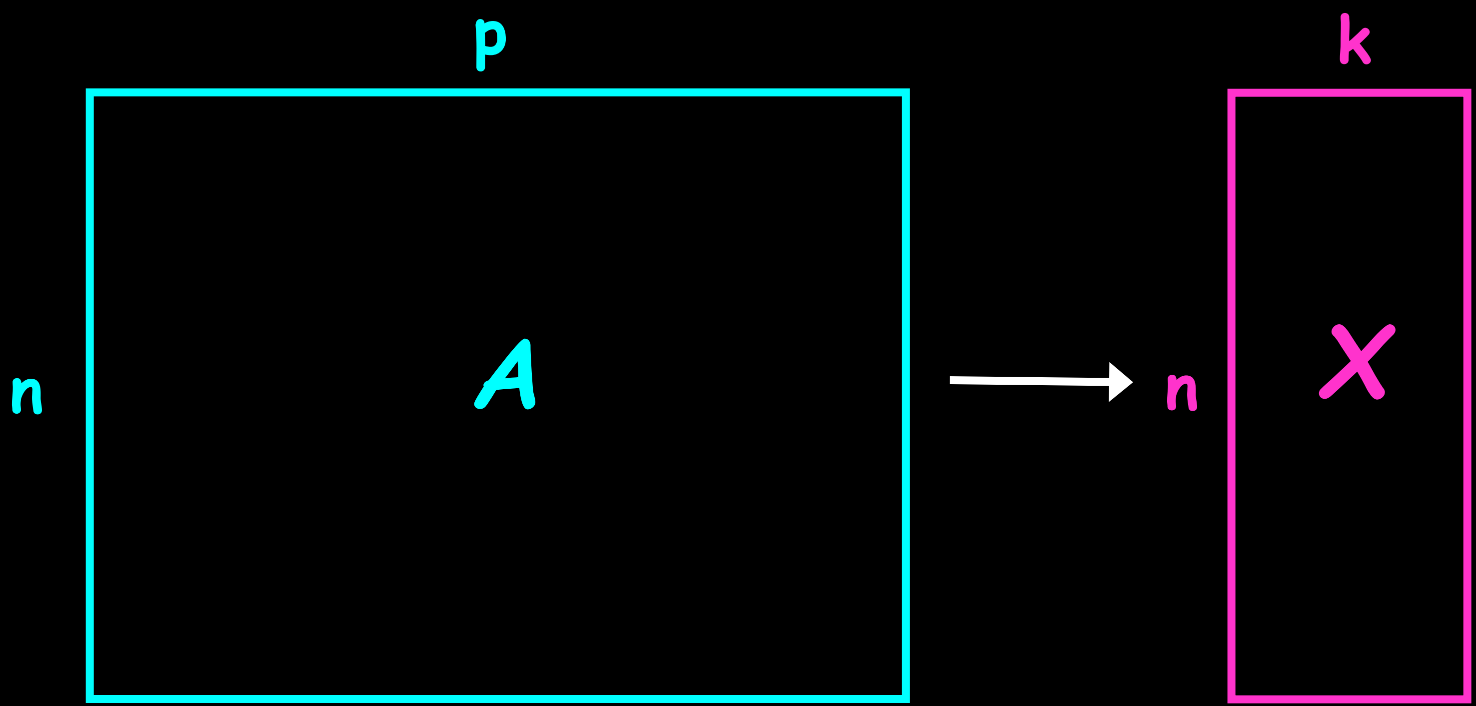
Principal Component Analysis (PCA)

Theory, Practice, and Examples

Data Reduction

- summarization of data with many (p) variables by a smaller set of (k) derived (synthetic, composite) variables.



Data Reduction

- "Residual" variation is information in *A* that is not retained in *X*
- balancing act between
  - clarity of representation, ease of understanding
  - oversimplification: loss of important or relevant information.

Principal Component Analysis (PCA)

- probably the most widely-used and well-known of the “standard” multivariate methods
- invented by Pearson (1901) and Hotelling (1933)
- first applied in ecology by Goodall (1954) under the name “factor analysis” (“principal factor analysis” is a synonym of PCA).

Principal Component Analysis (PCA)

- takes a data matrix of *n* objects by *p* variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original *p* variables
- the first *k* components display as much as possible of the variation among objects.

