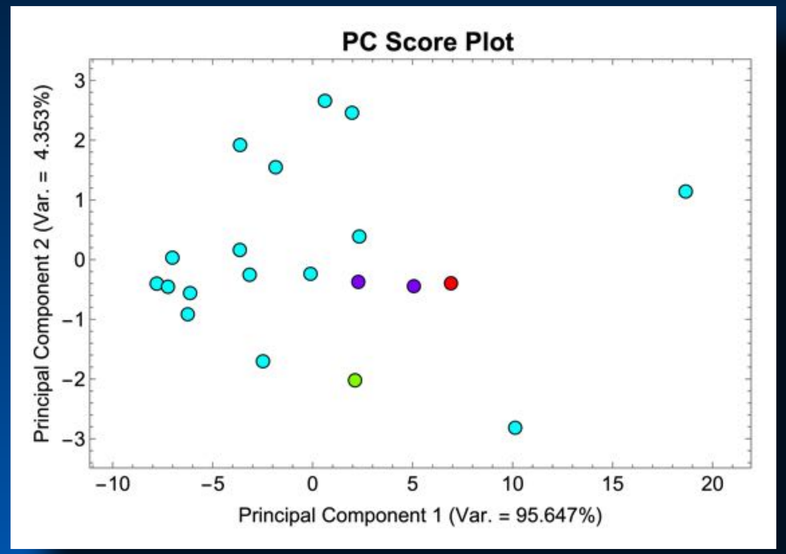
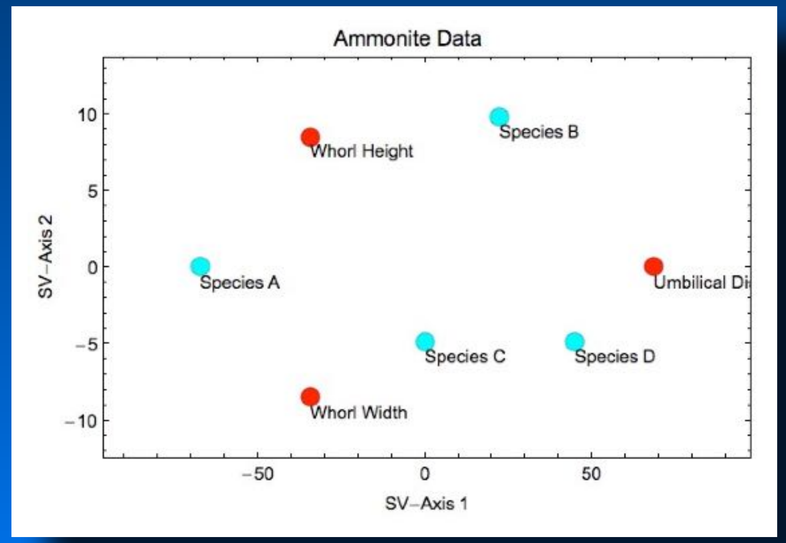
Dimensionality Reduction

Prof. Norman MacLeod
School of Earth Sciences & Engineering, Nanjing University



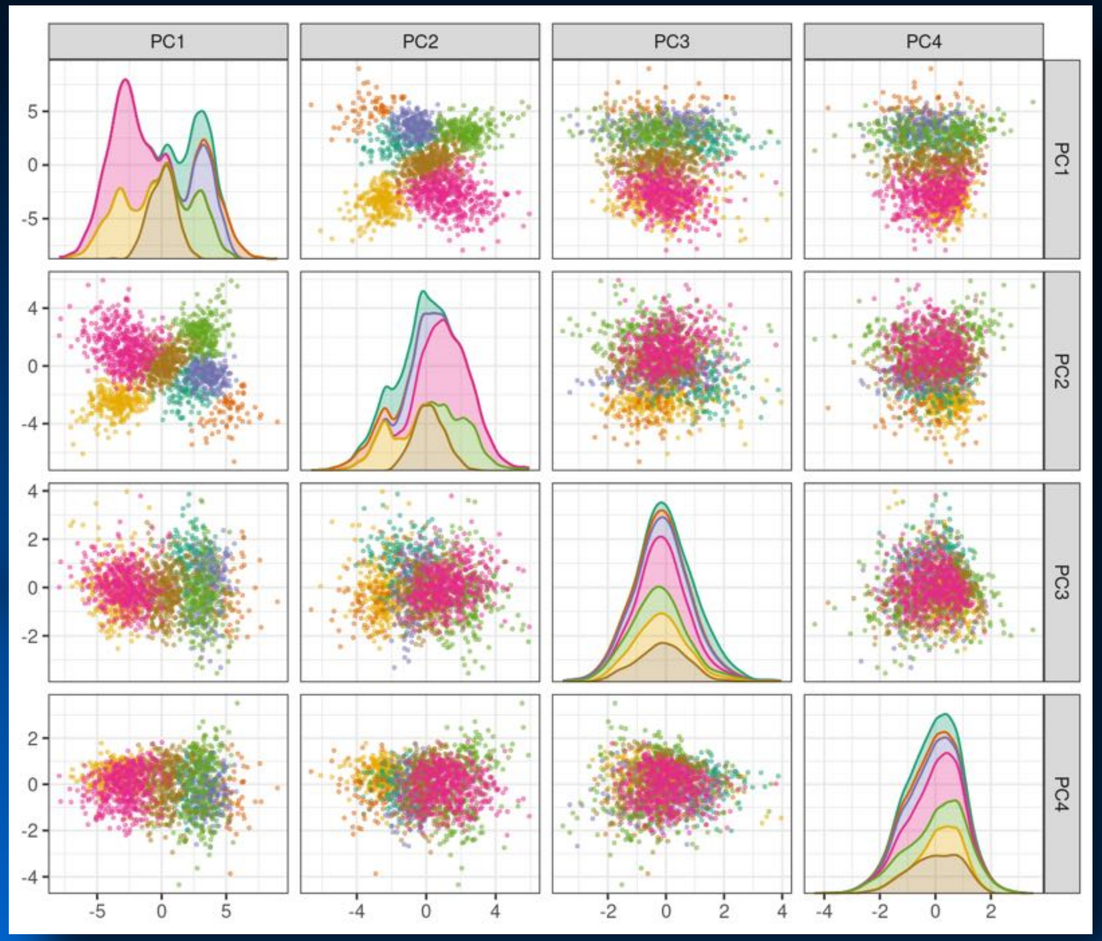
$$d_{ij} = \frac{\frac{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|}{Range_k}}{p}$$

$$X = \begin{pmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{pmatrix}$$



Dimensionality Reduction

The multivariate methods described in this lecture have been discussed traditionally under the heading of “ordination methods”. However, the more modern, and more accurate term for them is “dimensionality reduction methods”.



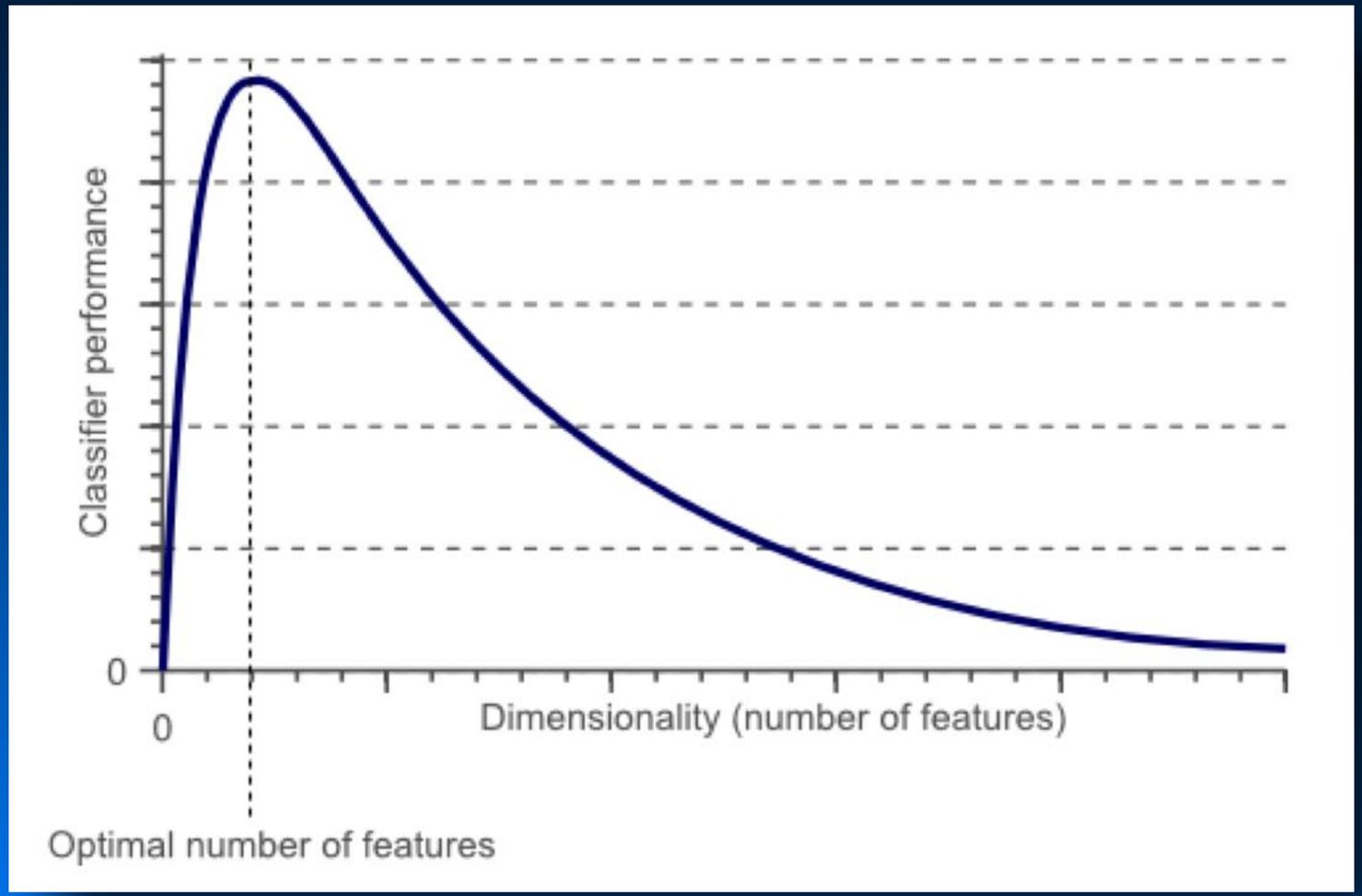
Their primary purpose is to take high-dimensional data and portray relations between objects and/or variables in a lower dimensional space that preserves the overall character of their high-dimensional structure. In addition to this reclassification, a substantial number of new DR methods have appeared of late since high-dimensional data have become very common in many areas of science while the need to visualize them is a constant challenge.

Dimensionality Reduction

Curse of Dimensionality

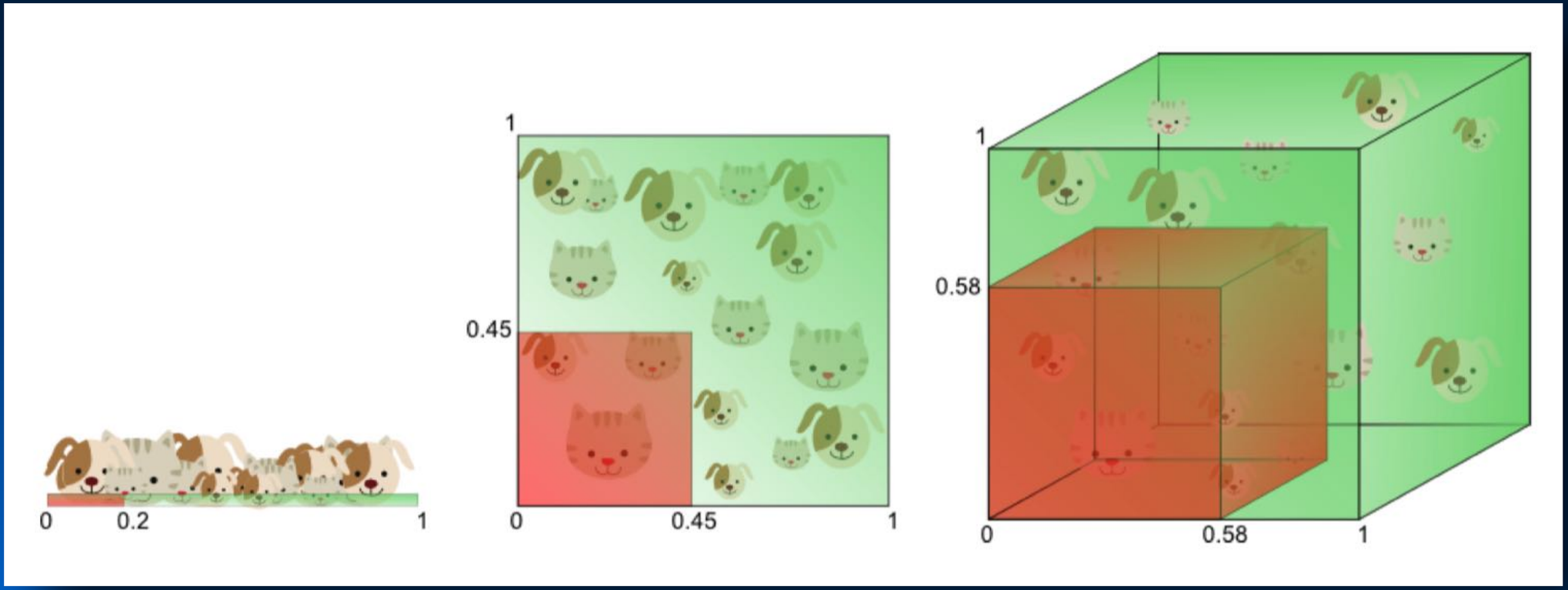
Dimensionality Reduction

Performance of Linear Classifiers



Dimensionality Reduction

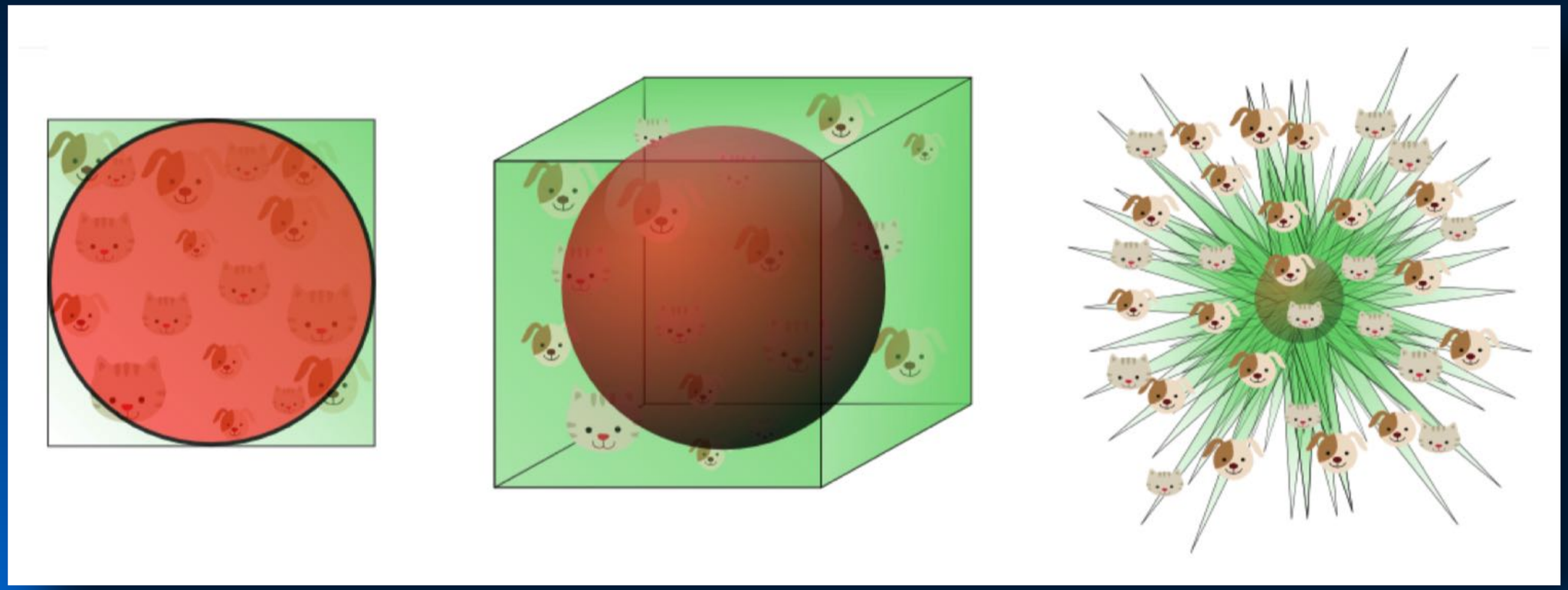
Curse of Dimensionality: Analysis



As the number of dimensions increases the number of samples required to provide reasonable coverage of any given proportion of the feature space increases exponentially with the volume of the feature space.

Dimensionality Reduction

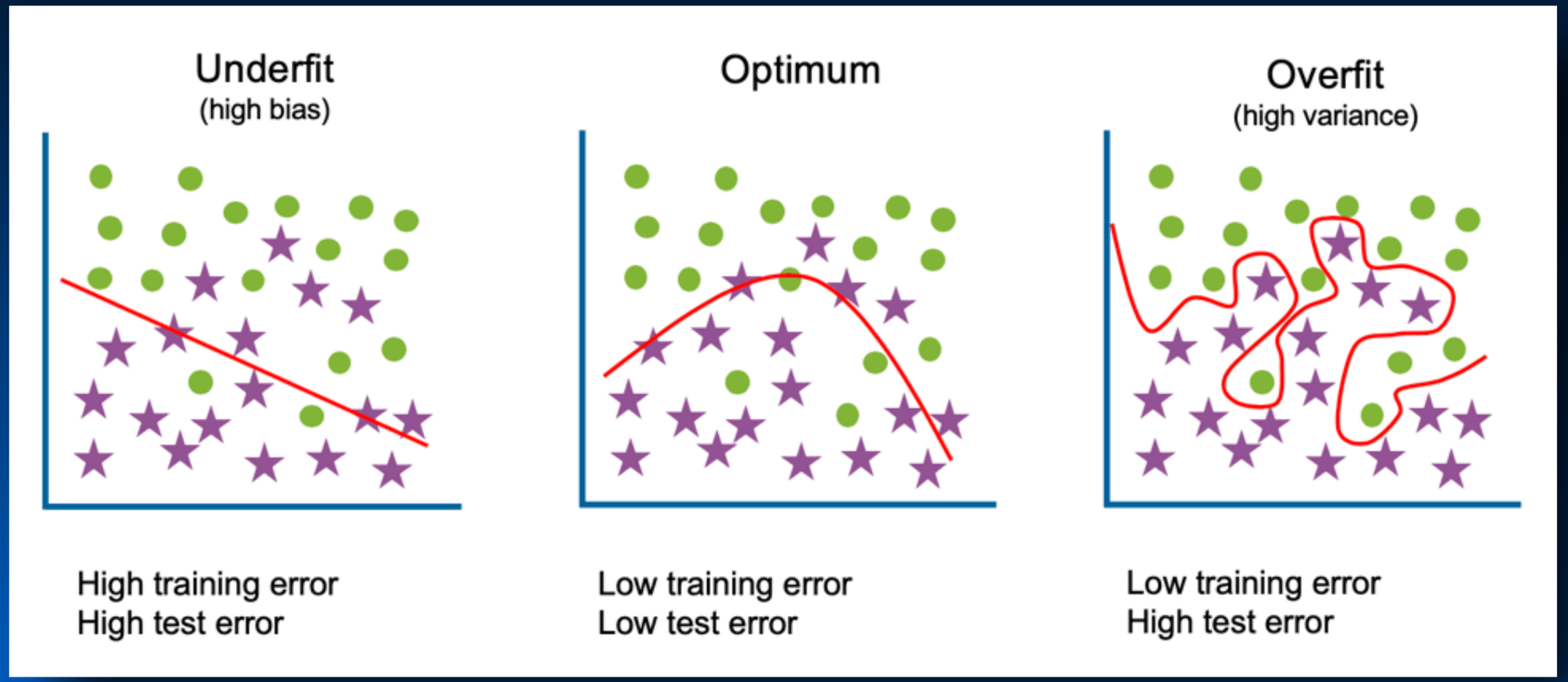
Curse of Dimensionality: Analysis



Moreover, as dimensionality increases, a greater proportion of data resides in the corners of the feature hyperspace rather than in its center. Linear classifiers require central points to create accurate models of variation.

Dimensionality Reduction

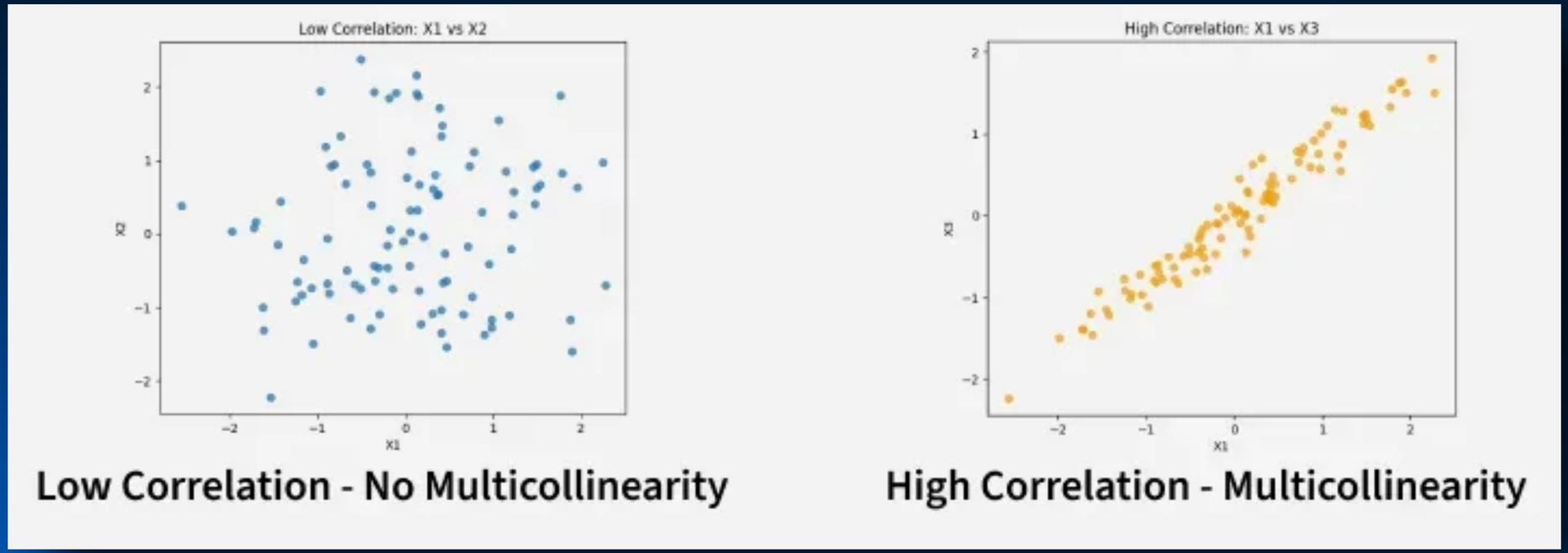
Overfitting



As dimensionality increases there is a greater tendency for multivariate data-analysis methods (esp. classifiers) to overfit the data. This leads both to poor generalizability in terms of the characterization of high dimensional data and poor classification performance on independent datasets.

