variates optimally separate different constructs (e.g. better individual correlations between original variables and latent variate) – better interpretability
  - Helps to optimize design of questionnaires seeking measurement of underlying constructs
- **Discriminant Analysis**
  - Useful for optimizing separation of known groups and/or sample clusters
  - Orthogonality not required
  - Apriori knowledge required for group classification and training set required (e.g. more samples needed)

*Latent Variates and Multicollinearity*



## VIII. Statistical Resources and Help

SON S:\Shared\Statistics\_MKHiggins

Shared resource for all of SON – faculty and students

Will continually update with tip sheets (for SPSS, SAS, and other software), lectures (PPTs and handouts), datasets, other resources and references

Website in development

[S:\Shared\Statistics\\_MKHiggins\website2\index.htm](S:\Shared\Statistics_MKHiggins\website2\index.htm)

Contact

Dr. Melinda Higgins

Melinda.higgins@emory.edu

Office: 404-727-5180 / Mobile: 404-434-1785

*Latent Variates and Multicollinearity*