Geometric Rationale of PCA

- objects are represented as a cloud of *n* points in a multidimensional space with an axis for each of the *p* variables
- the **centroid** of the points is defined by the mean of each variable
- the **variance** of each variable is the average squared deviation of its *n* values around the mean of that variable.

$$V_i = \frac{1}{n-1}\sum_{m=1}^{n}\left(X_{im} - \overline{X}_i\right)^2$$

Geometric Rationale of PCA

- degree to which the variables are linearly correlated is represented by their **covariances**.

$$C_{ij} = \frac{1}{n-1}\sum_{m=1}^{n}\left(X_{im} - \overline{X}_i\right)\left(X_{jm} - \overline{X}_j\right)$$

Covariance of variables *i* and *j*

Sum over all *n* objects

Value of variable *i* in object *m*

Mean of variable *i*

Value of variable *j* in object *m*

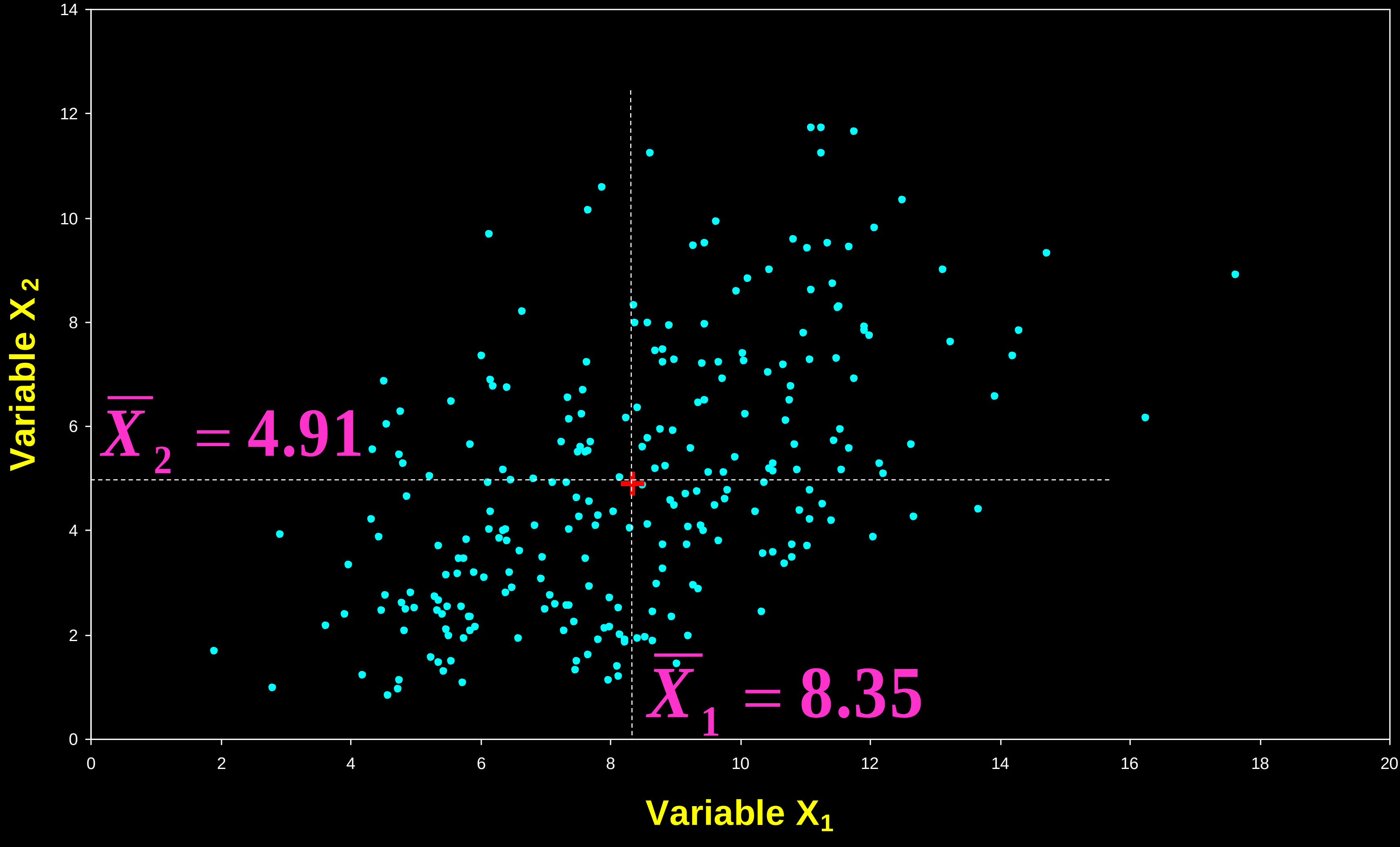
Mean of variable *j*

Geometric Rationale of PCA

- objective of PCA is to rigidly rotate the axes of this *p*-dimensional space to new positions (principal axes) that have the following properties:
  - ordered such that principal axis 1 has the highest variance, axis 2 has the next highest variance, ... , and axis *p* has the lowest variance
  - covariance among each pair of the principal axes is zero (the principal axes are uncorrelated).