Dimensionality Reduction

Mulicolinearity



As dimensionality increases there is also a greater likelihood of high correlations existing between variables or sets of variables. This existence of structured high correlations (multicollinearity) makes it difficult for the contributions of individual variables to any trends existing in the data to be assessed accurately. The existence of high correlations among variables also makes it difficult to an accurate representation of the data bivariate and trivariate spaces.

Dimensionality Reduction

Methods Taxonomy

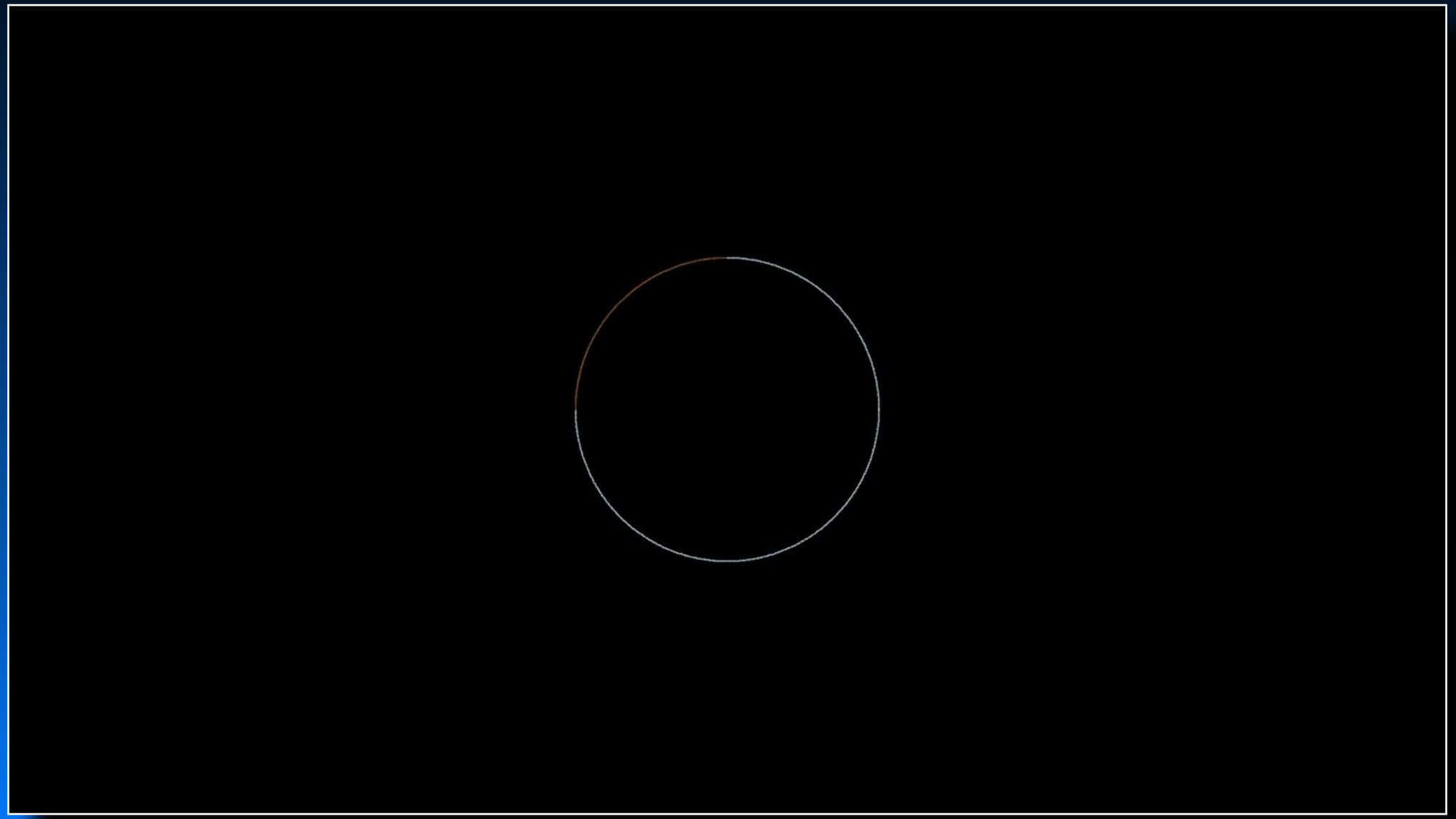
Pooled-Sample Methods - Methods that make no distinctions between any groups of data that may exist in the dataset.

- Principal Component Analysis (PCA)
- Principal Coordinate Analysis (PCoord)
- Singular-Value Decomposition (SVD)
- Multidimensional Scaling (MDS & NL-MDS)
- Uniform Manifold Approximation & Projection (UMAP)

Classification Methods - Methods that attempt to find distinctions between groups of data in the dataset so there can be used to achieve optimal group separation.

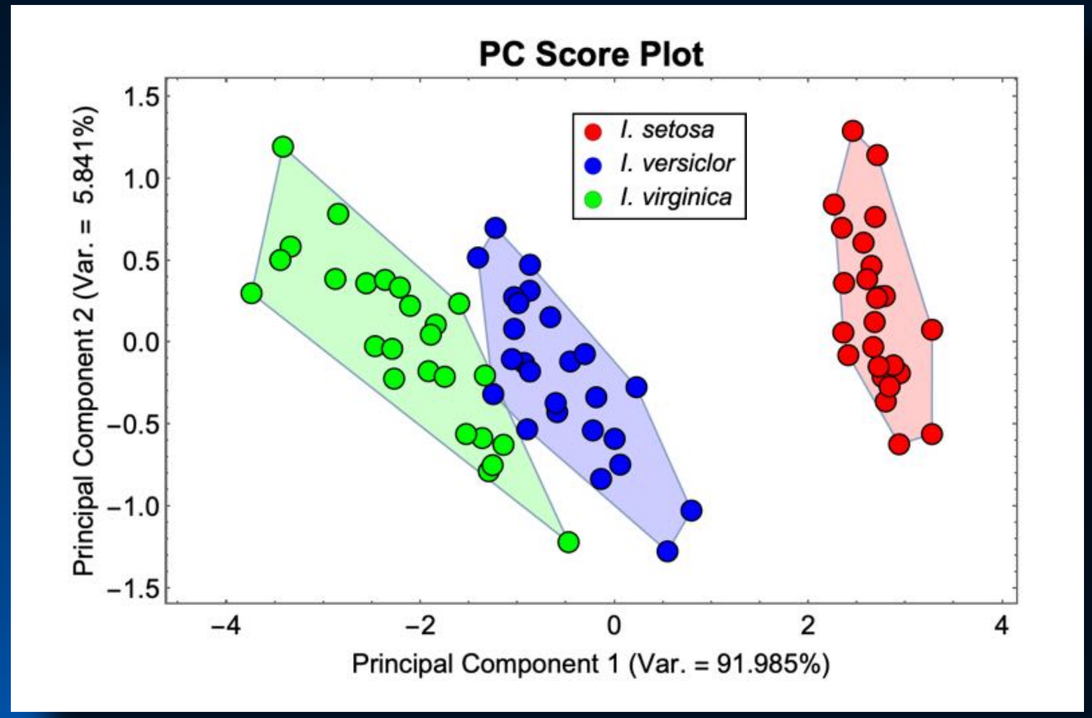
- Linear Discriminant Analysis (LDA)
- Canonical Variates Analysis (CVA)
- Gaussian Unmixing

Eigenvectors



Principal Components Analysis (PCA)

An orthogonal data transformation that projects a set of raw data values onto a set of linearly uncorrelated composite variables (= the principal components = the eigenvectors) the first of which is aligned with the direction of maximum linear variance and ordered such that each successive composite variable (= principal component) is aligned to the maximum residual variation subject to the constraint of orthogonality.



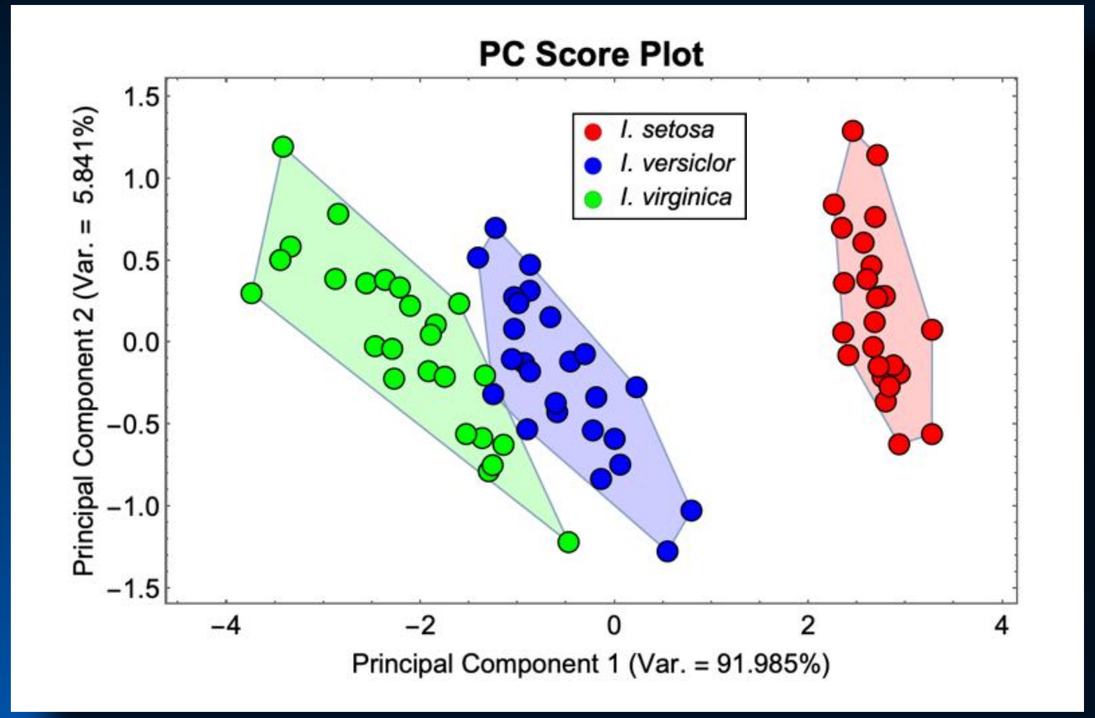
The first few eigenvectors are then used to define the axes of a variable space within which objects can be located by their projection coordinates.

Principal component analysis (PCA) is the multivariate extension of a major axis regression.

Principal Components Analysis (PCA)

Characteristics

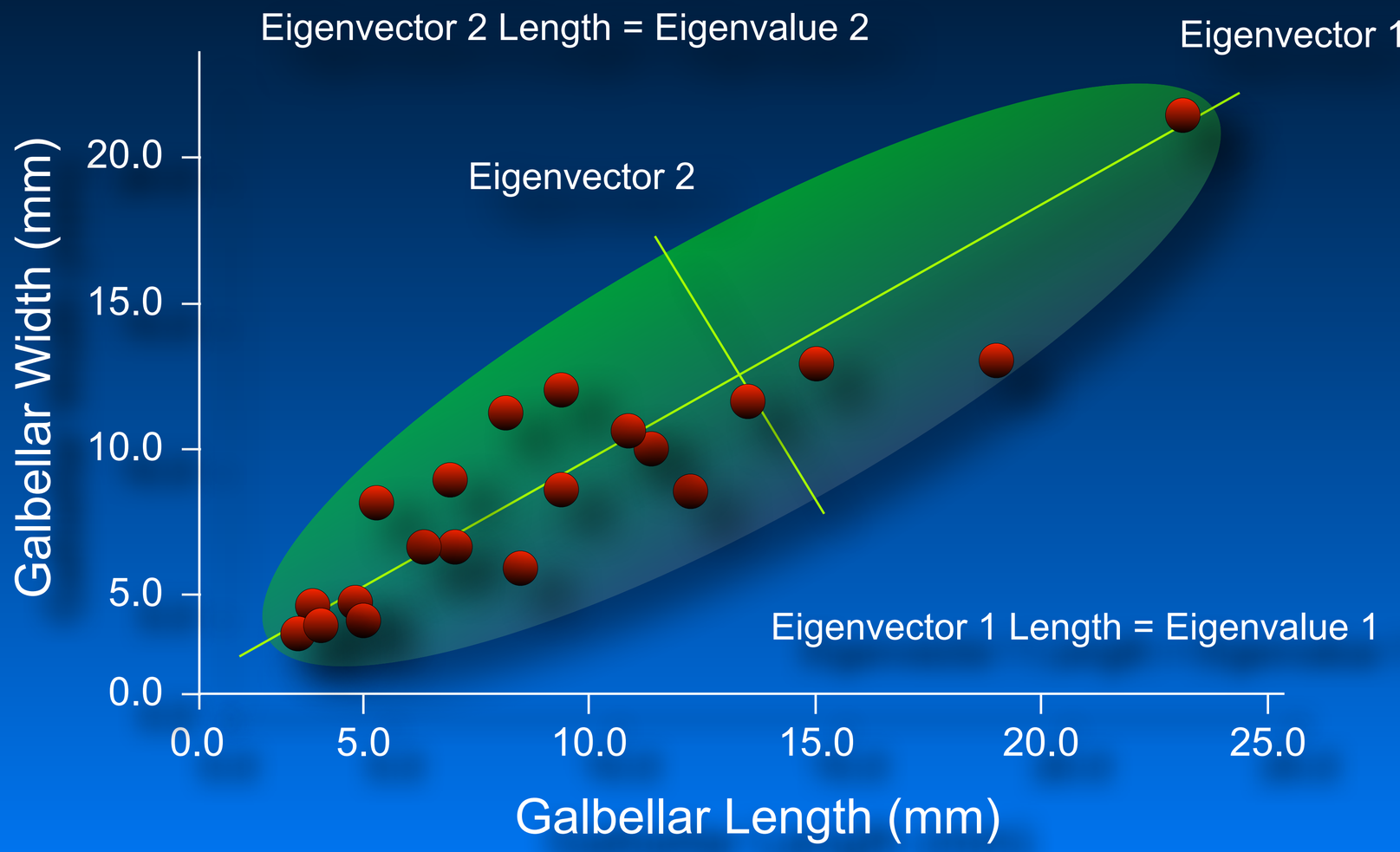
- A data transformation (not a statistical procedure).
- Principal components = eigenvector decomposition of a square, non-singular matrix that expresses covariance relations between variables.
- Involves no distributional assumptions.
- Assumes all observations comprise a single group (makes no between-group distinctions or optimizations).



- Perhaps the most commonly applied multivariate data-analysis procedure.
- Principally used as a linear technique to achieve dimensionality reduction such that information loss is minimized.

Principal Components Analysis

A more-or-less straightforward application of eigenvectors



Principal Components Analysis

What are principal components?

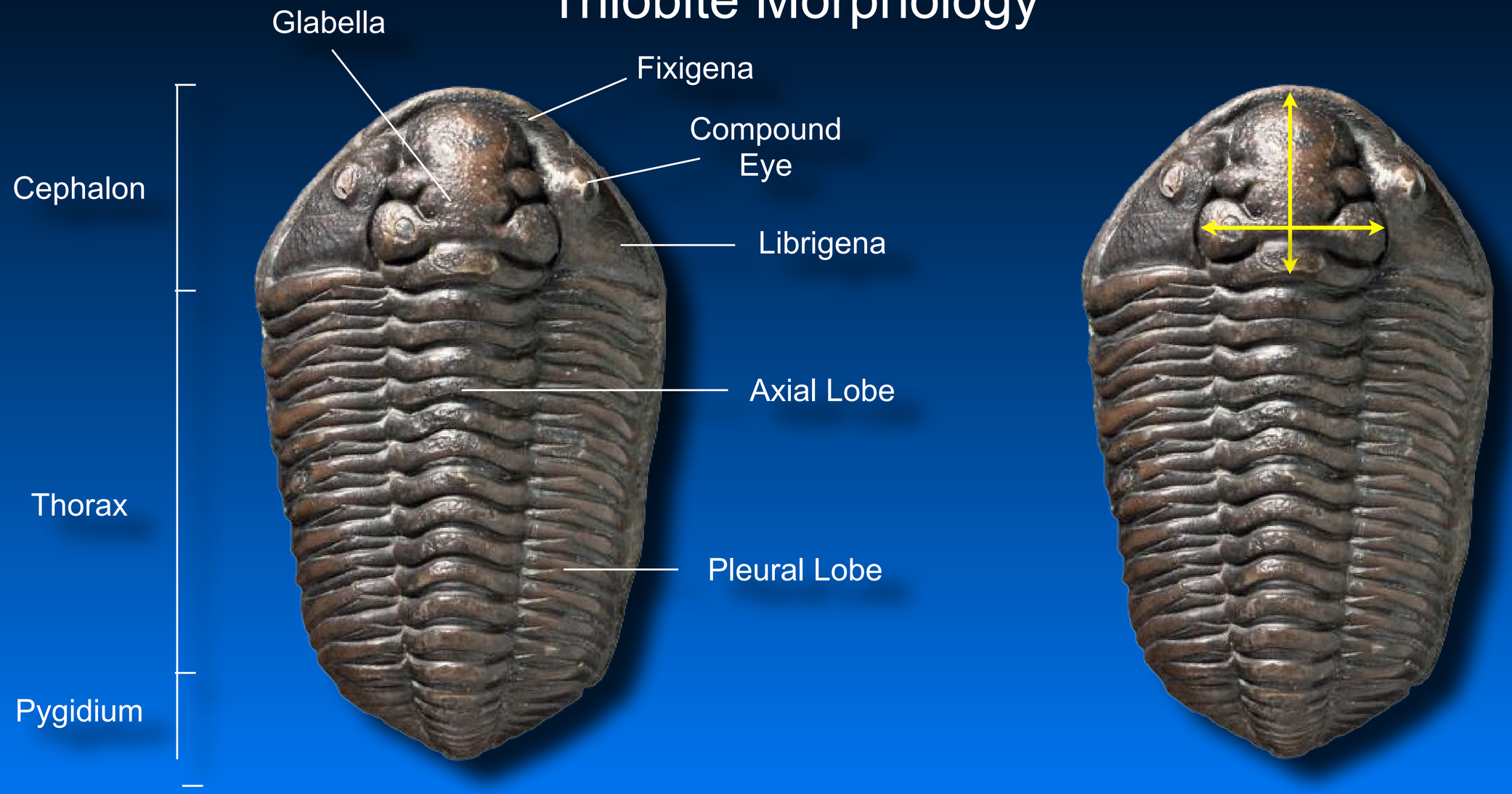
Alternative Concepts

- Pearson Model - uses covariance/correlation relations among variables to estimate underlying pattern-generation ‘factors’ or ‘components’. In other words, the Pearson school believes the principal components are “real” aspects of the data that, as such, can constitute the subjects of deterministic interpretation. This model is a derivative of the model commonly applied to “factor analysis” which is a related multivariate method.

- Morrison Model - uses covariance/correlation relations among variables to create composite variables that are unrelated to one another and optimized along trends expressing maximum variance. In other words, the Morisson school believes principal components are simply the results of a useful mathematical transform-ation of the data.

Principal Components Analysis

Trilobite Morphology



Principal Components Analysis

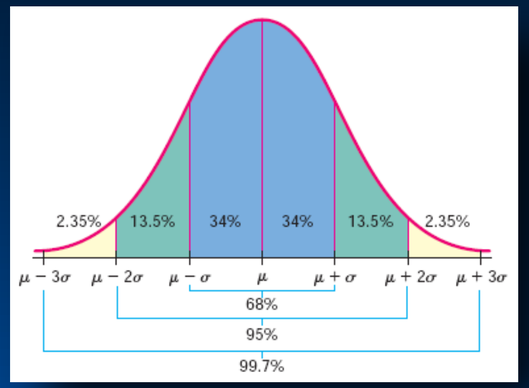
Trilobite Morphology

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78

Principal Components Analysis

Trilobite Morphology

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78



Variance

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

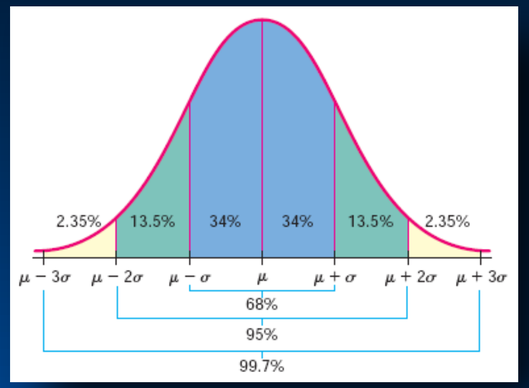
s^2_{length} = 27.33mm²

s^2_{width} = 19.27mm²

Principal Components Analysis

Trilobite Morphology

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78



Std. Deviation

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

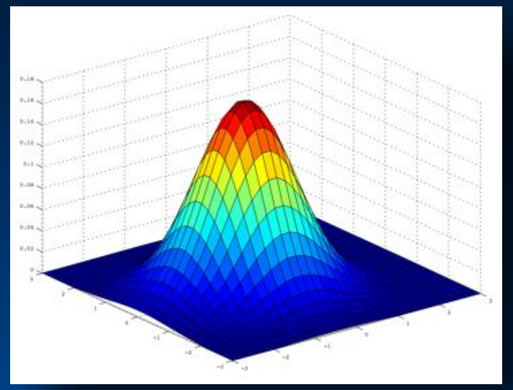
s_{length} = 5.23mm

s_{width} = 4.27mm

Principal Components Analysis

Trilobite Morphology

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78



Covariance

$$COV_{jk} = \frac{\sum_{i=1}^{n} x_{ij}x_{ik} - (\sum_{i=1}^{n} x_{ij} \cdot \sum_{i=1}^{n} x_{ik}}{n(n-1)}$$

$$COV_{length,width} = 20.315$$

Principal Components Analysis

Trilobite Morphology

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78

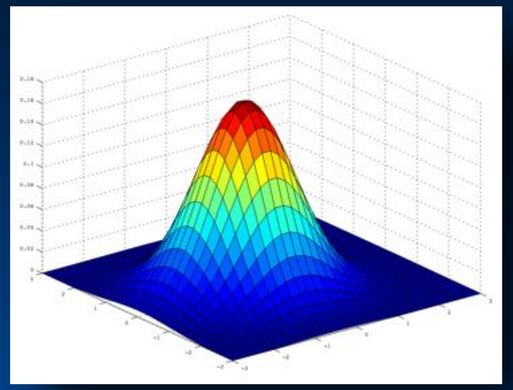
Covariance Matrix

	Glabella Length	Glabella Width
Glabella Length	27.33	20.32
Glabella Width	20.32	19.27

Principal Components Analysis

Trilobite Morphology

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78



Correlation

$$r_{jk} = \frac{\sum_{i=1}^{n} x_{ij} \cdot x_{jk} - \left(\frac{\left(\sum_{i=1}^{n} x_{ij} \cdot \sum_{i=1}^{n} x_{ik}\right)}{n}\right)}{\sqrt{\left(\sum_{i=1}^{n} x_{ij}^2 - \frac{\left(\sum_{i=1}^{n} x_{ij}\right)^2}{n}\right) \cdot \left(\sum_{i=1}^{n} x_{ik}^2 - \left(\frac{\left(\sum_{i=1}^{n} x_{ik}\right)^2}{n}\right)\right)}}$$

$r_{length,width} = 0.91$

Principal Components Analysis

Trilobite Morphology

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78

Covariance Matrix

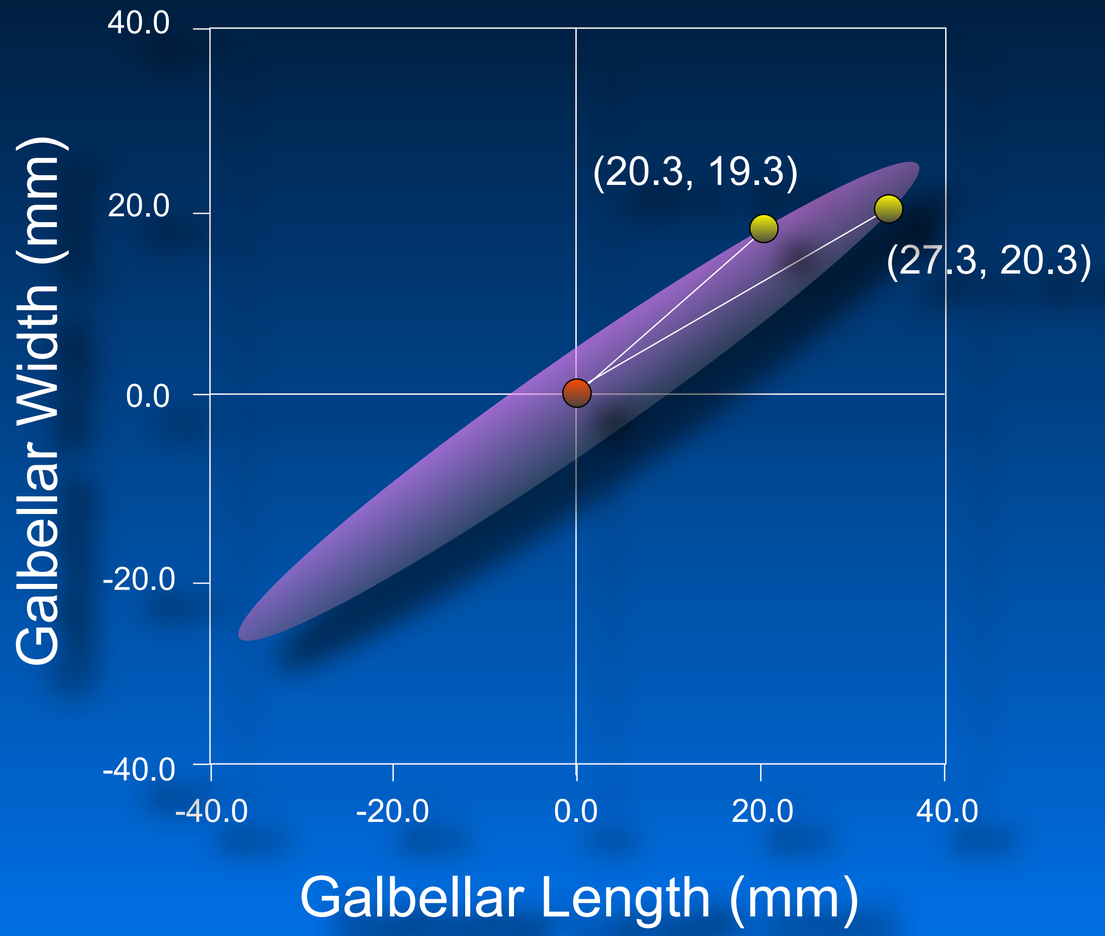
	Glabella Length	Glabella Width
Glabella Length	27.33	20.32
Glabella Width	20.32	19.27

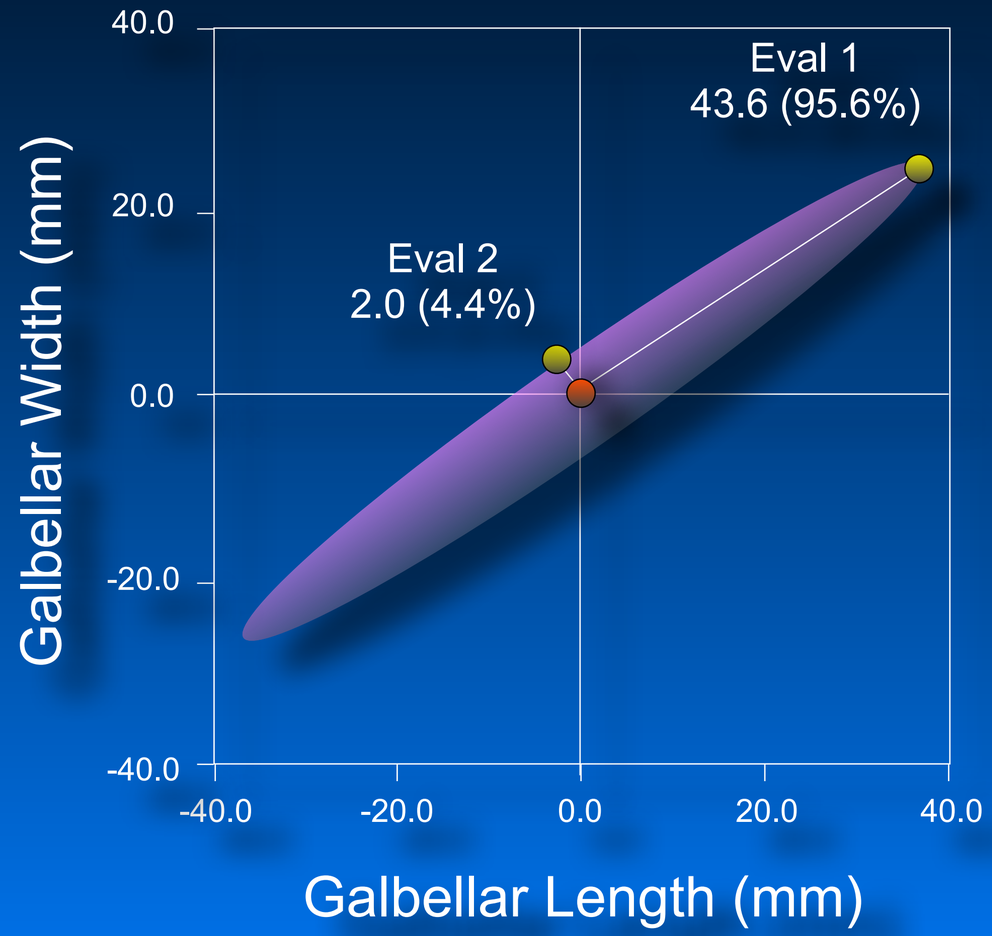
Correlation Matrix

	Glabella Length	Glabella Width
Glabella Length	1.00	0.91
Glabella Width	0.91	1.00

Principal Components Analysis

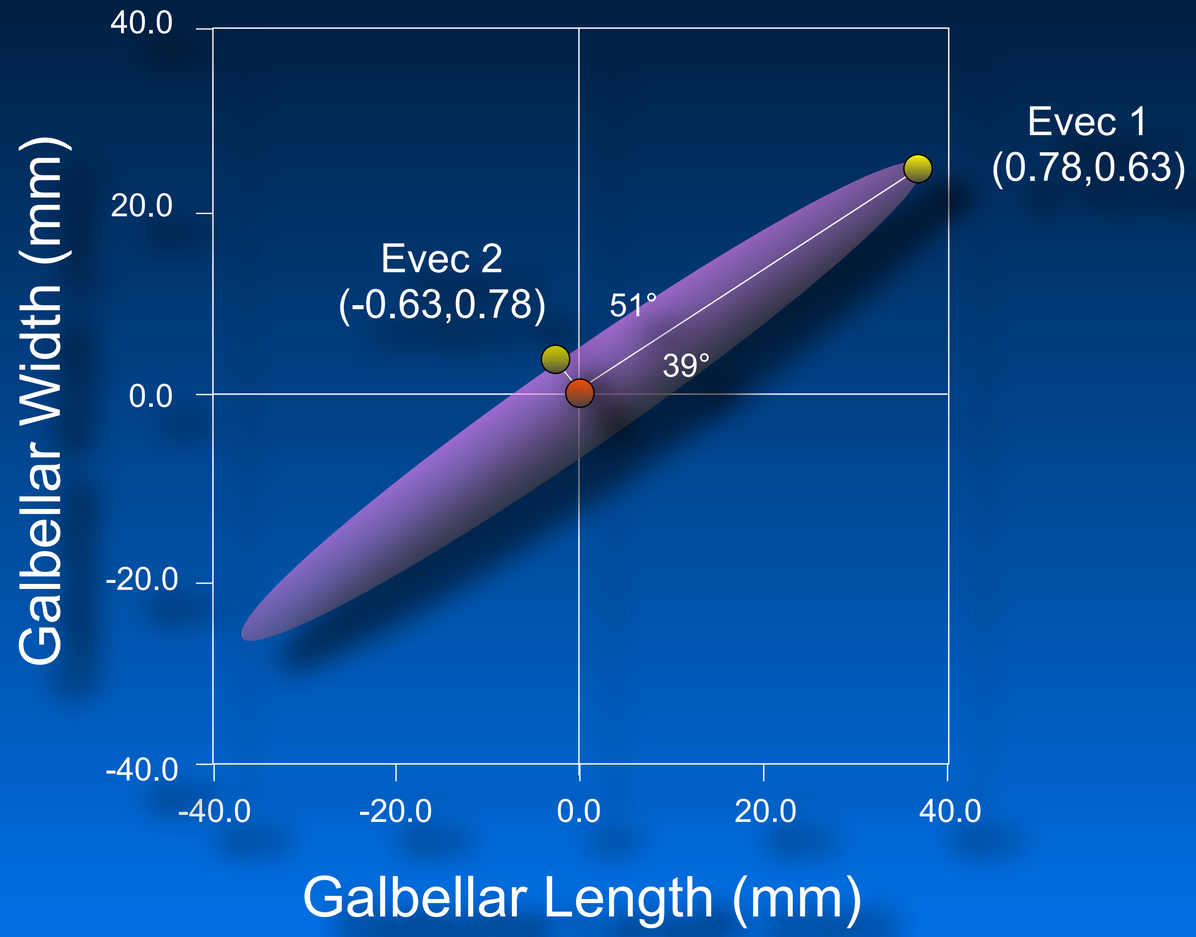
Covariance Structure





Principal Components Analysis

Covariance Structure



Equations of PCA Axes

$$PC_1 = 0.78x_1 + 0.63x_2$$
$$PC_2 = -0.63x_1 + 0.78x_2$$

Principal Components Analysis

Calculation of PCA Scores

$$X = S \,.\, A'^{-1}$$

Where: X = data matrix;

S = matrix of projected data scores in eigenvector space;

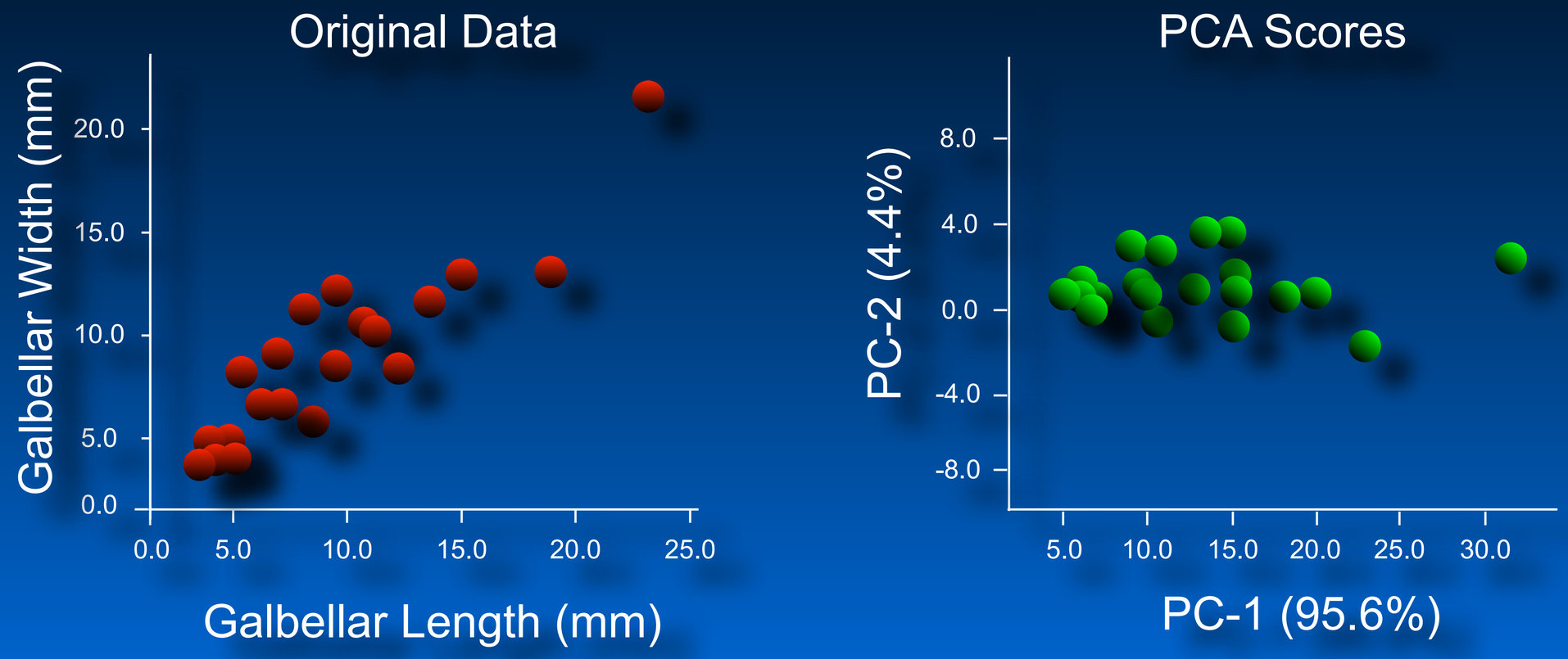
A = eigenvectors of X;

. = the dot product.

$$S = X \,.\, A'$$

Principal Components Analysis

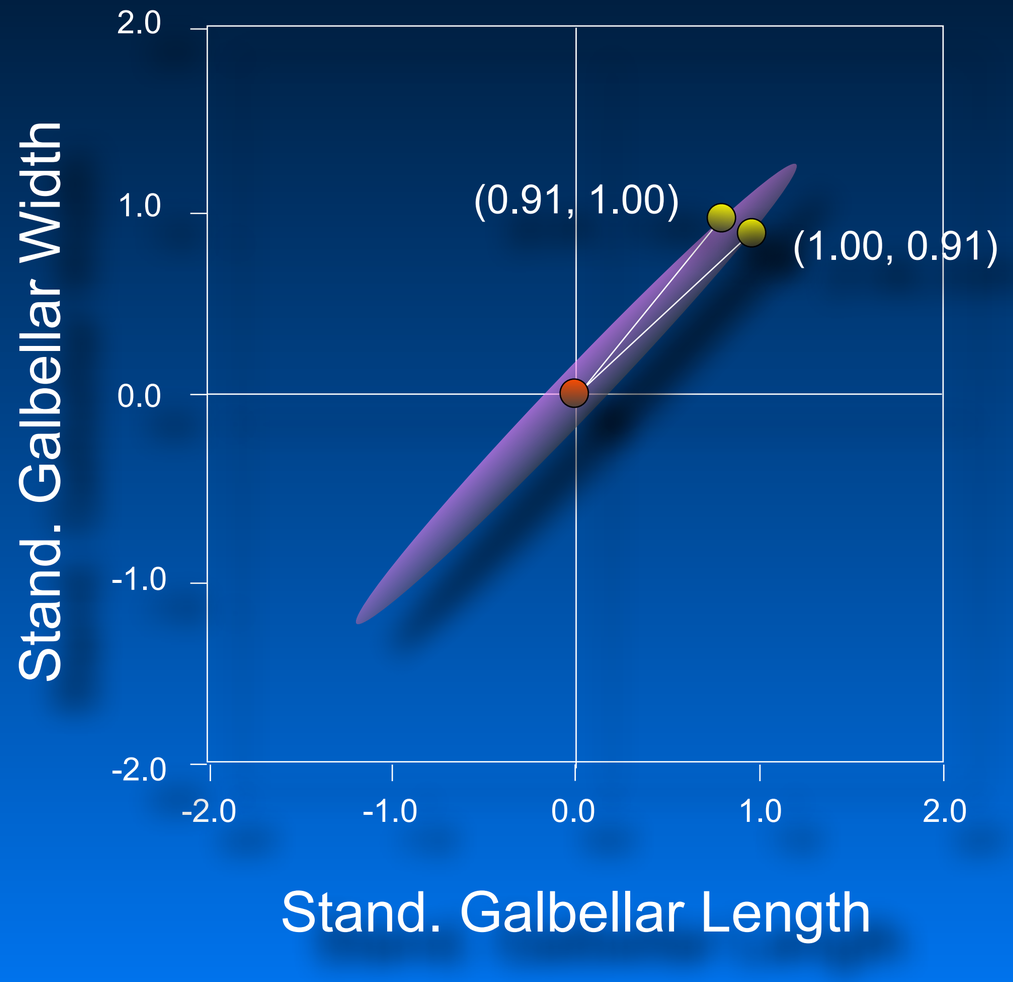
PCA Ordination Space



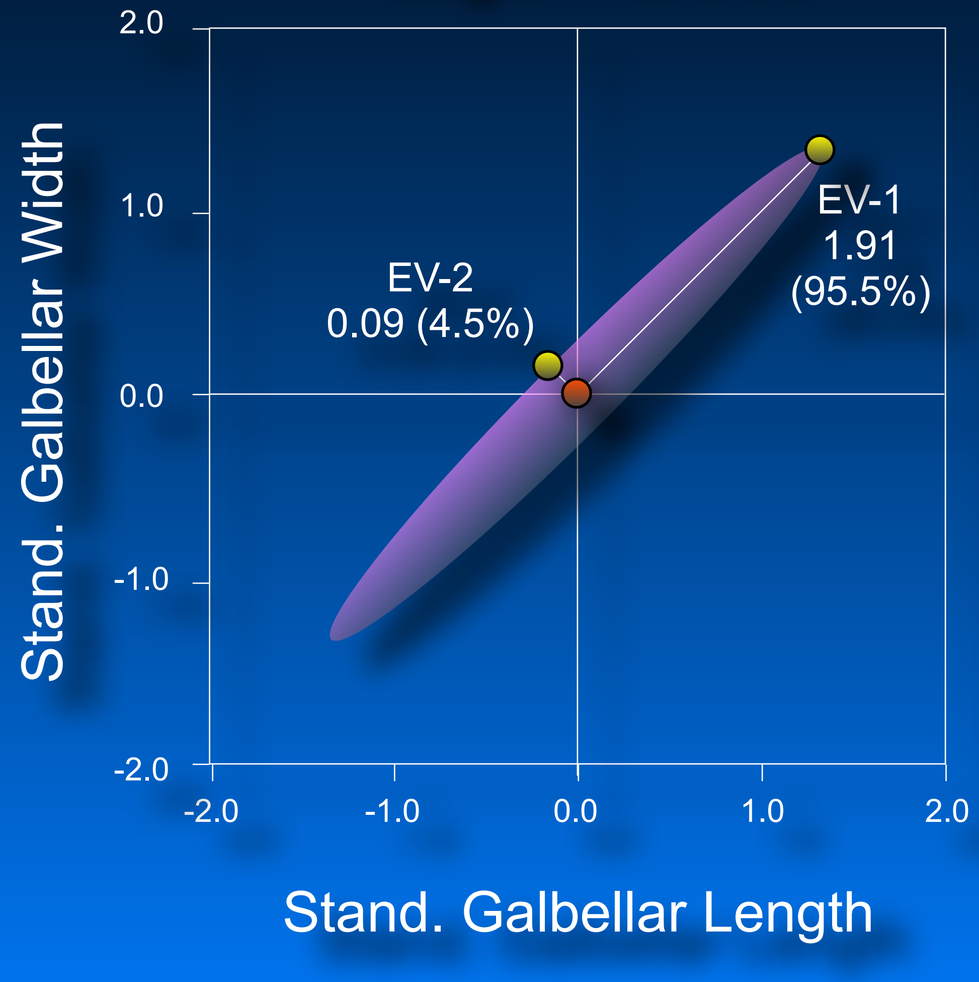
Note projection of PC scores into the geometric space formed by the eigenvector axes does not change the relative positions of the points, only the orientation and dimensions of the point cloud as a whole.