2D Example of PCA

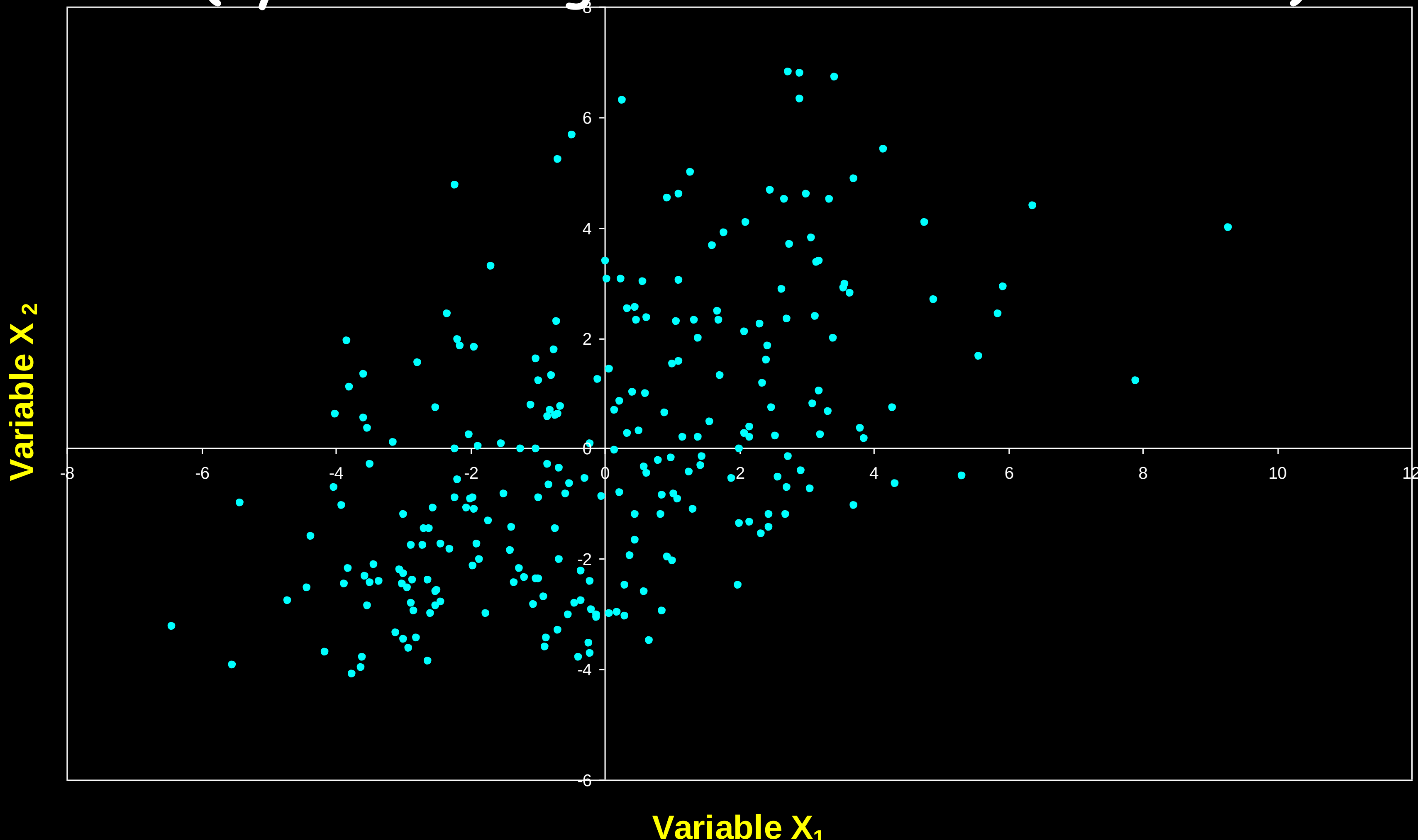
- variables X_1 and X_2 have positive covariance & each has a similar variance.



$$V_1 = 6.67 \qquad V_2 = 6.24 \qquad C_{1,2} = 3.42$$

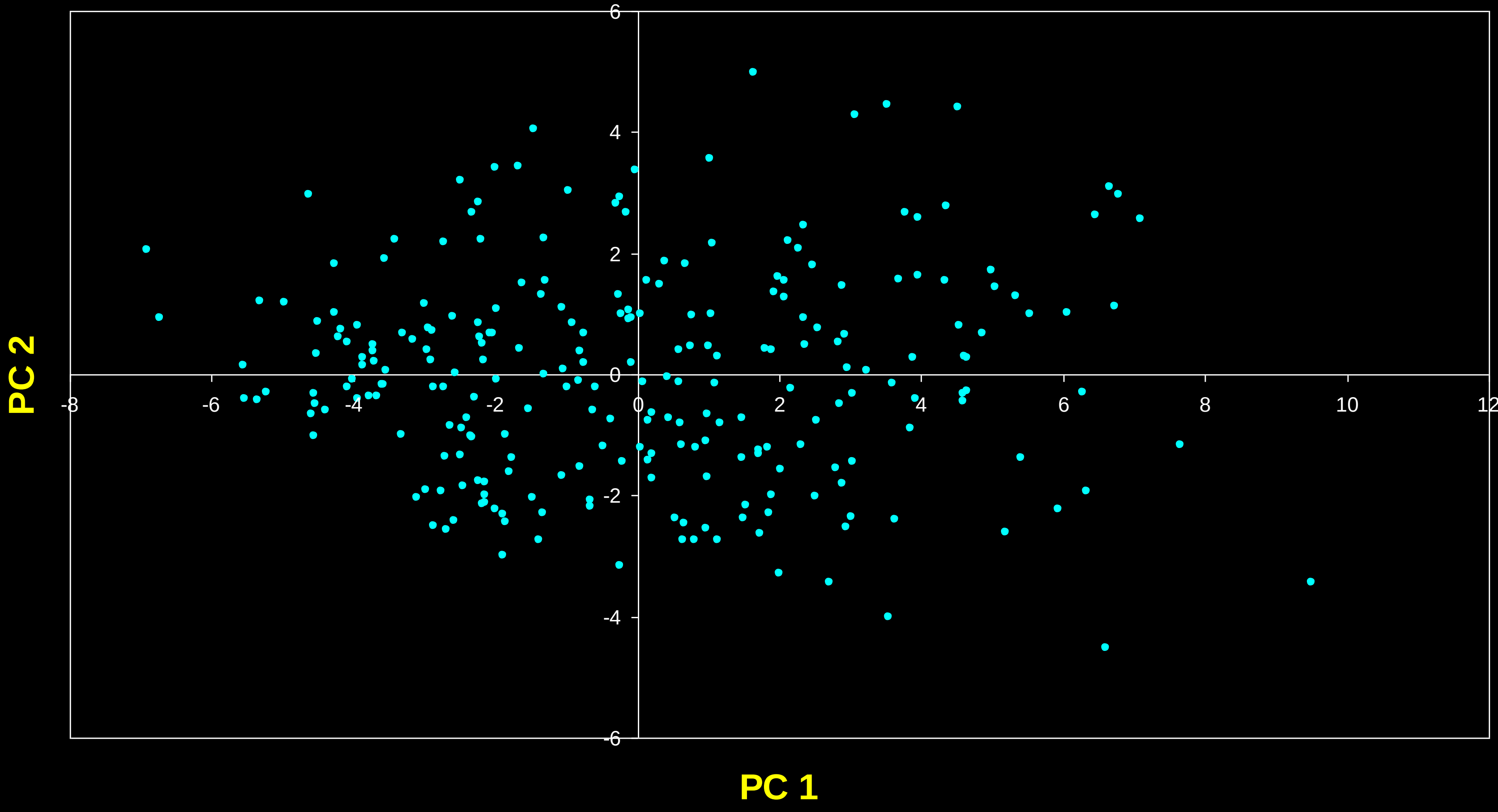
Configuration is Centered

- **each variable is adjusted to a mean of zero** (by subtracting the mean from each value).



Principal Components are Computed

- PC 1 has the highest possible variance (9.88)
- PC 2 has a variance of 3.03
- PC 1 and PC 2 have zero covariance.



The Dissimilarity Measure Used in PCA is Euclidean Distance

- PCA uses Euclidean Distance calculated from the *p* variables as the measure of dissimilarity among the n objects
- PCA derives the best possible *k* dimensional ($k < p$) representation of the Euclidean distances among objects.