Principal Components Analysis

Correlation Structure

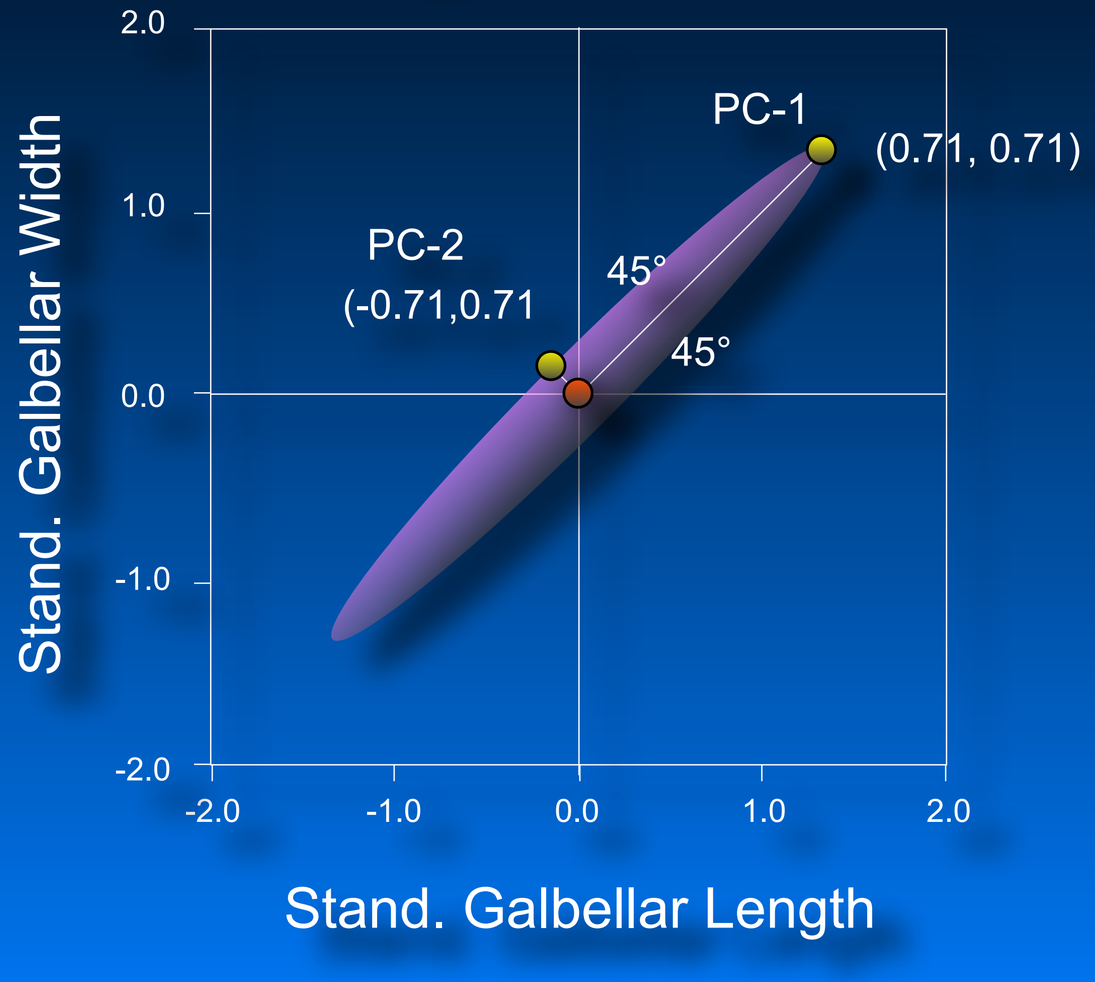


Eigenvalues



Principal Components Analysis

Eigenvectors



Correlation Matrix

$$PC_1 = 0.71x_1 + 0.71x_2$$
$$PC_2 = -0.71x_1 + 0.71x_2$$

Trilobite Glabella Dataset

Equations of PCA Axes

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78

Covariance Matrix

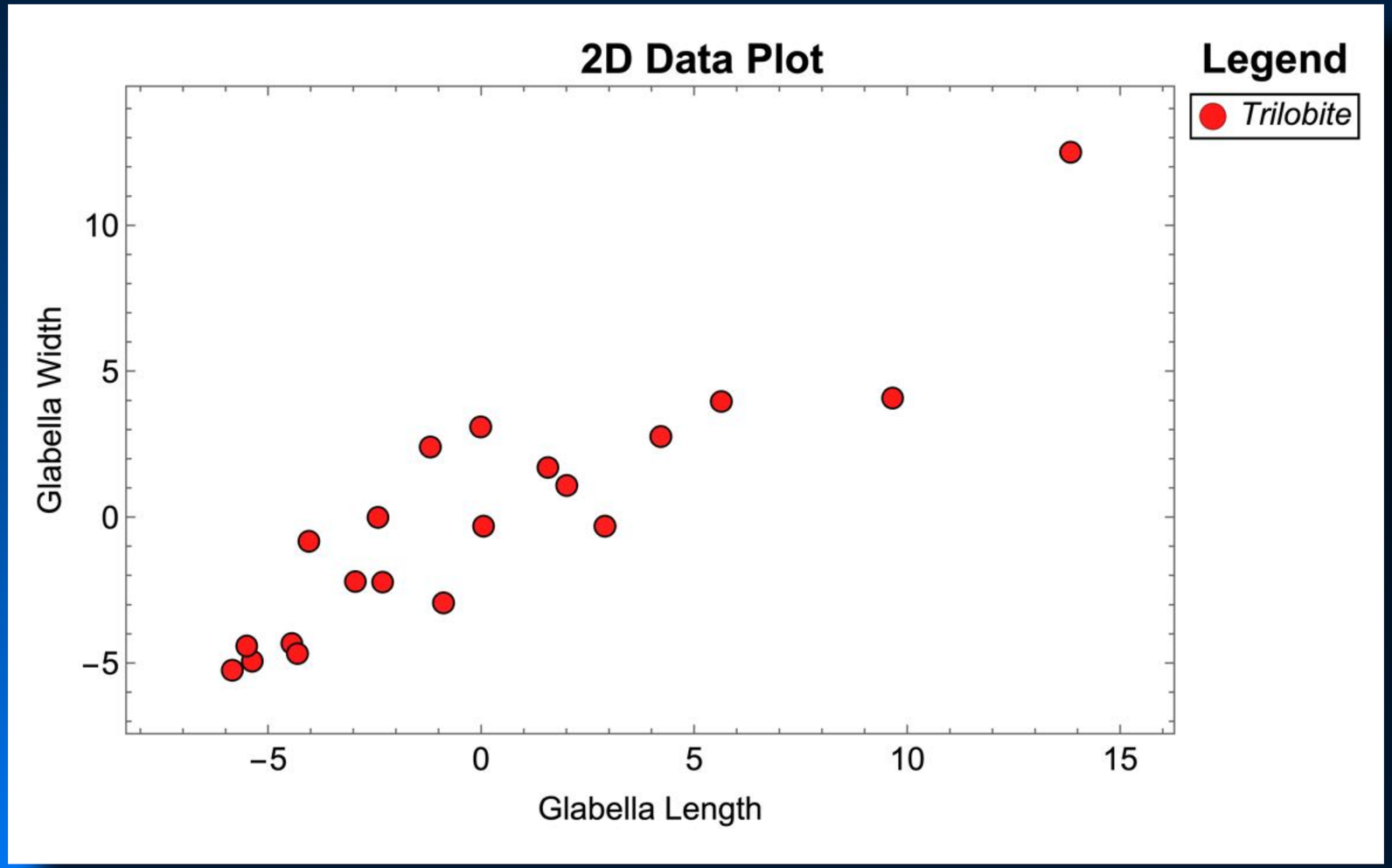
$$PC_1 = 0.78x_1 + 0.63x_2$$
$$PC_2 = -0.63x_1 + 0.78x_2$$

Correlation Matrix

$$PC_1 = 0.71x_1 + 0.71x_2$$
$$PC_2 = -0.71x_1 + 0.71x_2$$

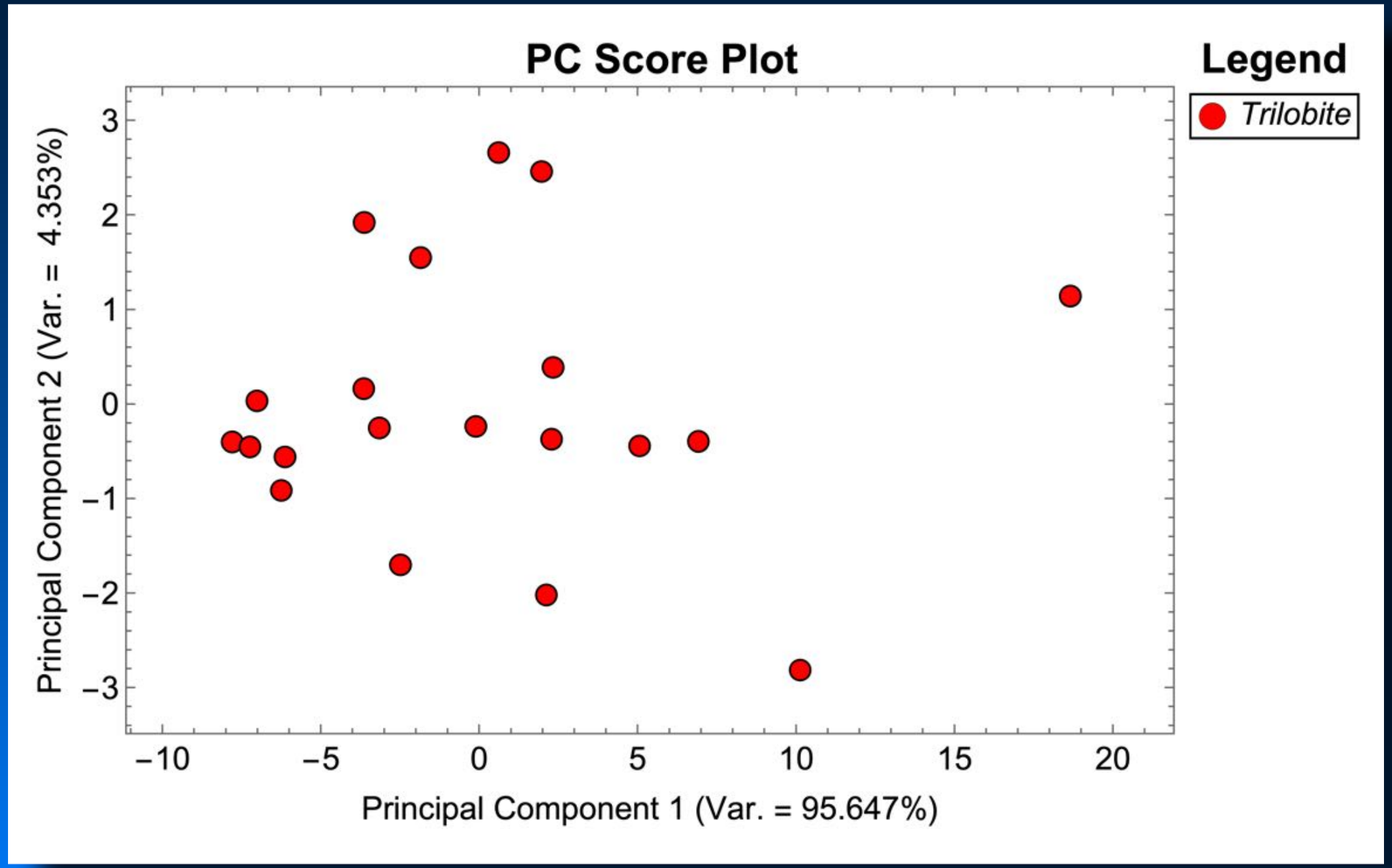
Principal Components Analysis

Original Variable Space



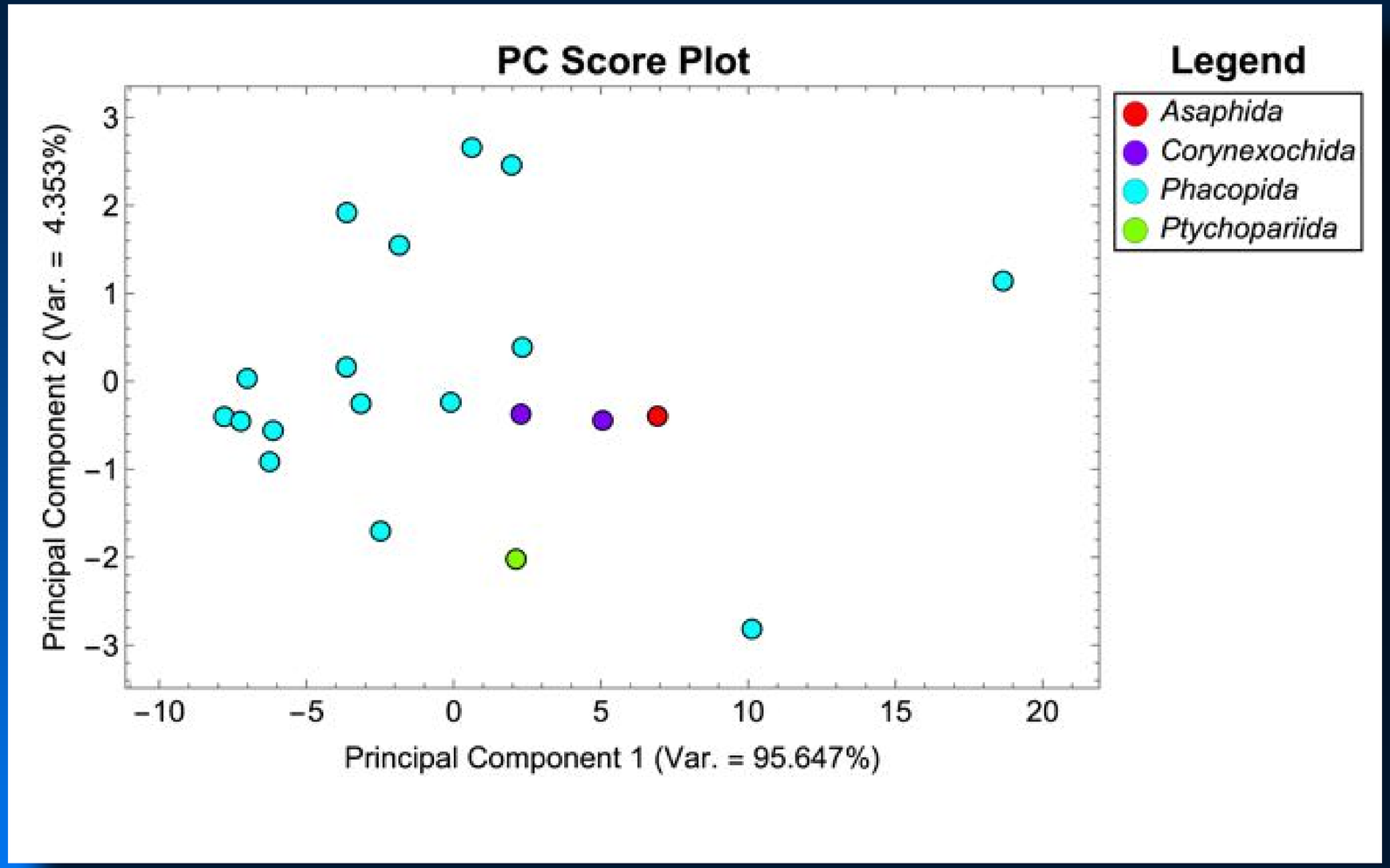
Principal Components Analysis

PCA Ordination Space



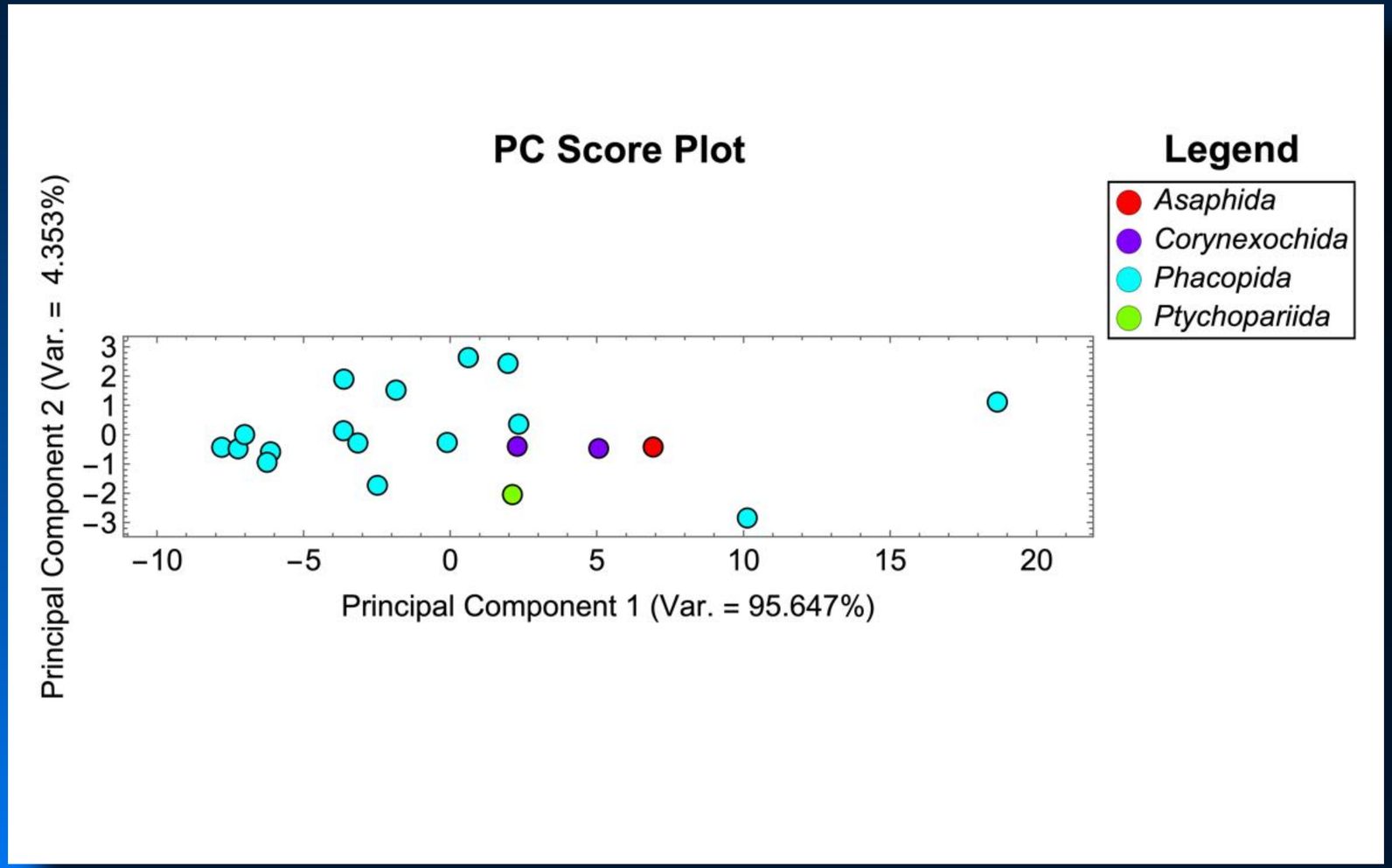
Principal Components Analysis

PCA Ordination Space with Groups



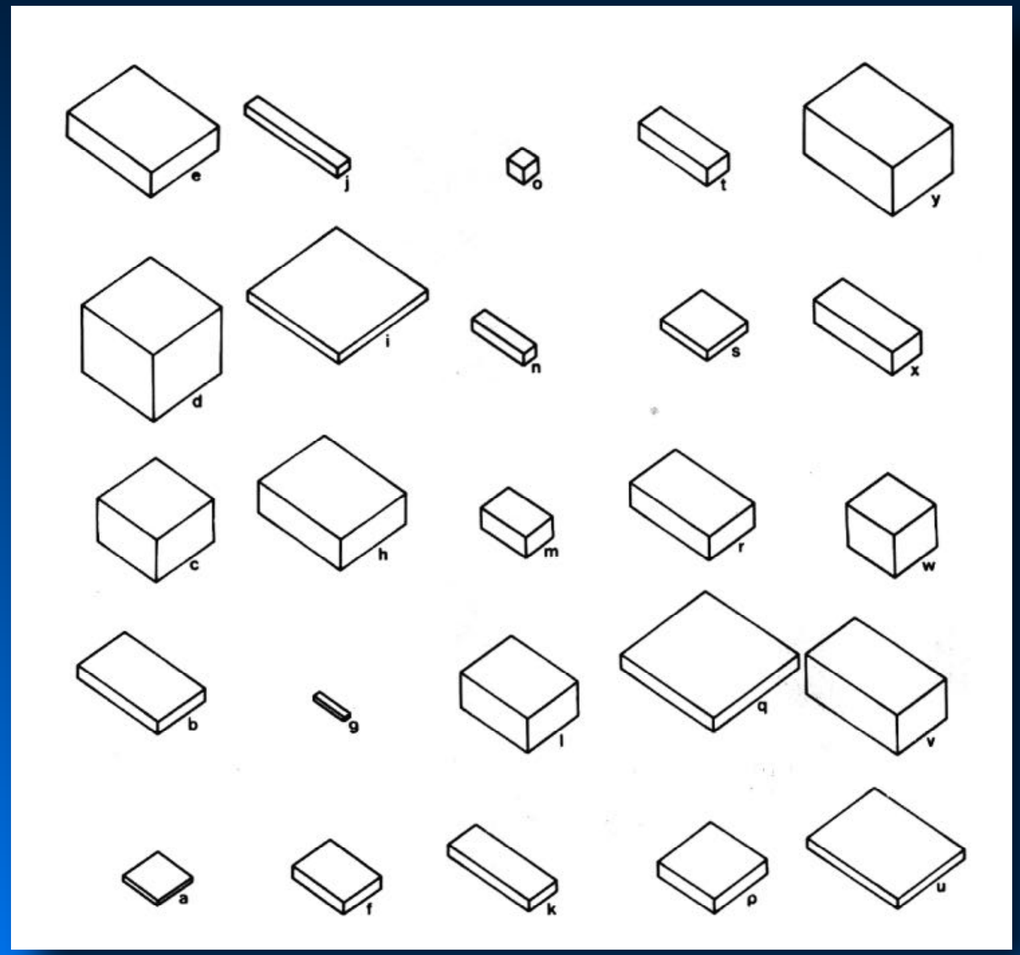
Principal Components Analysis

PCA Ordination Space with Groups & True-Scale Axes



Principal Components Analysis

A More Realistic Example: The Davis (2002) Boxes Dataset



X_1 = long axis

X_2 = intermediate axis

X_3 = short axis

X_4 = longest diagonal

$$X_5 = \text{ratio}\ \frac{\text{radius of smallest circumscribed sphere}}{\text{radius of largest inscribed sphere}}$$

$$X_6 = \text{ratio}\ \frac{\text{long axis} + \text{intermediate axis}}{\text{short axis}}$$

$$X_7 = \text{ratio}\ \frac{\text{surface area}}{\text{volume}}$$

Principal Components Analysis

A More Realistic Example: The Davis (2002) Boxes Dataset

Boxes

Object	Long Axis	Inter. Axis	Short Axis	Long. Diagonal	Large/Small Spheres	Long+Inter/ Short Axes	Area/ Volume
a	3.760	3.660	0.540	5.275	9.768	13.741	4.782
b	8.590	4.990	1.340	10.022	7.500	10.162	2.130
c	6.220	6.140	4.520	9.842	2.175	2.732	1.089
d	7.570	7.280	7.070	12.662	1.791	2.101	0.822
e	9.030	7.080	2.590	11.762	4.539	6.217	1.276
f	5.510	3.980	1.300	6.924	5.326	7.304	2.403
g	3.270	0.620	0.440	3.357	7.629	8.838	8.389
h	8.740	7.000	3.310	11.675	3.529	4.757	1.119
i	9.640	9.490	1.030	13.567	13.133	18.519	2.354
j	9.730	1.330	1.000	9.871	9.871	11.064	3.704
k	8.590	2.980	1.170	9.170	7.851	9.908	2.616
l	7.120	5.490	3.680	9.716	2.642	3.430	1.189
m	4.690	3.101	2.170	5.983	2.760	3.554	2.013
n	5.510	1.340	1.270	5.808	4.566	5.382	3.427
o	1.660	1.610	1.570	2.799	1.783	2.087	3.716
p	5.900	5.760	1.550	8.388	5.395	7.497	1.973