Generalization to *p*-dimensions

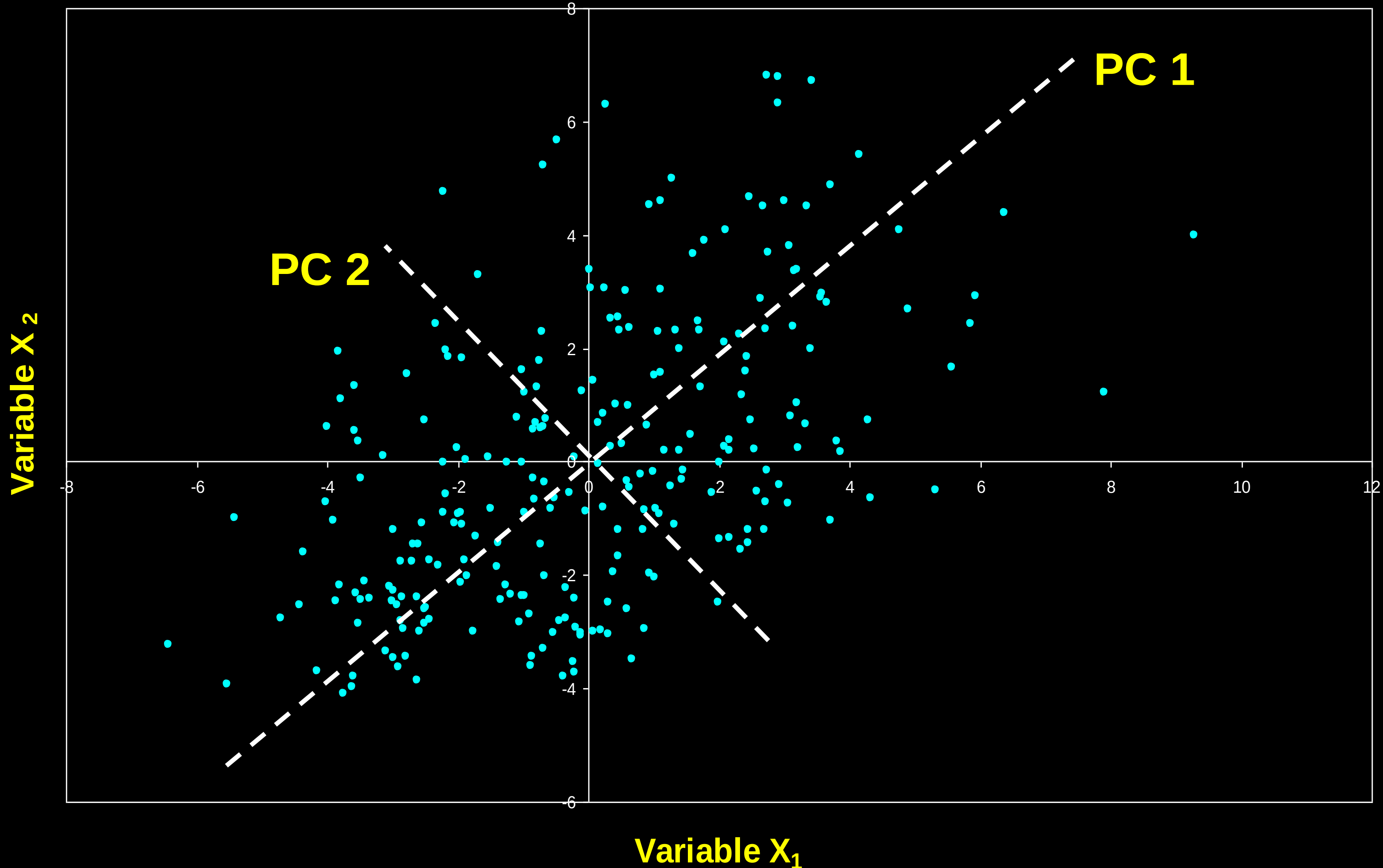
- In practice nobody uses PCA with only 2 variables
- The algebra for finding principal axes readily generalizes to *p* variables
- PC 1 is the direction of maximum variance in the *p*-dimensional cloud of points
- PC 2 is in the direction of the next highest variance, subject to the constraint that it has zero covariance with PC 1.

Generalization to *p*-dimensions

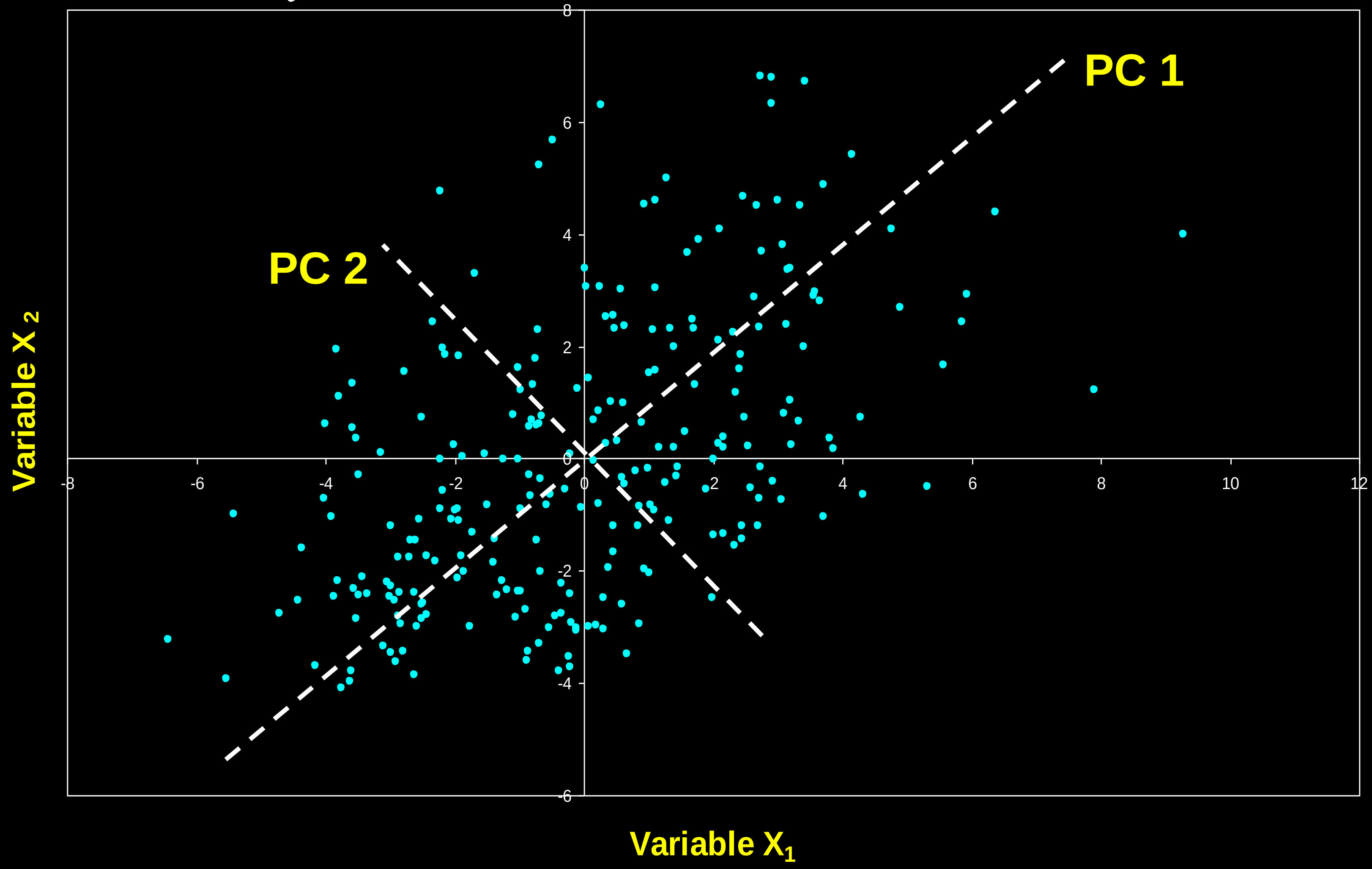
- PC 3 is in the direction of the next highest variance, subject to the constraint that it has zero covariance with both PC 1 and PC 2

- and so on... up to PC *p*

- each principal axis is a linear combination of the original two variables
- extended to p dimensions: $PC_i = a_{i1}X_1 + a_{i2}X_2 + \dots a_{in}X_p$
- a_{ij}'s are the coefficients for PC factor i, multiplied by the measured value for variable j



- PC axes are a rigid rotation of the original variables
- PC 1 is simultaneously the direction of maximum variance and a least-squares "line of best fit" (squared distances of points away from PC 1 are minimized).