Principal Components Analysis

A More Realistic Example: The Davis (2002) Boxes Dataset

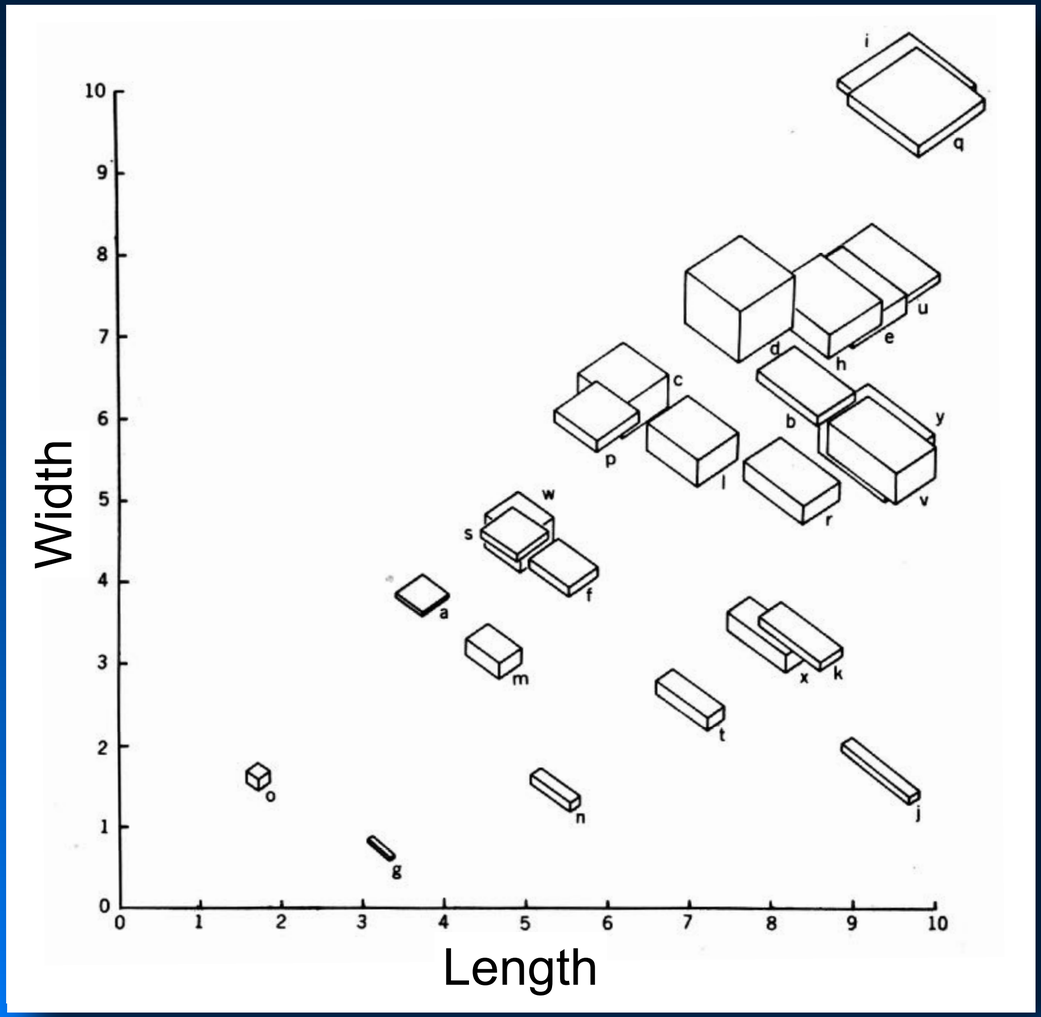


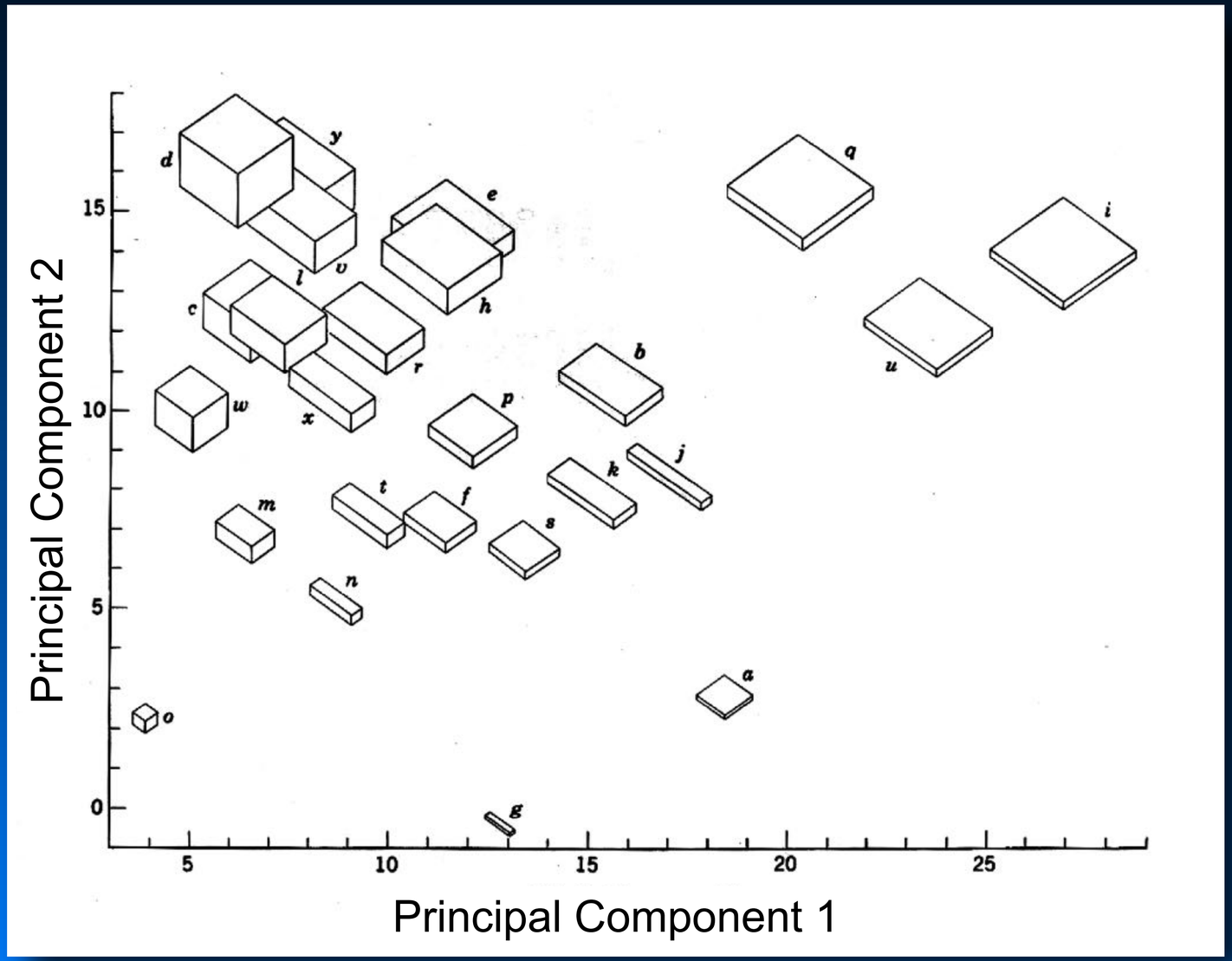
TABLE 6.21 Variance–Covariance Matrix of Seven Variables Measured on 25 Blocks; Only the Lower Half of the Symmetric Matrix is Shown

	Variable						
Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	5.400						
X_2	3.260	5.846					
X_3	0.7785	1.465	2.774				
X_4	6.391	6.083	2.204	9.107			
X_5	2.155	1.312	−3.839	1.610	10.710		
X_6	3.035	2.877	−5.167	2.782	14.770	20.780	
X_7	−1.996	−2.370	−1.740	−3.283	2.252	2.622	2.594

Matrix of Eigenvectors

	Eigenvector						
Variable	I	II	III	IV	V	VI	VII
X_1	0.164	0.422	0.645	−0.090	0.225	0.415	−0.385
X_2	0.142	0.447	−0.713	−0.050	0.395	0.066	−0.329
X_3	−0.173	0.257	−0.130	0.629	−0.607	0.280	−0.211
X_4	0.170	0.650	0.146	0.212	0.033	−0.403	0.565
X_5	0.546	−0.135	0.105	0.165	−0.161	−0.596	−0.513
X_6	0.768	−0.133	−0.149	−0.062	−0.207	0.465	0.327
X_7	0.073	−0.313	0.065	0.719	0.596	0.107	0.092
Eigenvalues	34.490	19.000	2.540	0.810	0.340	0.033	0.003
Percentage of Total Variance Contributed by Each Eigenvalue	60.290	33.210	4.440	1.410	0.600	0.060	0.004

Principal Components Analysis



Principal Components Analysis

The Closure Problem

N	Group	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	Example	0.167	0.189	0.030	0.144	0.015	0.136	0.136	0.076	0.023	0.083
2	Example	0.104	0.141	0.007	0.007	0.185	0.044	0.156	0.074	0.104	0.178
3	Example	0.085	0.074	0.117	0.053	0.074	0.053	0.053	0.170	0.074	0.245
4	Example	0.166	0.159	0.097	0.048	0.034	0.124	0.110	0.138	0.028	0.097
5	Example	0.177	0.177	0.031	0.138	0.069	0.062	0.100	0.185	0.023	0.038
6	Example	0.078	0.070	0.164	0.016	0.102	0.133	0.070	0.109	0.125	0.133
7	Example	0.114	0.179	0.064	0.043	0.093	0.086	0.064	0.171	0.157	0.029
8	Example	0.183	0.145	0.061	0.183	0.084	0.183	0.023	0.084	0.031	0.023
9	Example	0.182	0.164	0.173	0.036	0.127	0.055	0.027	0.118	0.100	0.018
10	Example	0.167	0.061	0.083	0.098	0.174	0.167	0.106	0.045	0.061	0.038
11	Example	0.073	0.127	0.127	0.064	0.082	0.118	0.173	0.173	0.045	0.018
12	Example	0.106	0.050	0.050	0.135	0.078	0.163	0.156	0.149	0.092	0.021
13	Example	0.114	0.065	0.154	0.138	0.098	0.195	0.138	0.008	0.049	0.041
14	Example	0.090	0.090	0.174	0.160	0.118	0.042	0.174	0.104	0.035	0.014
15	Example	0.031	0.078	0.078	0.047	0.141	0.109	0.078	0.164	0.133	0.141

Principal Components Analysis

The Closure Problem

N	Group	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	Σ
1	Example	0.167	0.189	0.030	0.144	0.015	0.136	0.136	0.076	0.023	0.083	1.000
2	Example	0.104	0.141	0.007	0.007	0.185	0.044	0.156	0.074	0.104	0.178	
3	Example	0.085	0.074	0.117	0.053	0.074	0.053	0.053	0.170	0.074	0.245	
4	Example	0.166	0.159	0.097	0.048	0.034	0.124	0.110	0.138	0.028	0.097	
5	Example	0.177	0.177	0.031	0.138	0.069	0.062	0.100	0.185	0.023	0.038	
6	Example	0.078	0.070	0.164	0.016	0.102	0.133	0.070	0.109	0.125	0.133	
7	Example	0.114	0.179	0.064	0.043	0.093	0.086	0.064	0.171	0.157	0.029	
8	Example	0.183	0.145	0.061	0.183	0.084	0.183	0.023	0.084	0.031	0.023	
9	Example	0.182	0.164	0.173	0.036	0.127	0.055	0.027	0.118	0.100	0.018	
10	Example	0.167	0.061	0.083	0.098	0.174	0.167	0.106	0.045	0.061	0.038	
11	Example	0.073	0.127	0.127	0.064	0.082	0.118	0.173	0.173	0.045	0.018	
12	Example	0.106	0.050	0.050	0.135	0.078	0.163	0.156	0.149	0.092	0.021	
13	Example	0.114	0.065	0.154	0.138	0.098	0.195	0.138	0.008	0.049	0.041	
14	Example	0.090	0.090	0.174	0.160	0.118	0.042	0.174	0.104	0.035	0.014	
15	Example	0.031	0.078	0.078	0.047	0.141	0.109	0.078	0.164	0.133	0.141	

Principal Components Analysis

The Closure Problem

N	Group	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	Σ
1	Example	0.167	0.189	0.030	0.144	0.015	0.136	0.136	0.076	0.023	0.083	1.000
2	Example	0.104	0.141	0.007	0.007	0.185	0.044	0.156	0.074	0.104	0.178	1.000
3	Example	0.085	0.074	0.117	0.053	0.074	0.053	0.053	0.170	0.074	0.245	1.000
4	Example	0.166	0.159	0.097	0.048	0.034	0.124	0.110	0.138	0.028	0.097	1.000
5	Example	0.177	0.177	0.031	0.138	0.069	0.062	0.100	0.185	0.023	0.038	1.000
6	Example	0.078	0.070	0.164	0.016	0.102	0.133	0.070	0.109	0.125	0.133	1.000
7	Example	0.114	0.179	0.064	0.043	0.093	0.086	0.064	0.171	0.157	0.029	1.000
8	Example	0.183	0.145	0.061	0.183	0.084	0.183	0.023	0.084	0.031	0.023	1.000
9	Example	0.182	0.164	0.173	0.036	0.127	0.055	0.027	0.118	0.100	0.018	1.000
10	Example	0.167	0.061	0.083	0.098	0.174	0.167	0.106	0.045	0.061	0.038	1.000
11	Example	0.073	0.127	0.127	0.064	0.082	0.118	0.173	0.173	0.045	0.018	1.000
12	Example	0.106	0.050	0.050	0.135	0.078	0.163	0.156	0.149	0.092	0.021	1.000
13	Example	0.114	0.065	0.154	0.138	0.098	0.195	0.138	0.008	0.049	0.041	1.000
14	Example	0.090	0.090	0.174	0.160	0.118	0.042	0.174	0.104	0.035	0.014	1.000
15	Example	0.031	0.078	0.078	0.047	0.141	0.109	0.078	0.164	0.133	0.141	1.000

Principal Components Analysis

The Closure Problem

The problem with closed datasets is that such data do not represent the relative magnitude of some variables relative to others. Instead, they represent variable parts of those magnitudes that have been scaled (inconsistently) by the variable sums of the cross-variable row magnitudes. Ideally, we would like to remove that effect.

Other (related) issues include:

- Imprecision in estimation of variances.
- Imprecision in estimation of covariances and/or correlations.

Principal Components Analysis

The Closure Problem

Eigenvalues

Component	Eigenvalue	Variance (%)	Cum. Variance (%)
1	0.009	31.194	31.194
2	0.005	17.747	48.940
3	0.004	12.906	61.846
4	0.003	11.194	73.040
5	0.003	9.650	82.690
6	0.002	7.596	90.286
7	0.002	6.621	96.907
8	0.001	2.116	99.023
9	0.000	0.977	100.000
10	0.000	0.000	100.000

Principal Components Analysis

The Closure Problem

Eigenvalues

Component	Eigenvalue	Variance (%)	Cum. Variance (%)
1	0.009	31.194	31.194
2	0.005	17.747	48.940
3	0.004	12.906	61.846
4	0.003	11.194	73.040
5	0.003	9.650	82.690
6	0.002	7.596	90.286
7	0.002	6.621	96.907
8	0.001	2.116	99.023
9	0.000	0.977	100.000
10	0.000	0.000	100.000

Principal Components Analysis

The Closure Problem

There is no way to remove the effect of conversion to proportional data completely. However, strategies are available that can be employed to minimize its effect.

Centered Log-Ratio Transform

$$Cov_{j,k} = Cov\left\{ ln\frac{x_j}{g_j}, ln\frac{x_k}{g_k} \right\}$$

$$g_i = (\prod_{i=1}^{m} x_i)^{\frac{1}{m}}$$

Principal Components Analysis

The Closure Problem

N	Group	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	Example	0.800	0.928	-0.905	0.653	-1.598	0.599	0.599	0.011	-1.193	0.107
2	Example	0.463	0.769	-2.176	-2.176	1.043	-0.384	0.869	0.127	0.463	1.002
3	Example	-0.026	-0.159	0.293	-0.496	-0.159	-0.496	-0.496	0.667	-0.159	1.030
4	Example	0.659	0.616	0.120	-0.573	-0.910	0.371	0.253	0.477	-1.133	0.120
5	Example	0.808	0.808	-0.941	0.563	-0.130	-0.248	0.237	0.851	-1.229	-0.718
6	Example	-0.105	-0.211	0.637	-1.715	0.157	0.425	-0.211	0.231	0.365	0.425
7	Example	0.284	0.731	-0.291	-0.697	0.077	-0.003	-0.291	0.690	0.603	-1.102
8	Example	0.884	0.651	-0.214	0.884	0.104	0.884	-1.195	0.104	-0.907	-1.195
9	Example	0.856	0.750	0.804	-0.754	0.499	-0.348	-1.042	0.425	0.258	-1.447
10	Example	0.643	-0.369	-0.050	0.117	0.688	0.643	0.191	-0.656	-0.369	-0.839
11	Example	-0.146	0.413	0.413	-0.280	-0.028	0.339	0.719	0.719	-0.616	-1.533
12	Example	0.220	-0.542	-0.542	0.457	-0.090	0.648	0.603	0.557	0.077	-1.389
13	Example	0.400	-0.160	0.705	0.594	0.246	0.939	0.594	-2.239	-0.448	-0.630
14	Example	0.131	0.131	0.785	0.702	0.399	-0.642	0.785	0.274	-0.824	-1.741
15	Example	-1.052	-0.136	-0.136	-0.647	0.452	0.201	-0.136	0.606	0.395	0.452

Principal Components Analysis

The Closure Problem

N	Group	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	Σ
1	Example	0.800	0.928	-0.905	0.653	-1.598	0.599	0.599	0.011	-1.193	0.107	0.000
2	Example	0.463	0.769	-2.176	-2.176	1.043	-0.384	0.869	0.127	0.463	1.002	0.000
3	Example	-0.026	-0.159	0.293	-0.496	-0.159	-0.496	-0.496	0.667	-0.159	1.030	0.000
4	Example	0.659	0.616	0.120	-0.573	-0.910	0.371	0.253	0.477	-1.133	0.120	0.000
5	Example	0.808	0.808	-0.941	0.563	-0.130	-0.248	0.237	0.851	-1.229	-0.718	0.000
6	Example	-0.105	-0.211	0.637	-1.715	0.157	0.425	-0.211	0.231	0.365	0.425	0.000
7	Example	0.284	0.731	-0.291	-0.697	0.077	-0.003	-0.291	0.690	0.603	-1.102	0.000
8	Example	0.884	0.651	-0.214	0.884	0.104	0.884	-1.195	0.104	-0.907	-1.195	0.000
9	Example	0.856	0.750	0.804	-0.754	0.499	-0.348	-1.042	0.425	0.258	-1.447	0.000
10	Example	0.643	-0.369	-0.050	0.117	0.688	0.643	0.191	-0.656	-0.369	-0.839	0.000
11	Example	-0.146	0.413	0.413	-0.280	-0.028	0.339	0.719	0.719	-0.616	-1.533	0.000
12	Example	0.220	-0.542	-0.542	0.457	-0.090	0.648	0.603	0.557	0.077	-1.389	0.000
13	Example	0.400	-0.160	0.705	0.594	0.246	0.939	0.594	-2.239	-0.448	-0.630	0.000
14	Example	0.131	0.131	0.785	0.702	0.399	-0.642	0.785	0.274	-0.824	-1.741	0.000
15	Example	-1.052	-0.136	-0.136	-0.647	0.452	0.201	-0.136	0.606	0.395	0.452	0.000

Principal Coordinates Analysis

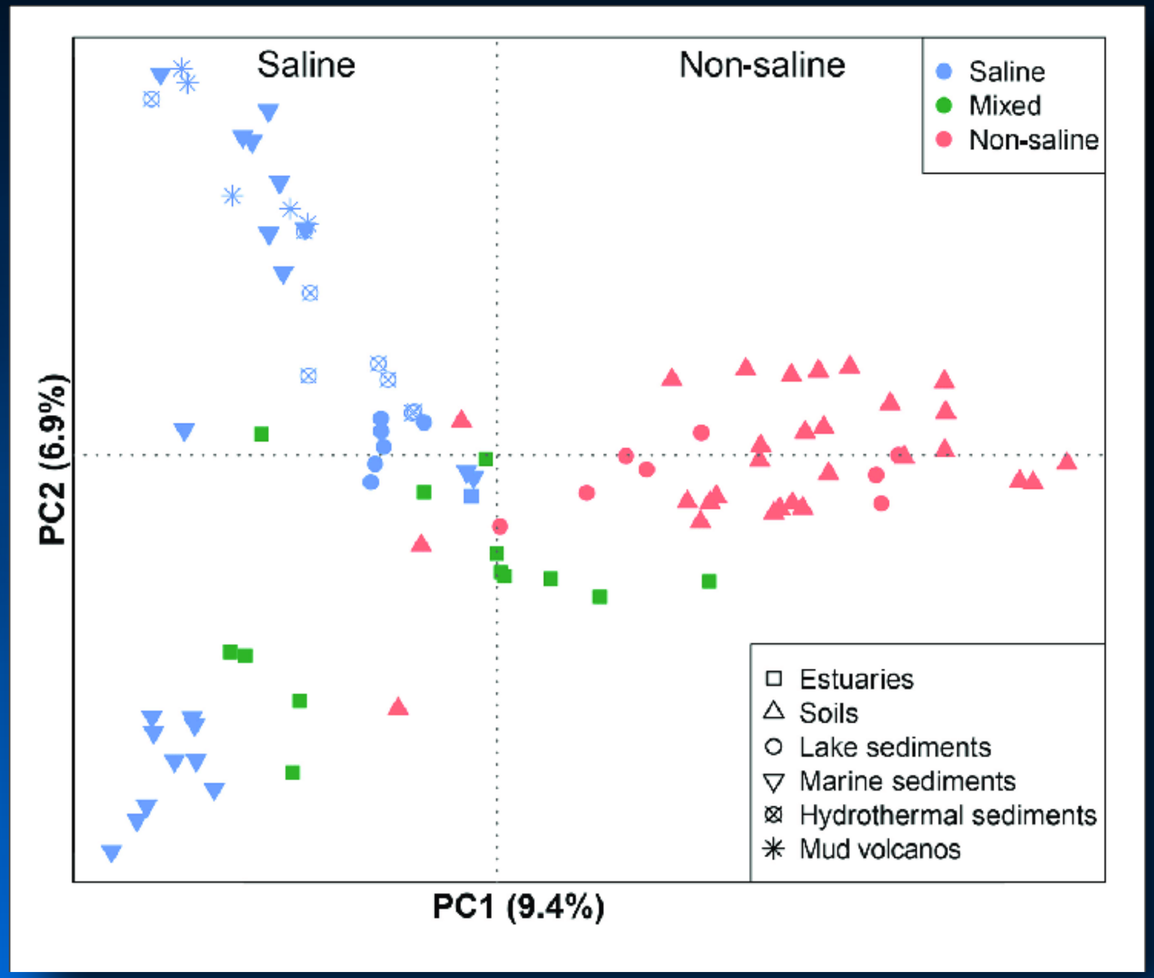
Trilobite Morphology

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78

Principal Coordinates Analysis

An orthogonal data transformation that projects a set of raw data values onto a set of linearly uncorrelated composite variables (the principal coordinates), the first of which is aligned with the direction of maximum linear distance or similarity variation. Subsequent coordinate axes are ordered such that each is aligned to the directions of maximum residual variation subject to the constraint of orthogonality.

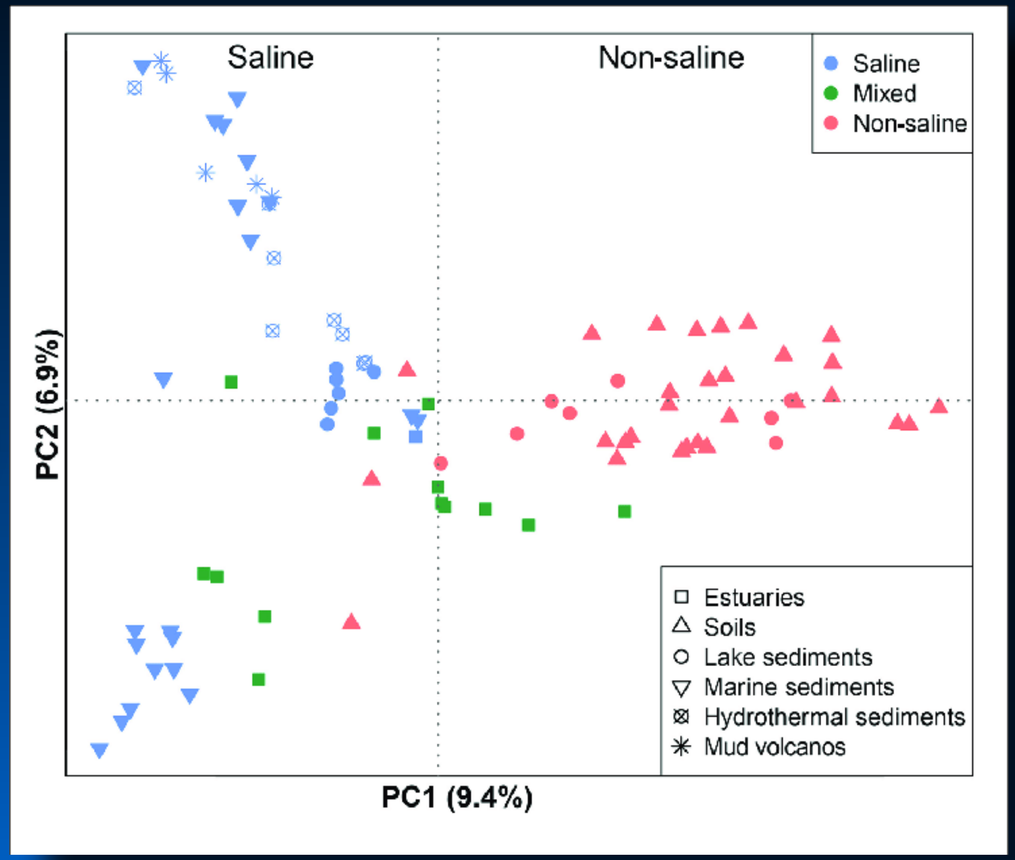
Principal coordinates analysis (PCoord or PCoA) is the eigenanalysis of object or sample similarity-dissimilarity matrices.



Principal Coordinates Analysis

Characteristics

- A data transformation (not a statistical procedure)
- Principal coordinates = eigenvector de-composition of a square, non-singular matrix that expresses distance or similarity relations between observations or objects.
- Assumes all observations comprise a single group (makes no between-group distinctions or optimizations).
- Involves no distributional assumptions.
- Can be used to reduce data dimensionality, with minimal loss of information.
- Under certain settings can be regarded as the duel of principal components analysis.



Principal Coordinates Analysis

r-Mode and Q-Mode Orientations and Duality

r-Mode Analyses - Data analyses that seek to focus on and, to the extent allowed by the procedure, preserve the variance structure of the variables measured or collected from a sample of individuals. The term comes from the mathematical symbol for the correlation matrix ($r=X'X$). Example: principal components analysis (PCA).

Q-Mode Analyses - Data analyses that seek to focus on and, to the extent allowed by the procedure, preserve the information content and or distance/similarity structure of the observations made, measured or collected from a sample of individuals. The term comes from the mathematical symbol (q) for the cosine θ similarity matrix ($q=XX'$). Example: principal coordinates analysis (PCoord).

r-Mode Basis Matrices

These matrices quantify and scale relations between variables.

Genus	Glabella Length	Glabella Width
Acaste	3.50	3.77
Balizoma	3.97	4.08
Calymene	10.91	10.72
Ceraurus	4.90	4.69
Cheirurus	9.33	12.11
Cybantyx	11.35	10.10
Cybeloides	6.39	6.81
Dalmanites	8.46	6.08
Deiphon	6.92	9.01
Ormathops	5.03	4.34
Phacopidina	7.03	6.79
Phacops	5.30	8.19
Placoparia	9.40	8.71
Pricyclopyge	14.98	12.98
Ptychoparia	12.25	8.71
Rhenops	19.00	13.10
Sphaerexochus	3.84	4.60
Toxochasmops	8.15	11.42
Trimerus	23.18	21.52
Zacanthoides	13.56	11.78

Covariance Matrix

	Glabella Length	Glabella Width
Glabella Length	27.33	20.32
Glabella Width	20.32	19.27

Correlation Matrix

	Glabella Length	Glabella Width
Glabella Length	1.00	0.91
Glabella Width	0.91	1.00

Q-Mode Basis (Distance) Matrices

These indices are used quantify and scale relations between objects.

Euclidean Distance: $$d_{jk} = \sum_{i=0}^{n} \sqrt{(x_{ij} - x_{ik})^2}$$

Squared Euclidean Distance: $$d_{jk}^2 = \sum_{i=0}^{n} (x_{ij} - x_{ik})^2$$

Squared Mahalanobis Distance: $$d_{ij}^2 = \left[\bar{X}_i - \bar{X}_j\right]' \cdot \left[s_{ij}^2\right] \cdot \left[\bar{X}_i - \bar{X}_j\right]$$

Mahalanobis Distance: $$d_{ij} = \sqrt{d_{ij}^2}$$

Cosine θ:

$$\theta_{jk} = \frac{\sum_{i=1}^{n} x_{ik} x_{jk}}{\sqrt{\sum_{i=1}^{n} x_{ik}^2 \cdot \sum_{i=1}^{n} x_{jk}^2}}$$

Gower Distance:

$$d_{ij} = \frac{\frac{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|}{Range_k}}{p}$$

Q-Mode Basis (Distance) Matrix

Squared Euclidean Distance Matrix

Genus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0.32	103.21	2.81	103.54	101.69	17.59	29.94	39.15	2.67	21.58	22.78	59.21	216.61	100.97	327.30	0.80	80.15	702.36	165.36
2	0.32	0	92.25	1.24	93.21	90.70	13.31	24.16	33.01	1.19	16.71	18.66	50.92	200.43	90.00	307.26	0.29	71.35	673.18	151.26
3	103.21	92.25	0	72.48	4.43	0.58	35.72	27.53	18.84	75.28	30.50	37.87	6.32	21.67	5.84	71.11	87.44	8.11	267.19	8.15
4	2.81	1.24	72.48	0	74.68	70.87	6.71	14.61	22.74	0.14	8.95	12.41	36.41	170.33	70.18	269.54	1.13	55.86	617.41	125.26
5	103.54	93.21	4.43	74.68	0	8.12	36.73	37.12	15.42	78.86	33.59	31.61	11.56	32.68	20.09	94.49	86.54	1.87	280.37	18.00
6	101.69	90.70	0.58	70.87	8.12	0	35.43	24.51	20.81	73.12	29.62	40.25	5.73	21.47	2.74	67.52	86.65	11.98	270.37	7.71
7	17.59	13.31	35.72	6.71	36.73	35.43	0	4.82	5.12	7.95	0.41	3.09	12.67	111.86	37.95	198.58	11.39	24.35	498.29	76.11
8	29.94	24.16	27.53	14.61	37.12	24.51	4.82	0	10.96	14.79	2.55	14.44	7.80	90.12	21.28	160.37	23.53	28.61	455.07	58.50
9	39.15	33.01	18.84	22.74	15.42	20.81	5.12	10.96	0	25.38	4.94	3.30	6.24	80.72	28.50	162.65	28.93	7.32	420.89	51.76
10	2.67	1.19	75.28	0.14	78.86	73.12	7.95	14.79	25.38	0	10.00	14.90	38.19	173.65	71.23	271.90	1.48	59.86	624.57	128.11
11	21.58	16.71	30.50	8.95	33.59	29.62	0.41	2.55	4.94	10.00	0	4.95	9.30	101.52	30.93	183.10	14.97	22.69	477.80	67.54
12	22.78	18.66	37.87	12.41	31.61	40.25	3.09	14.44	3.30	14.90	4.95	0	17.08	116.65	48.57	211.80	15.02	18.56	497.38	81.12
13	59.21	50.92	6.32	36.41	11.56	5.73	12.67	7.80	6.24	38.19	9.30	17.08	0	49.37	8.12	111.43	47.81	8.91	353.98	26.73
14	216.61	200.43	21.67	170.33	32.68	21.47	111.86	90.12	80.72	173.65	101.52	116.65	49.37	0	25.69	16.17	194.32	49.08	140.17	3.46
15	100.97	90.00	5.84	70.18	20.09	2.74	37.95	21.28	28.50	71.23	30.93	48.57	8.12	25.69	0	64.83	87.62	24.15	283.56	11.14
16	327.30	307.26	71.11	269.54	94.49	67.52	198.58	160.37	162.65	271.90	183.10	211.80	111.43	16.17	64.83	0	302.08	120.54	88.37	31.34
17	0.80	0.29	87.44	1.13	86.54	86.65	11.39	23.53	28.93	1.48	14.97	15.02	47.81	194.32	87.62	302.08	0	65.09	660.32	146.03
18	80.15	71.35	8.11	55.86	1.87	11.98	24.35	28.61	7.32	59.86	22.69	18.56	8.91	49.08	24.15	120.54	65.09	0	327.91	29.40
19	702.36	673.18	267.19	617.41	280.37	270.37	498.29	455.07	420.89	624.57	477.80	497.38	353.98	140.17	283.56	88.37	660.32	327.91	0	187.41
20	165.36	151.26	8.15	125.26	18.00	7.71	76.11	58.50	51.76	128.11	67.54	81.12	26.73	3.46	11.14	31.34	146.03	29.40	187.41	0

Principal Coordinates Analysis

Data Normalization Step

$$a_{ij} = \frac{(\bar{d}_i + \bar{d}_j - \bar{d}_{GM} - d_{ij})}{2}$$

Where: $\bar{d}_i$ = mean of row i;

$\bar{d}_j$ = mean of column j;

$\bar{d}_{GM}$ = grand mean of all rows and columns.

The effect of this operation is to center and “close” the dataset such that the sum of each row and each column is 0.0.

Q-Mode Basis (Distance) Matrix

Normalized Squared Euclidean Distance Matrix

Genus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	61.58	57.21	-18.11	48.57	-16.07	-17.47	28.79	20.43	14.22	49.63	25.13	28.00	1.22	-53.78	-15.52	-78.01	55.27	-5.55	-146.38	-39.19
2	57.21	53.15	-16.85	45.14	-15.12	-16.19	26.71	19.10	13.08	46.15	23.35	25.85	1.15	-49.90	-14.25	-72.20	51.31	-5.36	-136.01	-36.35
3	-18.11	-16.85	5.41	-14.35	5.40	5.00	-8.36	-6.45	-3.71	-14.76	-7.41	-7.63	-0.42	15.61	3.96	22.00	-16.14	2.38	43.11	11.33
4	48.57	45.14	-14.35	38.37	-13.24	-13.66	22.62	16.49	10.82	39.29	19.84	21.58	1.01	-42.24	-11.73	-60.73	43.50	-5.01	-115.51	-30.75
5	-16.07	-15.12	5.40	-13.24	9.83	3.44	-6.66	-9.04	0.21	-14.35	-6.75	-2.29	-0.83	12.31	-0.95	12.52	-13.48	7.71	38.73	8.61
6	-17.47	-16.19	5.00	-13.66	3.44	5.18	-8.33	-5.06	-4.81	-13.80	-7.09	-8.94	-0.24	15.59	5.39	23.68	-15.86	0.33	41.41	11.43
7	28.79	26.71	-8.36	22.62	-6.66	-8.33	13.58	8.99	7.24	22.99	11.72	13.85	0.49	-25.40	-8.01	-37.65	25.98	-1.65	-68.35	-18.56
8	20.43	19.10	-6.45	16.49	-9.04	-5.06	8.99	9.22	2.14	17.38	8.47	5.99	0.74	-16.71	-1.86	-20.73	17.72	-5.96	-48.92	-11.94
9	14.22	13.08	-3.71	10.82	0.21	-4.81	7.24	2.14	6.02	10.49	5.67	9.96	-0.08	-13.61	-7.07	-23.47	13.42	3.08	-33.43	-10.17
10	49.63	46.15	-14.76	39.29	-14.35	-13.80	22.99	17.38	10.49	40.35	20.30	21.33	1.11	-42.91	-11.26	-60.93	44.31	-6.02	-118.11	-31.18
11	25.13	23.35	-7.41	19.84	-6.75	-7.09	11.72	8.47	5.67	20.30	10.26	11.26	0.52	-21.89	-6.16	-31.57	22.52	-2.48	-59.76	-15.94
12	28.00	25.85	-7.63	21.58	-2.29	-8.94	13.85	5.99	9.96	21.33	11.26	17.20	0.10	-25.98	-11.51	-42.45	25.97	3.06	-66.08	-19.26
13	1.22	1.15	-0.42	1.01	-0.83	-0.24	0.49	0.74	-0.08	1.11	0.52	0.10	0.07	-0.91	0.15	-0.83	1.01	-0.68	-2.95	-0.63
14	-53.78	-49.90	15.61	-42.24	12.31	15.59	-25.40	-16.71	-13.61	-42.91	-21.89	-25.98	-0.91	47.48	15.07	70.50	-48.55	2.93	127.66	34.71
15	-15.52	-14.25	3.96	-11.73	-0.95	5.39	-8.01	-1.86	-7.07	-11.26	-6.16	-11.51	0.15	15.07	8.35	26.61	-14.76	-4.17	36.40	11.30
16	-78.01	-72.20	22.00	-60.73	12.52	23.68	-37.65	-20.73	-23.47	-60.93	-31.57	-42.45	-0.83	70.50	26.61	109.70	-71.31	-1.69	184.67	51.88
17	55.27	51.31	-16.14	43.50	-13.48	-15.86	25.98	17.72	13.42	44.31	22.52	25.97	1.01	-48.55	-14.76	-71.31	49.75	-3.93	-131.28	-35.44
18	-5.55	-5.36	2.38	-5.01	7.71	0.33	-1.65	-5.96	3.08	-6.02	-2.48	3.06	-0.68	2.93	-4.17	-1.69	-3.93	7.47	13.79	1.74
19	-146.38	-136.01	43.11	-115.51	38.73	41.41	-68.35	-48.92	-33.43	-118.11	-59.76	-66.08	-2.95	127.66	36.40	184.67	-131.28	13.79	348.01	93.00
20	-39.19	-36.35	11.33	-30.75	8.61	11.43	-18.56	-11.94	-10.17	-31.18	-15.94	-19.26	-0.63	34.71	11.30	51.88	-35.44	1.74	93.00	25.40

Q-Mode Basis (Distance) Matrix

Normalized Squared Euclidean Distance Matrix

Genus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	**61.58**	57.21	-18.11	48.57	-16.07	-17.47	28.79	20.43	14.22	49.63	25.13	28.00	1.22	-53.78	-15.52	-78.01	55.27	-5.55	-146.38	-39.19
2	57.21	**53.15**	-16.85	45.14	-15.12	-16.19	26.71	19.10	13.08	46.15	23.35	25.85	1.15	-49.90	-14.25	-72.20	51.31	-5.36	-136.01	-36.35
3	-18.11	-16.85	**5.41**	-14.35	5.40	5.00	-8.36	-6.45	-3.71	-14.76	-7.41	-7.63	-0.42	15.61	3.96	22.00	-16.14	2.38	43.11	11.33
4	48.57	45.14	-14.35	**38.37**	-13.24	-13.66	22.62	16.49	10.82	39.29	19.84	21.58	1.01	-42.24	-11.73	-60.73	43.50	-5.01	-115.51	-30.75
5	-16.07	-15.12	5.40	-13.24	**9.83**	3.44	-6.66	-9.04	0.21	-14.35	-6.75	-2.29	-0.83	12.31	-0.95	12.52	-13.48	7.71	38.73	8.61
6	-17.47	-16.19	5.00	-13.66	3.44	**5.18**	-8.33	-5.06	-4.81	-13.80	-7.09	-8.94	-0.24	15.59	5.39	23.68	-15.86	0.33	41.41	11.43
7	28.79	26.71	-8.36	22.62	-6.66	-8.33	**13.58**	8.99	7.24	22.99	11.72	13.85	0.49	-25.40	-8.01	-37.65	25.98	-1.65	-68.35	-18.56
8	20.43	19.10	-6.45	16.49	-9.04	-5.06	8.99	**9.22**	2.14	17.38	8.47	5.99	0.74	-16.71	-1.86	-20.73	17.72	-5.96	-48.92	-11.94
9	14.22	13.08	-3.71	10.82	0.21	-4.81	7.24	2.14	**6.02**	10.49	5.67	9.96	-0.08	-13.61	-7.07	-23.47	13.42	3.08	-33.43	-10.17
10	49.63	46.15	-14.76	39.29	-14.35	-13.80	22.99	17.38	10.49	**40.35**	20.30	21.33	1.11	-42.91	-11.26	-60.93	44.31	-6.02	-118.11	-31.18
11	25.13	23.35	-7.41	19.84	-6.75	-7.09	11.72	8.47	5.67	20.30	**10.26**	11.26	0.52	-21.89	-6.16	-31.57	22.52	-2.48	-59.76	-15.94
12	28.00	25.85	-7.63	21.58	-2.29	-8.94	13.85	5.99	9.96	21.33	11.26	**17.20**	0.10	-25.98	-11.51	-42.45	25.97	3.06	-66.08	-19.26
13	1.22	1.15	-0.42	1.01	-0.83	-0.24	0.49	0.74	-0.08	1.11	0.52	0.10	**0.07**	-0.91	0.15	-0.83	1.01	-0.68	-2.95	-0.63
14	-53.78	-49.90	15.61	-42.24	12.31	15.59	-25.40	-16.71	-13.61	-42.91	-21.89	-25.98	-0.91	**47.48**	15.07	70.50	-48.55	2.93	127.66	34.71
15	-15.52	-14.25	3.96	-11.73	-0.95	5.39	-8.01	-1.86	-7.07	-11.26	-6.16	-11.51	0.15	15.07	**8.35**	26.61	-14.76	-4.17	36.40	11.30
16	-78.01	-72.20	22.00	-60.73	12.52	23.68	-37.65	-20.73	-23.47	-60.93	-31.57	-42.45	-0.83	70.50	26.61	**109.70**	-71.31	-1.69	184.67	51.88
17	55.27	51.31	-16.14	43.50	-13.48	-15.86	25.98	17.72	13.42	44.31	22.52	25.97	1.01	-48.55	-14.76	-71.31	**49.75**	-3.93	-131.28	-35.44
18	-5.55	-5.36	2.38	-5.01	7.71	0.33	-1.65	-5.96	3.08	-6.02	-2.48	3.06	-0.68	2.93	-4.17	-1.69	-3.93	**7.47**	13.79	1.74
19	-146.38	-136.01	43.11	-115.51	38.73	41.41	-68.35	-48.92	-33.43	-118.11	-59.76	-66.08	-2.95	127.66	36.40	184.67	-131.28	13.79	**348.01**	93.00
20	-39.19	-36.35	11.33	-30.75	8.61	11.43	-18.56	-11.94	-10.17	-31.18	-15.94	-19.26	-0.63	34.71	11.30	51.88	-35.44	1.74	93.00	**25.40**

Principal Coordinates Analysis

Perform Eigenanalysis of Matrix A

Eigenvalues - the number of positive eigenvalues will often be less than the number of variables and/or number of objects/samples (whichever is smaller). This is because of the steps taken to close the matrix's rows and columns. All eigenvectors associated with zero eigenvalues can be ignored.

All non-zero eigenvectors can be collected into a matrix (V) and used to project the raw data into the principal coordinate space.

$$S = V \cdot \Lambda^{0.5}$$

Where: S = coordinates of raw data projected into PCoord space;

V = matrix of eigenvectors;

Λ = matrix of eigenvalues.

Principal Coordinates Analysis

Principal Coordinates Eigensystem

Eigenvectors	1	2
Acaste	-0.2723	-0.0634
Balizoma	-0.2528	-0.0719
Calymene	0.0796	0.0651
Ceraurus	-0.2143	-0.0891
Cheirurus	0.0669	0.4026
Cybantyx	0.0780	-0.0585
Cybeloides	-0.1279	0.0286
Dalmanites	-0.0876	-0.2750
Deiphon	-0.0657	0.2541
Ormathops	-0.2184	-0.1468
Phacopidina	-0.1110	-0.0392
Phacops	-0.1275	0.3149
Placoparia	-0.0050	-0.0365
Pricyclopyge	0.2390	-0.0622
Ptychoparia	0.0722	-0.3268
Rhenops	0.3506	-0.4563
Sphaerexochus	-0.2450	0.0074
Toxochasmops	0.0200	0.4351
Trimerus	0.6468	0.1879
Zacanthoides	0.1744	-0.0701

Eigenvector	1	2
Eigenvalue	828.67	37.71
Percent Variance	95.65	4.35
Cum. Percent Variance	95.65	100.00

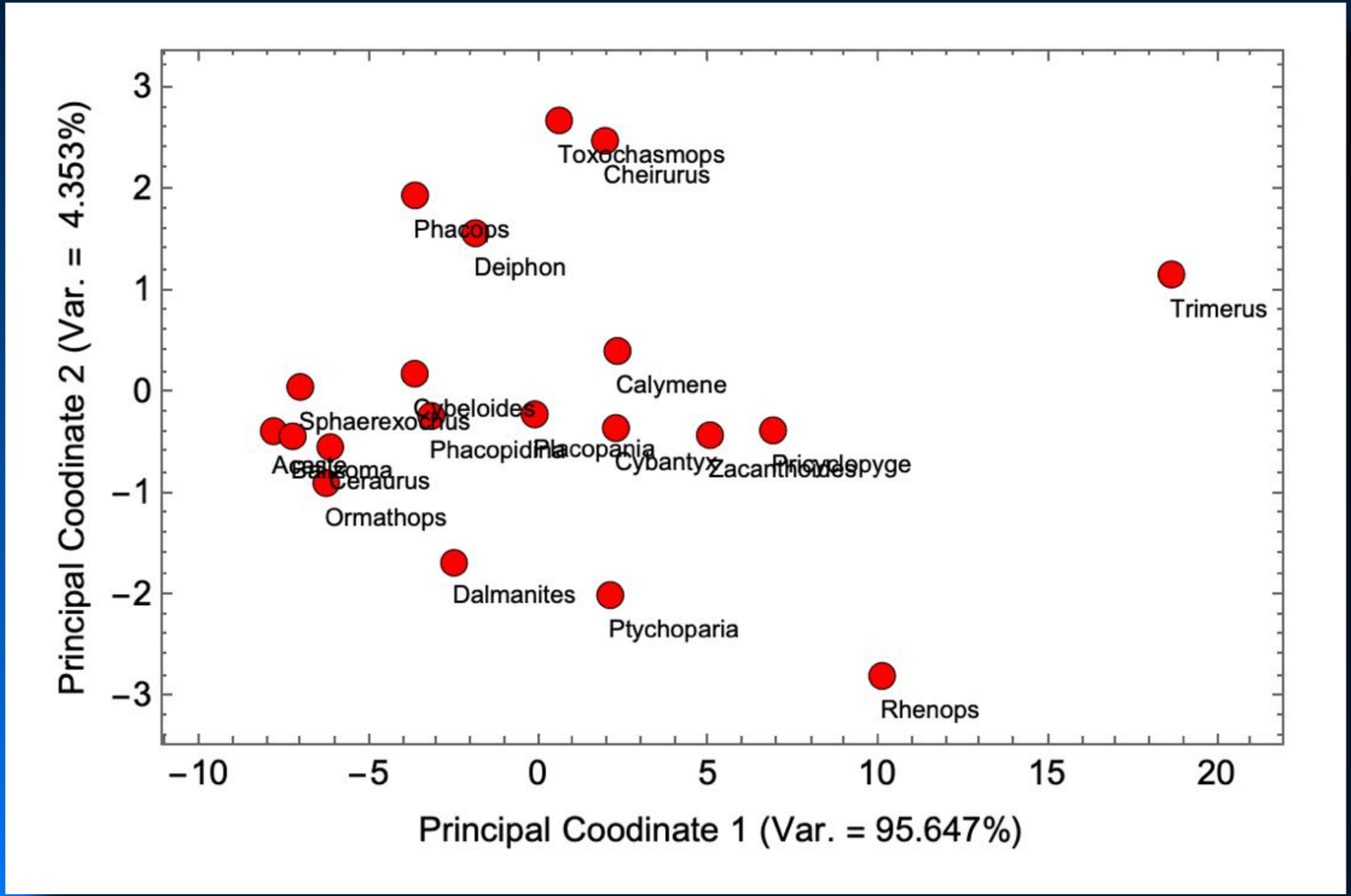
Principal Coordinates Analysis

Principal Coordinate Scores

Eigenvectors	PCoord-1	PCoord-2
Acaste	-7.8378	-0.3894
Balizoma	-7.2772	-0.4414
Calymene	2.2907	0.3997
Ceraurus	-6.1700	-0.5470
Cheirurus	1.9271	2.4725
Cybantyx	2.2463	-0.3592
Cybeloides	-3.6816	0.1754
Dalmanites	-2.5228	-1.6888
Deiphon	-1.8922	1.5607
Ormathops	-6.2875	-0.9015
Phacopidina	-3.1947	-0.2404
Phacops	-3.6691	1.9339
Placoparia	-0.1446	-0.2244
Pricyclopyge	6.8800	-0.3820
Ptychoparia	2.0793	-2.0067
Rhenops	10.0919	-2.8024
Sphaerexochus	-7.0535	0.0456
Toxochasmops	0.5748	2.6720
Trimerus	18.6194	1.1538
Zacanthoides	5.0215	-0.4304