Generalization to *p*-dimensions

- if we take the first *k* principal components, they define the *k*-dimensional "hyperplane of best fit" to the point cloud
- of the total variance of all *p* variables:
  - PCs 1 to *k* represent the maximum possible proportion of that variance that can be displayed in *k* dimensions
  - *i.e.* the squared Euclidean distances among points calculated from their coordinates on PCs 1 to *k* are the best possible representation of their squared Euclidean distances in the full *p* dimensions.

Covariance vs Correlation

- using covariances among variables only makes sense if they are measured in the same units
- even then, variables with high variances will dominate the principal components
- these problems are generally avoided by standardizing each variable to unit variance and zero mean.

$$X'_{im} = \frac{\left(X_{im} - \bar{X}_i\right)}{\mathrm{SD}_i}$$

Mean variable *i*

Standard deviation of variable *i*

Covariance vs Correlation

- covariances between the standardized variables are correlations
- after standardization, each variable has a variance of 1.000
- correlations can be also calculated from the variances and covariances:

Correlation between variables *i* and *j*

Covariance of variables *i* and *j*

$$r_{ij} = \frac{C_{ij}}{\sqrt{V_i V_j}}$$

Variance of variable *i*

Variance of variable *j*

The Algebra of PCA

- first step is to calculate the **cross-products matrix** of variances and covariances (or correlations) among every pair of the *p* variables
- square, symmetric matrix
- diagonals are the variances, off-diagonals are the covariances.

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Variance-covariance Matrix

	X_1	X_2
X_1	1.0000	0.5297
X_2	0.5297	1.0000

Correlation Matrix

The Algebra of PCA

- in matrix notation, this is computed as

$$\mathbf{S} = \mathbf{X}'\mathbf{X}$$

- where X is the *n* x *p* data matrix, with each variable centered (also standardized by SD if using correlations).

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Variance-covariance Matrix

	X_1	X_2
X_1	1.0000	0.5297
X_2	0.5297	1.0000

Correlation Matrix

Manipulating Matrices

- transposing: could change the columns to rows or the rows to columns

$$X = \begin{pmatrix} 10 & 0 & 4 \\ 7 & 1 & 2 \end{pmatrix} \qquad X' = \begin{pmatrix} 10 & 7 \\ 0 & 1 \\ 4 & 2 \end{pmatrix}$$

- multiplying matrices
  - must have the same number of columns in the premultiplicand matrix as the number of rows in the postmultiplicand matrix

The Algebra of PCA

- sum of the diagonals of the variance-covariance matrix is called the **trace**
- it represents the **total variance** in the data
- it is the mean squared Euclidean distance between each object and the centroid in *p*-dimensional space.

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Trace = 12.9091

	X_1	X_2
X_1	1.0000	0.5297
X_2	0.5297	1.0000

Trace = 2.0000

The Algebra of PCA

- finding the principal axes involves eigenanalysis of the cross-products matrix (S)

- the eigenvalues (latent roots) of S are solutions (λ) to the characteristic equation

$$\left|\mathbf{S} - \lambda\mathbf{I}\right| = \mathbf{0}$$

- S=UΛW^{-1} (Singular Value Decomposition)
  - U,W are orthogonal, store the Eigenvectors
  - Λ is a diagonal matrix, stores the Eigenvalues

The Algebra of PCA

- the eigenvalues, λ_1, λ_2, ... λ_p are the variances of the coordinates on each principal component axis
- the sum of all *p* eigenvalues equals the trace of *S* (the sum of the variances of the original variables).

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Trace = 12.9091

$\lambda_1 = 9.8783$

$\lambda_2 = 3.0308$

Note: $\lambda_1 + \lambda_2 = 12.9091$

The Algebra of PCA

- each eigenvector consists of *p* values which represent the "contribution" of each variable to the principal component axis

- eigenvectors are uncorrelated (orthogonal)
  - their cross-products are zero.