Principal Coordinates Analysis

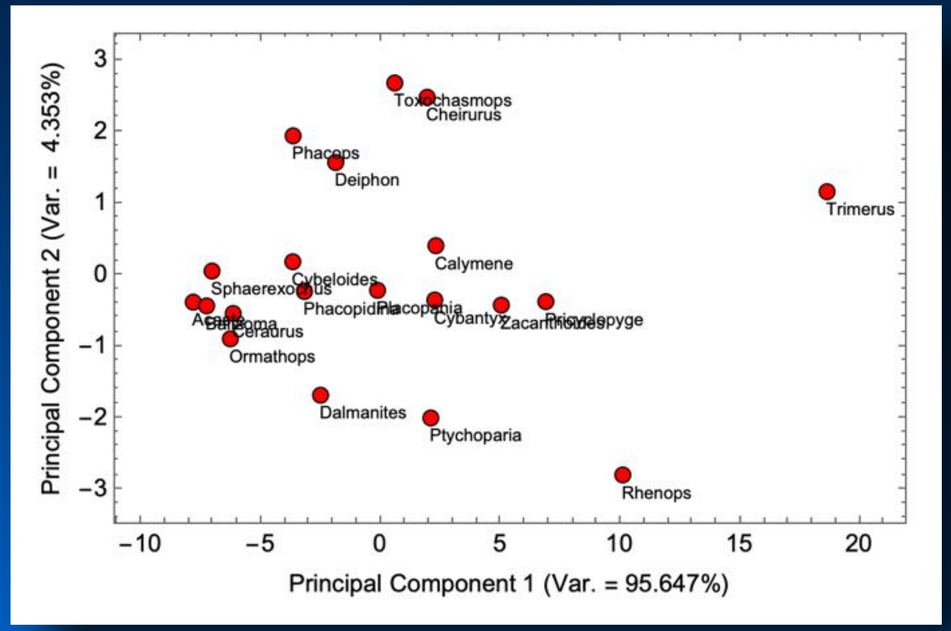
Principal Coordinate Ordination Space



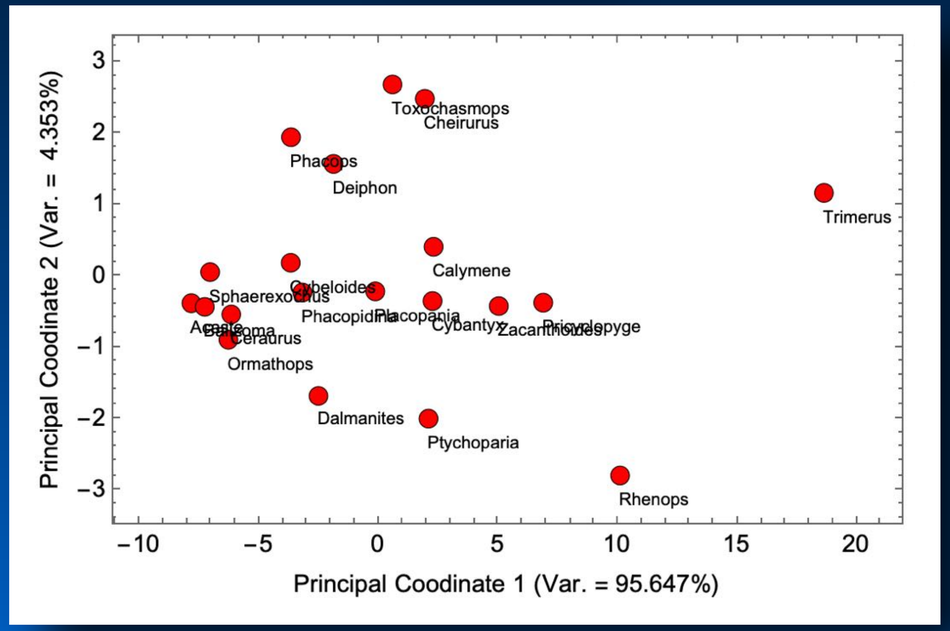
Principal Coordinates Analysis

Comparison of PCA & PCoord Results

PCA



PCoord



Because the projected scores for a covariance-based principal components analysis and a squared Euclidean distance-based principal coordinates analysis are identical, these two procedures are said to be “duals” of one another. However, this designation is restricted to these two methods calculated in these specific ways.

Principal Coordinates Analysis

What About Mixed-Mode Data?

Genus	Body Length (mm)	Glabella Length (mm)	Glabella Wdth (mm)	Pleural Lobes (n)
Acaste	23.14	3.50	3.77	11
Balizoma	14.32	3.97	4.08	11
Calymene	51.69	10.91	10.72	13
Ceraurus	21.15	4.90	4.69	9
Cheirurus	31.74	9.33	12.11	12
Cybantyx	36.81	11.35	10.10	9
Cybeloides	25.13	6.39	6.81	12
Dalmanites	32.93	8.46	6.08	11
Ormathops	13.88	5.03	4.34	9
Phacopidina	21.43	7.03	6.79	10
Phacops	27.23	5.30	8.19	11
Placoparia	38.15	9.40	8.71	12
Pricyclopyge	40.11	14.98	12.98	5
Ptychoparia	62.17	12.25	8.71	12
Rhenops	55.94	19.00	13.10	10
Sphaerexochus	23.31	3.84	4.60	11

Q-Mode Basis (Distance) Matrices

These indices are used quantify and scale relations between objects.

Euclidean Distance: $$d_{jk} = \sum_{i=0}^{n} \sqrt{(x_{ij} - x_{ik})^2}$$

Squared Euclidean Distance: $$d_{jk}^2 = \sum_{i=0}^{n} (x_{ij} - x_{ik})^2$$

Squared Mahalanobis Distance: $$d_{ij}^2 = \left[\bar{X}_i - \bar{X}_j\right]' \cdot \left[s_{ij}^2\right] \cdot \left[\bar{X}_i - \bar{X}_j\right]$$

Mahalanobis Distance: $$d_{ij} = \sqrt{d_{ij}^2}$$

Cosine θ:

$$\theta_{jk} = \frac{\sum_{i=1}^{n} x_{ik} x_{jk}}{\sqrt{\sum_{i=1}^{n} x_{ik}^2 \cdot \sum_{i=1}^{n} x_{jk}^2}}$$

Gower Distance:

$$d_{ij} = \frac{\frac{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|}{Range_k}}{p}$$

Q-Mode Basis (Distance) Matrices

These indices are used quantify and scale relations between objects.

Euclidean Distance: $$d_{jk} = \sum_{i=0}^{n} \sqrt{(x_{ij} - x_{ik})^2}$$

Squared Euclidean Distance: $$d_{jk}^2 = \sum_{i=0}^{n} (x_{ij} - x_{ik})^2$$

Squared Mahalanobis Distance: $$d_{ij}^2 = \left[\bar{X}_i - \bar{X}_j\right]' \cdot \left[s_{ij}^2\right] \cdot \left[\bar{X}_i - \bar{X}_j\right]$$

Mahalanobis Distance: $$d_{ij} = \sqrt{d_{ij}^2}$$

Cosine θ:

$$\theta_{jk} = \frac{\sum_{i=1}^{n} x_{ik} x_{jk}}{\sqrt{\sum_{i=1}^{n} x_{ik}^2 \cdot \sum_{i=1}^{n} x_{jk}^2}}$$

Gower Distance:

$$d_{ij} = \frac{\frac{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|}{Range_k}}{p}$$

Q-Mode Basis (Distance) Matrix

Gower Distance Matrix

Genus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	0.06	0.52	0.12	0.39	0.43	0.17	0.19	0.15	0.18	0.17	0.34	0.71	0.51	0.70	0.03
2	0.06	0	0.55	0.13	0.42	0.46	0.20	0.22	0.09	0.19	0.20	0.37	0.74	0.54	0.73	0.06
3	0.52	0.55	0	0.54	0.20	0.23	0.35	0.32	0.59	0.42	0.35	0.18	0.44	0.16	0.31	0.49
4	0.12	0.13	0.54	0	0.42	0.33	0.20	0.22	0.05	0.12	0.19	0.36	0.61	0.53	0.66	0.09
5	0.39	0.42	0.20	0.42	0	0.21	0.22	0.21	0.46	0.30	0.22	0.13	0.38	0.30	0.37	0.36
6	0.43	0.46	0.23	0.33	0.21	0	0.32	0.24	0.37	0.27	0.26	0.17	0.28	0.28	0.33	0.40
7	0.17	0.20	0.35	0.20	0.22	0.32	0	0.12	0.24	0.09	0.10	0.17	0.60	0.34	0.59	0.14
8	0.19	0.22	0.32	0.22	0.21	0.24	0.12	0	0.26	0.13	0.14	0.14	0.51	0.31	0.51	0.16
9	0.15	0.09	0.59	0.05	0.46	0.37	0.24	0.26	0	0.17	0.24	0.41	0.65	0.58	0.71	0.14
10	0.18	0.19	0.42	0.12	0.30	0.27	0.09	0.13	0.17	0	0.13	0.24	0.55	0.41	0.54	0.15
11	0.17	0.20	0.35	0.19	0.22	0.26	0.10	0.14	0.24	0.13	0	0.17	0.54	0.34	0.53	0.14
12	0.34	0.37	0.18	0.36	0.13	0.17	0.17	0.14	0.41	0.24	0.17	0	0.43	0.17	0.43	0.31
13	0.71	0.74	0.44	0.61	0.38	0.28	0.60	0.51	0.65	0.55	0.54	0.43	0	0.49	0.31	0.68
14	0.51	0.54	0.16	0.53	0.30	0.28	0.34	0.31	0.58	0.41	0.34	0.17	0.49	0	0.32	0.48
15	0.70	0.73	0.31	0.66	0.37	0.33	0.59	0.51	0.71	0.54	0.53	0.43	0.31	0.32	0	0.67
16	0.03	0.06	0.49	0.09	0.36	0.40	0.14	0.16	0.14	0.15	0.14	0.31	0.68	0.48	0.67	0

Principal Coordinates Analysis

Data Normalization Step

$$a_{ij} = \frac{(\bar{d}_i + \bar{d}_j - \bar{d}_{GM} - d_{ij}}{2}$$

Where: $\bar{d}_i$ = mean of row i;

$\bar{d}_j$ = mean of column j;

$\bar{d}_{GM}$ = grand mean of all rows and columns.

The effect of this operation is to center and “close” the dataset such that the sum of each row and each column is 0.0.

Q-Mode Basis (Distance) Matrix

Normalized Gower Distance Matrix

Genus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.14	0.12	-0.09	0.07	-0.06	-0.08	0.03	0.01	0.08	0.02	0.02	-0.05	-0.12	-0.08	-0.12	0.11
2	0.12	0.16	-0.10	0.08	-0.07	-0.09	0.02	0.01	0.12	0.03	0.02	-0.06	-0.12	-0.09	-0.12	0.10
3	-0.09	-0.10	0.20	-0.11	0.07	0.05	-0.03	-0.02	-0.11	-0.07	-0.04	0.06	0.05	0.12	0.11	-0.09
4	0.07	0.08	-0.11	0.13	-0.08	-0.03	0.01	-0.00	0.12	0.05	0.01	-0.07	-0.07	-0.10	-0.10	0.08
5	-0.06	-0.07	0.07	-0.08	0.13	0.03	-0.00	-0.00	-0.08	-0.04	-0.01	0.05	0.05	0.02	0.05	-0.06
6	-0.08	-0.09	0.05	-0.03	0.03	0.13	-0.05	-0.01	-0.04	-0.02	-0.03	0.03	0.10	0.03	0.06	-0.08
7	0.03	0.02	-0.03	0.01	-0.00	-0.05	0.09	0.02	0.01	0.04	0.03	0.01	-0.09	-0.02	-0.09	0.03
8	0.01	0.01	-0.02	-0.00	-0.00	-0.01	0.02	0.08	-0.01	0.02	0.01	0.01	-0.05	-0.02	-0.05	0.01
9	0.08	0.12	-0.11	0.12	-0.08	-0.04	0.01	-0.01	0.17	0.04	0.00	-0.07	-0.07	-0.10	-0.11	0.07
10	0.02	0.03	-0.07	0.05	-0.04	-0.02	0.04	0.02	0.04	0.09	0.02	-0.03	-0.06	-0.06	-0.06	0.03
11	0.02	0.02	-0.04	0.01	-0.01	-0.03	0.03	0.01	0.00	0.02	0.08	0.00	-0.06	-0.03	-0.06	0.03
12	-0.05	-0.06	0.06	-0.07	0.05	0.03	0.01	0.01	-0.07	-0.03	0.00	0.10	0.00	0.07	-0.00	-0.05
13	-0.12	-0.12	0.05	-0.07	0.05	0.10	-0.09	-0.05	-0.07	-0.06	-0.06	0.00	0.34	0.03	0.18	-0.11
14	-0.08	-0.09	0.12	-0.10	0.02	0.03	-0.02	-0.02	-0.10	-0.06	-0.03	0.07	0.03	0.21	0.11	-0.08
15	-0.12	-0.12	0.11	-0.10	0.05	0.06	-0.09	-0.05	-0.11	-0.06	-0.06	-0.00	0.18	0.11	0.33	-0.11
16	0.11	0.10	-0.09	0.08	-0.06	-0.08	0.03	0.01	0.07	0.03	0.03	-0.05	-0.11	-0.08	-0.11	0.12

Q-Mode Basis (Distance) Matrix

Normalized Gower Distance Matrix

Genus	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	**0.14**	0.12	-0.09	0.07	-0.06	-0.08	0.03	0.01	0.08	0.02	0.02	-0.05	-0.12	-0.08	-0.12	0.11
2	0.12	**0.16**	-0.10	0.08	-0.07	-0.09	0.02	0.01	0.12	0.03	0.02	-0.06	-0.12	-0.09	-0.12	0.10
3	-0.09	-0.10	**0.20**	-0.11	0.07	0.05	-0.03	-0.02	-0.11	-0.07	-0.04	0.06	0.05	0.12	0.11	-0.09
4	0.07	0.08	-0.11	**0.13**	-0.08	-0.03	0.01	-0.00	0.12	0.05	0.01	-0.07	-0.07	-0.10	-0.10	0.08
5	-0.06	-0.07	0.07	-0.08	**0.13**	0.03	-0.00	-0.00	-0.08	-0.04	-0.01	0.05	0.05	0.02	0.05	-0.06
6	-0.08	-0.09	0.05	-0.03	0.03	**0.13**	-0.05	-0.01	-0.04	-0.02	-0.03	0.03	0.10	0.03	0.06	-0.08
7	0.03	0.02	-0.03	0.01	-0.00	-0.05	**0.09**	0.02	0.01	0.04	0.03	0.01	-0.09	-0.02	-0.09	0.03
8	0.01	0.01	-0.02	-0.00	-0.00	-0.01	0.02	**0.08**	-0.01	0.02	0.01	0.01	-0.05	-0.02	-0.05	0.01
9	0.08	0.12	-0.11	0.12	-0.08	-0.04	0.01	-0.01	**0.17**	0.04	0.00	-0.07	-0.07	-0.10	-0.11	0.07
10	0.02	0.03	-0.07	0.05	-0.04	-0.02	0.04	0.02	0.04	**0.09**	0.02	-0.03	-0.06	-0.06	-0.06	0.03
11	0.02	0.02	-0.04	0.01	-0.01	-0.03	0.03	0.01	0.00	0.02	**0.08**	0.00	-0.06	-0.03	-0.06	0.03
12	-0.05	-0.06	0.06	-0.07	0.05	0.03	0.01	0.01	-0.07	-0.03	0.00	**0.10**	0.00	0.07	-0.00	-0.05
13	-0.12	-0.12	0.05	-0.07	0.05	0.10	-0.09	-0.05	-0.07	-0.06	-0.06	0.00	**0.34**	0.03	0.18	-0.11
14	-0.08	-0.09	0.12	-0.10	0.02	0.03	-0.02	-0.02	-0.10	-0.06	-0.03	0.07	0.03	**0.21**	0.11	-0.08
15	-0.12	-0.12	0.11	-0.10	0.05	0.06	-0.09	-0.05	-0.11	-0.06	-0.06	-0.00	0.18	0.11	**0.33**	-0.11
16	0.11	0.10	-0.09	0.08	-0.06	-0.08	0.03	0.01	0.07	0.03	0.03	-0.05	-0.11	-0.08	-0.11	**0.12**

Principal Coordinates Analysis

Principal Coordinates Eigensystem

Data

Genus	Body Length	Glabella Length	Glabella Wdth	Pleural Lobes (n)
Acaste	23.14	3.50	3.77	11
Balizoma	14.32	3.97	4.08	11
Calymene	51.69	10.91	10.72	13
Ceraurus	21.15	4.90	4.69	9
Cheirurus	31.74	9.33	12.11	12
Cybantyx	36.81	11.35	10.10	9
Cybeloides	25.13	6.39	6.81	12
Dalmanites	32.93	8.46	6.08	11
Ormathops	13.88	5.03	4.34	9
Phacopidina	21.43	7.03	6.79	10
Phacops	27.23	5.30	8.19	11
Placoparia	38.15	9.40	8.71	12
Pricyclopyge	40.11	14.98	12.98	5
Ptychoparia	62.17	12.25	8.71	12
Rhenops	55.94	19.00	13.10	10
Sphaerexochus	23.31	3.84	4.60	11

Eigenvalues

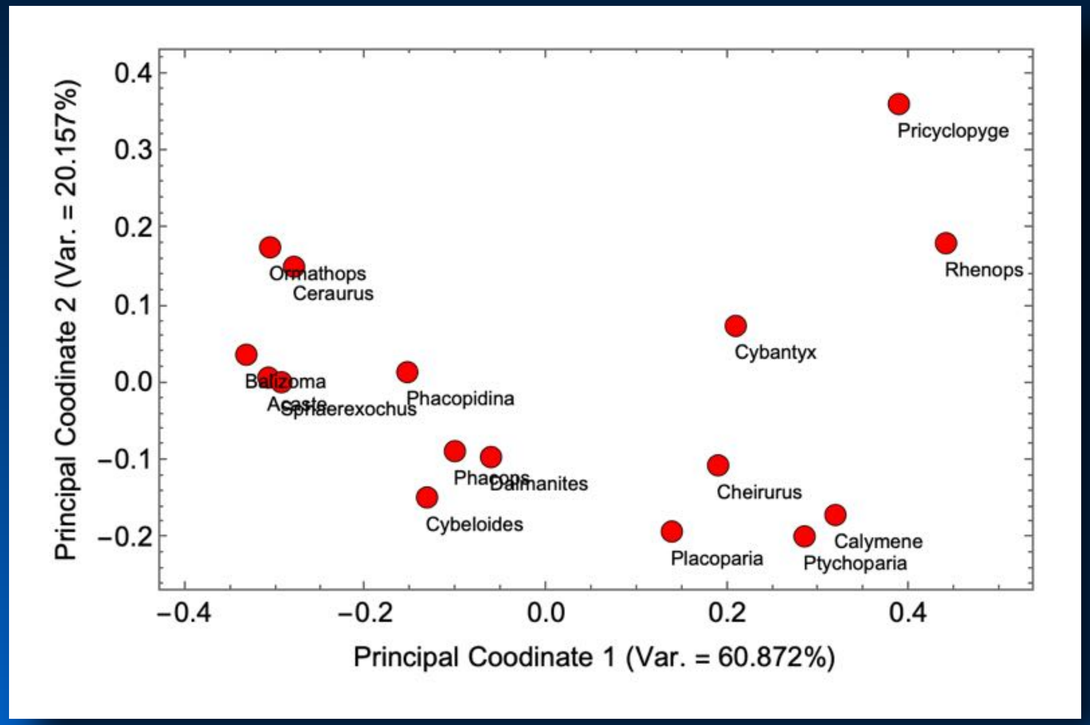
Eigenvector	1	2	3	4
Eigenvalue	1.15	0.38	0.21	0.14
Percent Variance	60.87	20.15	11.32	7.66
Cum. Percent Variance	60.87	81.02	92.34	100.00

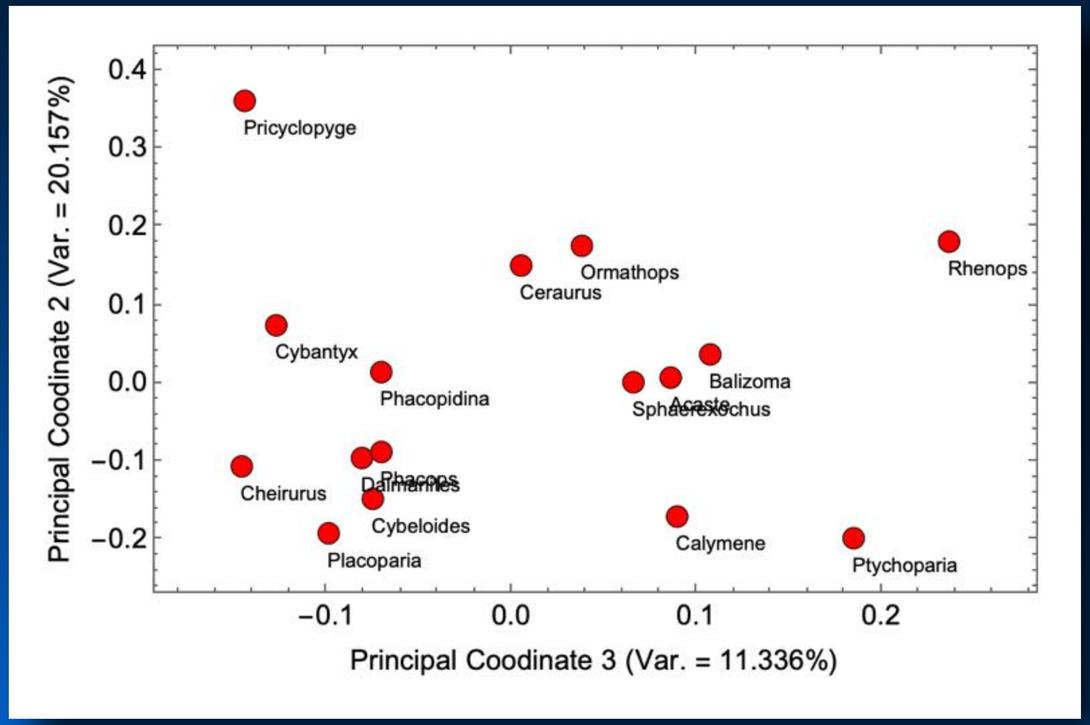
Eigenscores

Eigenvectors	1	2	3
Acaste	-0.288	0.011	0.186
Balizoma	-0.310	0.059	0.233
Calymene	0.298	-0.277	0.194
Ceraurus	-0.261	0.244	0.011
Cheirurus	0.177	-0.173	-0.316
Cybantyx	0.195	0.120	-0.276
Cybeloides	-0.124	-0.240	-0.163
Dalmanites	-0.058	-0.156	-0.175
Ormathops	-0.286	0.285	0.082
Phacopidina	-0.144	0.023	-0.153
Phacops	-0.095	-0.143	-0.153
Placoparia	0.129	-0.312	-0.214
Pricyclopyge	0.363	0.586	-0.313
Ptychoparia	0.266	-0.322	0.401
Rhenops	0.412	0.294	0.512
Sphaerexochus	-0.274	0.002	0.143

Principal Coordinates Analysis

Gower Principal Coordinate Space

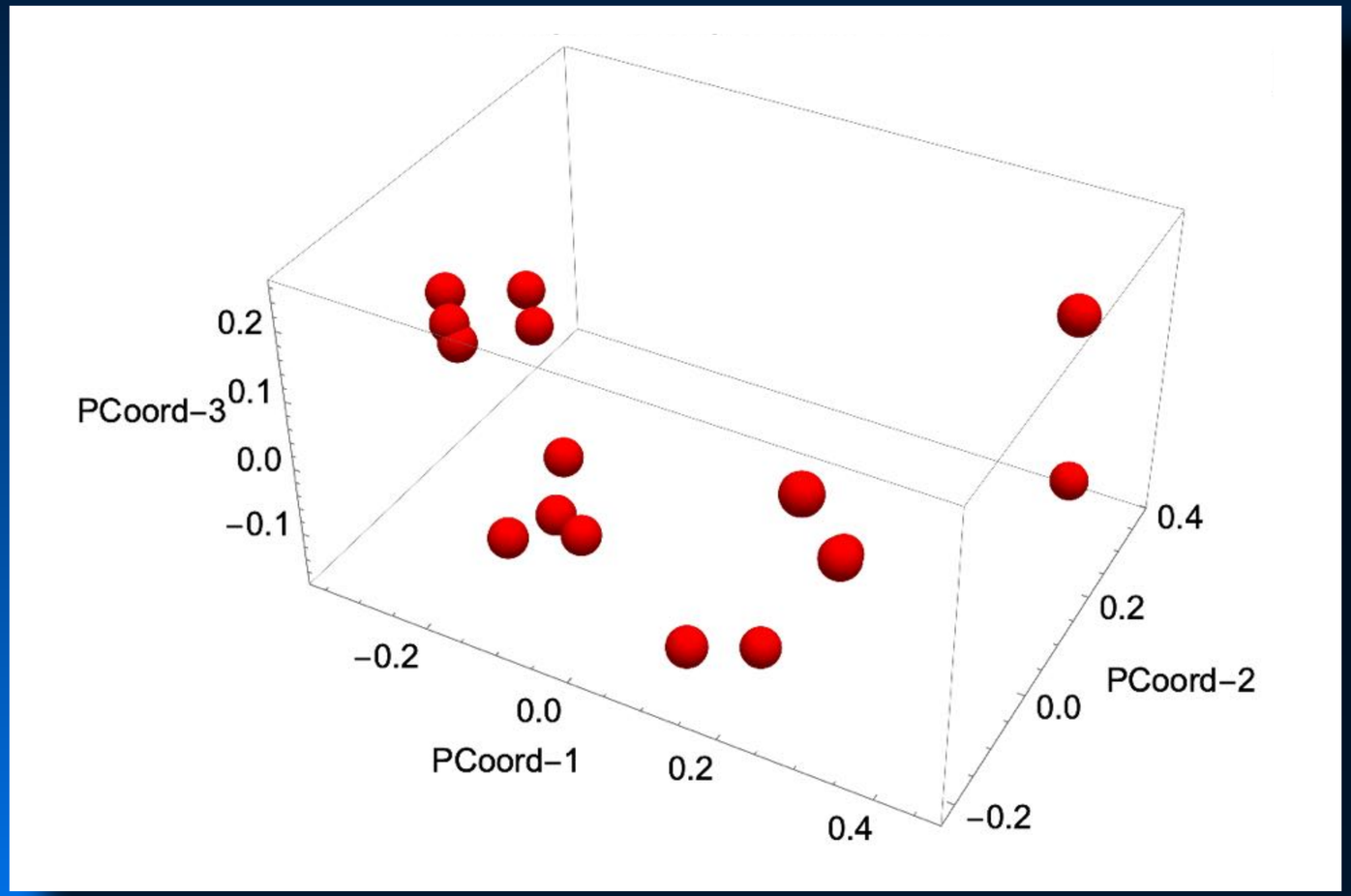




See if you can visualize the point cloud in 3D by comparing these plots. The plot on the right is a picture of the points looking down the PCoord-1 axis. Are *Pricyclopyge* and *Rhenops* really next to each other, or far apart, in this space?

Principal Coordinates Analysis

Gower Principal Coordinate Space



Singular Value Decomposition

Singular Value Decomposition

Eckart-Young Theorem

For any matrix of raw observations X of n rows and m columns there exists a set of matrices such that $X = ULV'$ where …

$$UU' = I \quad (r\text{-mode eigenvectors})$$
$$V'V = VV' = I \quad (Q\text{-mode eigenvectors})$$

… with the matrix L being the diagonal matrix of singular values (= square roots of the set of eigenvalues) arranged in decreasing order. These matrices can then be used to compute an estimator matrix B where …

$$B = UL_sV'$$

…whose values minimize the sum of the squared error between the elements of X and the corresponding elements of B for different values of s.

Singular Value Decomposition

Eckart-Young Theorem

Theory originally proposed by Carl Eckhart & Gale Young in 1936 in the first issue of the journal *Psychometrika*. The Algorithm for calculating these matrices wasoriginally devised by Golub and Reinsch (1971).

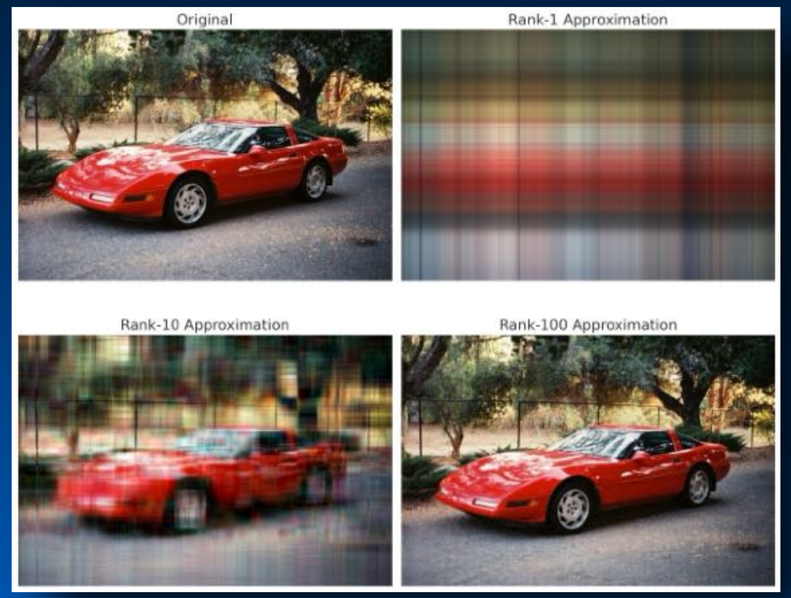
Used widely in computer algorithms for …

- … solving sets of simultaneous equations (esp. when these are represented by ill-conditioned matrices);
- … extraction of eigenvalues & eigenvectors from large data sets;
- … extraction of eigenvalues & eigenvectors from non-square matrices.

Carl Eckhart
(1902 – 1973)

Singular Value Decomposition

Practical Uses



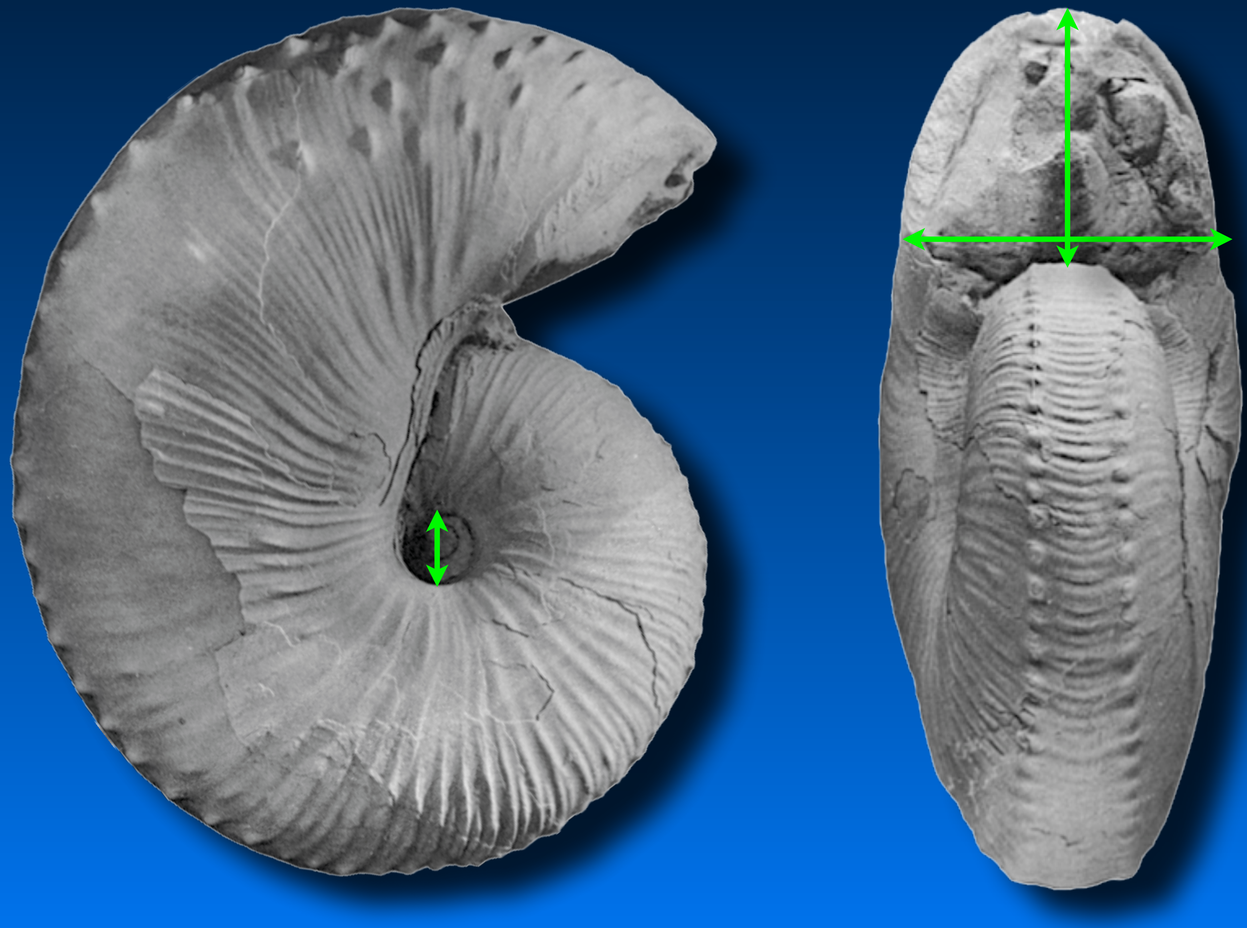
Data Compression

Recommendation Systems

Facial Recognition Systems

Singular Value Decomposition

Example: Ammonite Morphometrics



Species	Umbilical Diameter	Whorl Height	Whorl Width
A	4	27	18
B	12	25	12
C	10	23	16
D	14	21	14

Singular Value Decomposition

Example: Ammonite Morphometrics

Mean Center Data Matrix

$$X = \begin{pmatrix} 4 & 27 & 18 \\ 12 & 25 & 12 \\ 10 & 23 & 16 \\ 14 & 21 & 14 \end{pmatrix} \qquad X_{MC} = \begin{pmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{pmatrix}$$

$$\bar{X} = (10 \quad 24 \quad 15)$$

Singular Value Decomposition

Example: Ammonite Morphometrics

r-Mode Covariance Matrix

$$r = X' \cdot X$$

$$X = \begin{pmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{pmatrix}$$

$$X' = \begin{pmatrix} -6 & 2 & 0 & 4 \\ 3 & 1 & -1 & -3 \\ 3 & -3 & 1 & -1 \end{pmatrix}$$

$$r = \begin{pmatrix} 56 & -28 & -28 \\ -28 & 20 & 8 \\ -28 & 8 & 20 \end{pmatrix}$$

Singular Value Decomposition

Example: Ammonite Morphometrics

Q-Mode Similarity Matrix

$$Q = X \cdot X'$$

$$X = \begin{pmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{pmatrix} \qquad X' = \begin{pmatrix} -6 & 2 & 0 & 4 \\ 3 & 1 & -1 & -3 \\ 3 & -3 & 1 & -1 \end{pmatrix}$$

$$Q = \begin{pmatrix} 54 & -18 & 0 & -36 \\ -18 & 14 & -4 & 8 \\ 0 & -4 & 2 & 2 \\ -36 & 8 & 2 & 26 \end{pmatrix}$$

Singular Value Decomposition

Example: Ammonite Morphometrics

Singular Values & Eigenvalues

Singular Values

$$L^2 = \begin{pmatrix} 84 & 0 \\ 0 & 12 \end{pmatrix}$$

Eigenvalues

$$L = \begin{pmatrix} 9.1616 & 0 \\ 0 & 3.4641 \end{pmatrix}$$

Singular Value Decomposition

Example: Ammonite Morphometrics

Eigenvectors

r-Mode *(U)*

$$\begin{pmatrix} -0.8165 & 0.4082 \\ 0.0000 & -0.7071 \\ 0.5774 & 0.5774 \end{pmatrix}$$

Q-Mode *(V)*

$$\begin{pmatrix} 0.8018 & 0.0000 \\ -0.2673 & -0.8165 \\ 0.0000 & 0.4082 \\ -0.5345 & 0.4082 \end{pmatrix}$$

Singular Value Decomposition

Example: Ammonite Morphometrics

Loadings

r-Mode $(U \cdot L)$

$$\begin{pmatrix} -7.4833 & 0.0000 \\ 3.7413 & -2.4495 \\ 3.7313 & 2.4495 \end{pmatrix}$$

Q-Mode $(V \cdot L)$

$$\begin{pmatrix} 7.3485 & 0.0000 \\ -2.4495 & -2.8284 \\ 0.0000 & 1.4142 \\ -4.8990 & 1.4142 \end{pmatrix}$$

Singular Value Decomposition

Example: Ammonite Morphometrics

Projected Scores

r-Mode $(U \cdot L)$

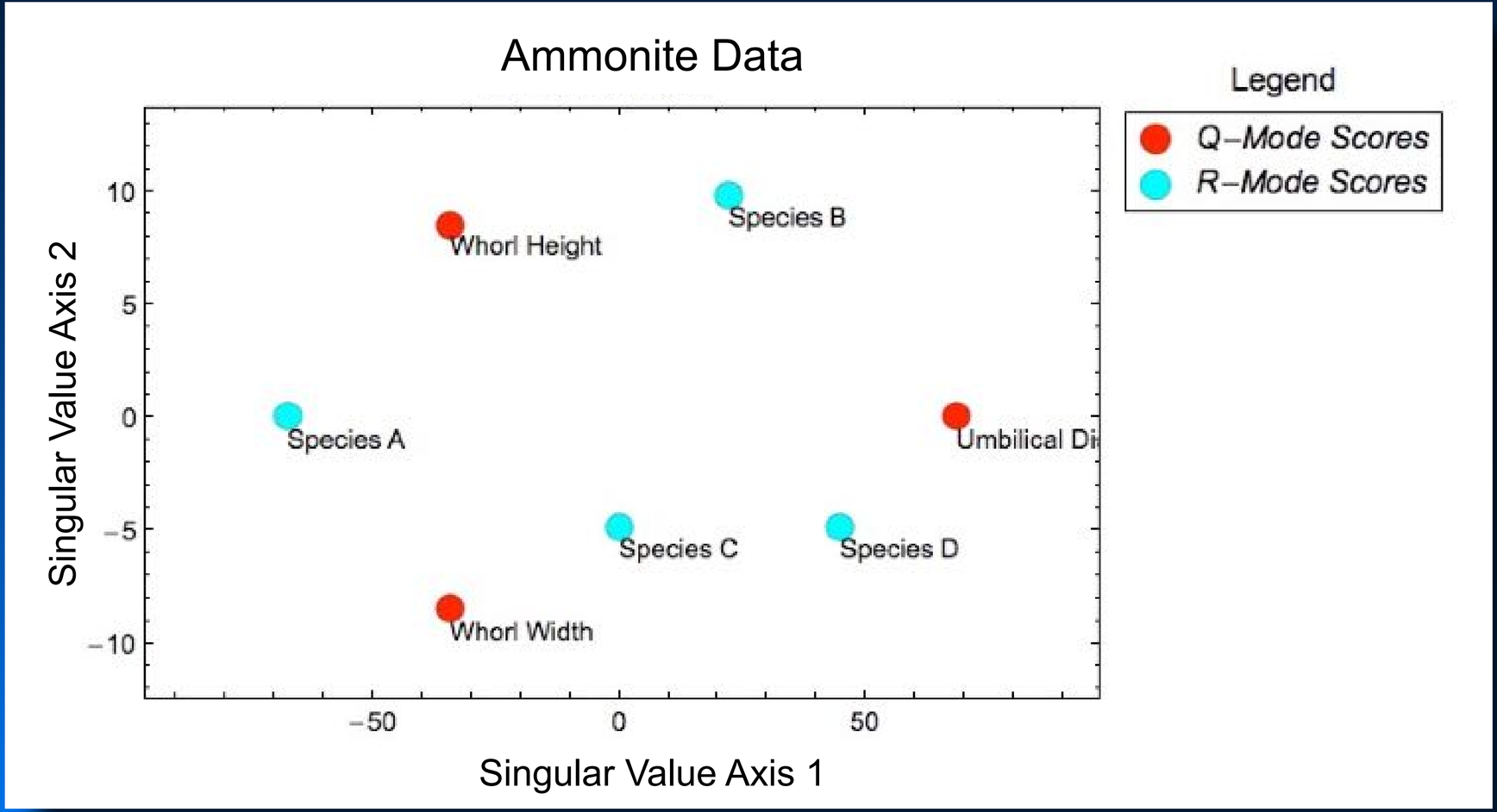
$$\begin{pmatrix} 67.5498 & 0.0000 \\ -22.4499 & -9.7980 \\ 0.0000 & 4.8990 \\ -44.8999 & 4.8990 \end{pmatrix}$$

Q-Mode $(V \cdot L)$

$$\begin{pmatrix} -68.5857 & 0.0000 \\ 34.2929 & -8.4853 \\ 34.2929 & 8.4853 \end{pmatrix}$$

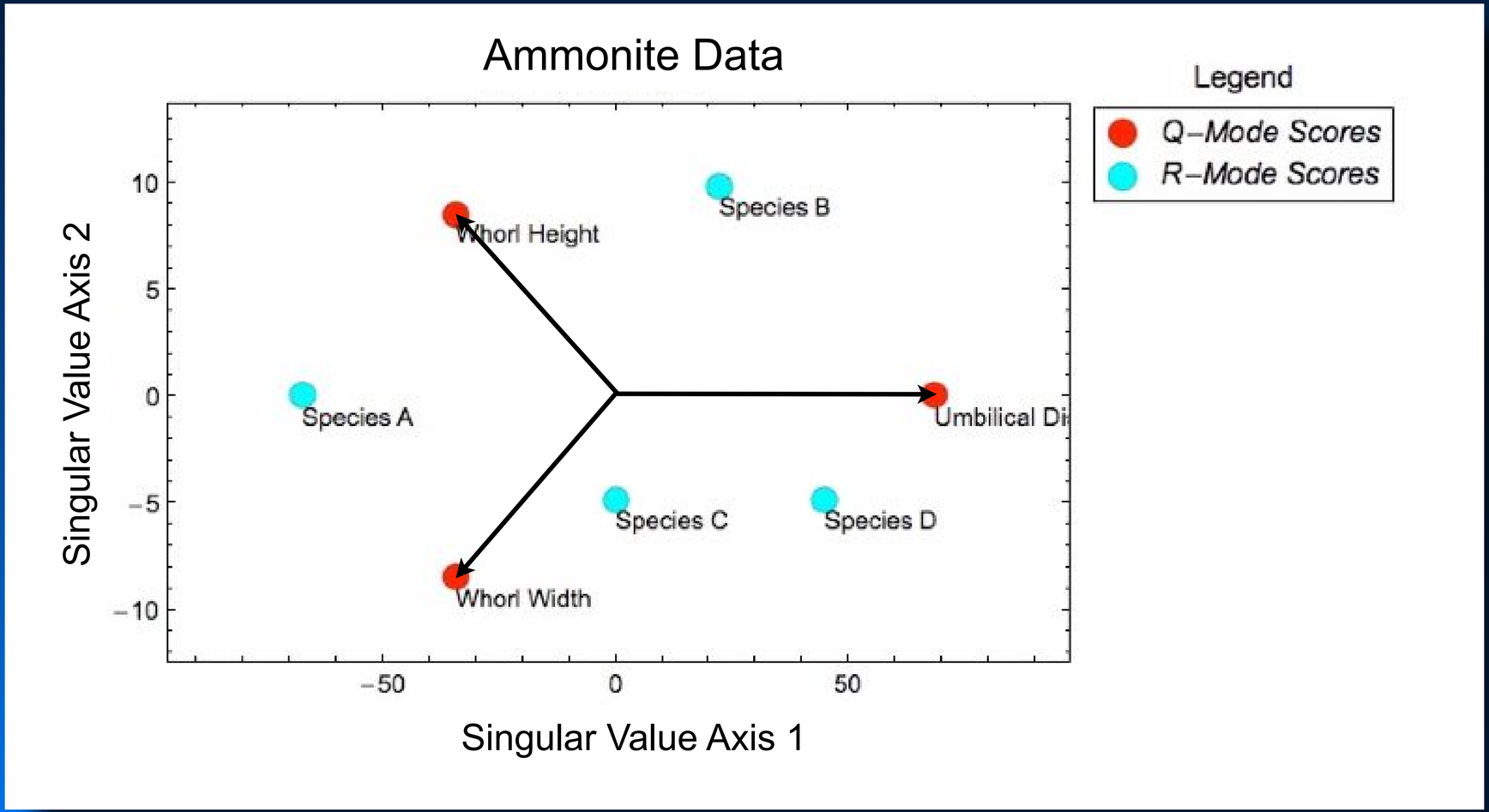
Singular Value Decomposition

Example: Ammonite Morphometrics



Singular Value Decomposition

Example: Ammonite Morphometrics



Singular Value Decomposition

Example: Ammonite Morphometrics

Proof

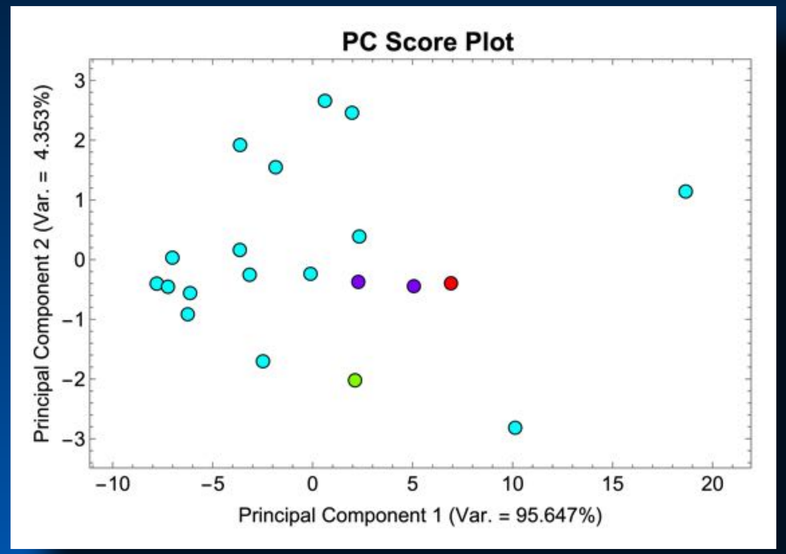
$$X = VLU'$$

$$X = \begin{pmatrix} 0.8018 & 0.0000 \\ -0.2673 & -0.8165 \\ 0.0000 & 0.4082 \\ -0.5345 & 0.4082 \end{pmatrix} \cdot \begin{pmatrix} 9.1616 & 0 \\ 0 & 3.4641 \end{pmatrix} \cdot \begin{pmatrix} -0.8165 & 0.0000 & 0.5774 \\ 0.4082 & -0.7071 & 0.5774 \end{pmatrix}$$

$$X = \begin{pmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{pmatrix}$$

Dimensionality Reduction

Prof. Norman MacLeod
School of Earth Sciences & Engineering, Nanjing University



$$d_{ij} = \frac{\frac{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|}{Range_k}}{p}$$

$$X = \begin{pmatrix} -6 & 3 & 3 \\ 2 & 1 & -3 \\ 0 & -1 & 1 \\ 4 & -3 & -1 \end{pmatrix}$$