Eigenvectors

	u_1	u_2
X_1	0.7291	-0.6844
X_2	0.6844	0.7291

$$0.7291*(-0.6844) + 0.6844*0.7291 = 0$$

The Algebra of PCA

- assume there are *n* data objects, each with p attributes → data matrix X

- the coordinates of each object *i* on the k^{th} principal axis, known as the scores on PC *k*, are computed as

$$z_{ki} = u_{1k}x_{1i} + u_{2k}x_{2i} + \cdots + u_{pk}x_{pi}$$

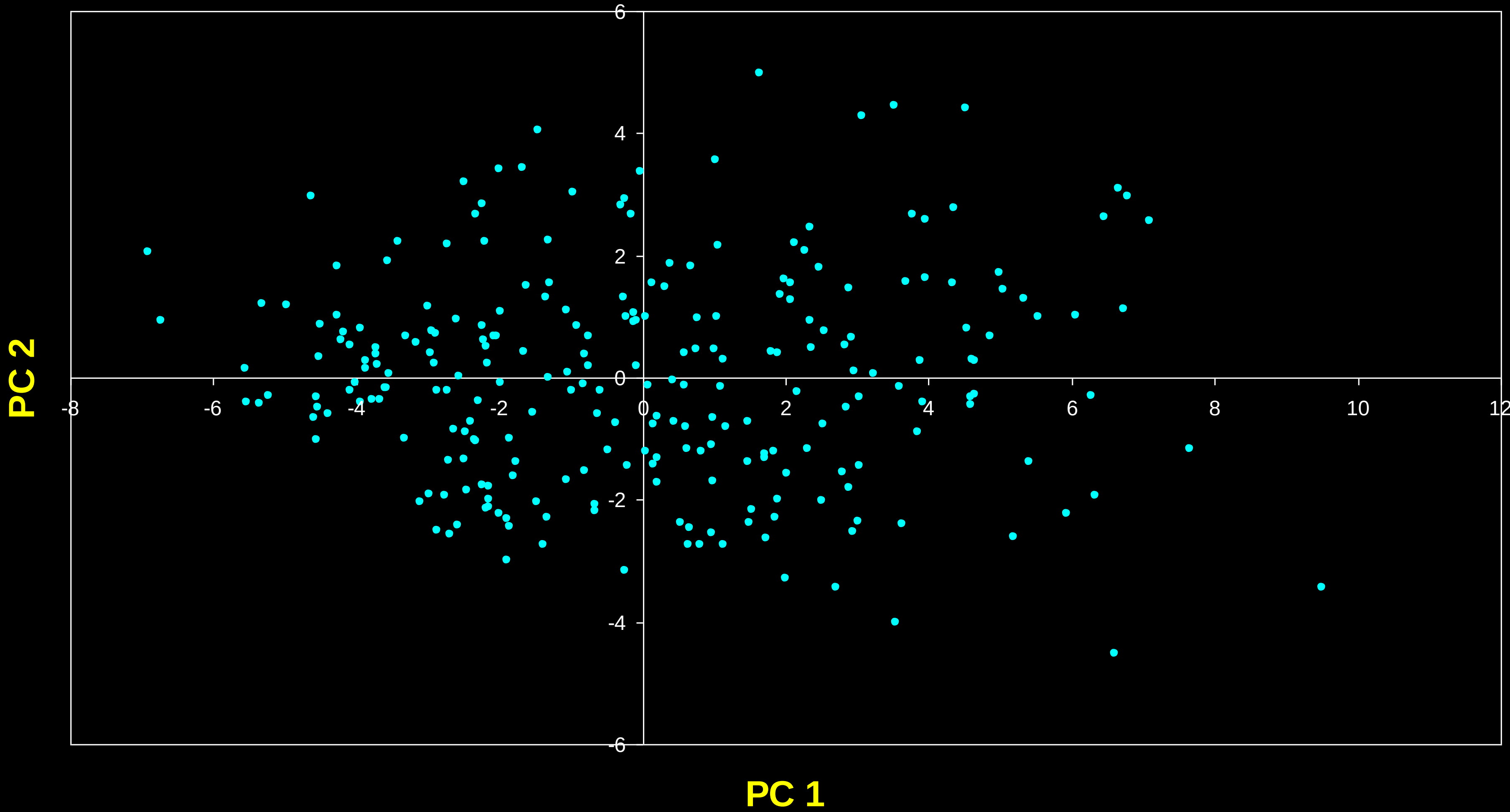
- where Z is the *n* x *k* matrix of PC scores, X is the *n* x *p* centered data matrix and U is the *p* x *k* matrix of eigenvectors.

The Algebra of PCA

- variance of the scores on each PC axis is equal to the corresponding eigenvalue for that axis
- the eigenvalue represents the variance displayed ("explained" or "extracted") by the k^{th} axis
- the sum of the first k eigenvalues is the variance explained by the k-dimensional ordination.

$\lambda_1 = 9.8783$ $\lambda_2 = 3.0308$ Trace = 12.9091

PC 1 displays ("explains")
9.8783/12.9091 = 76.5% of the total variance



The Algebra of PCA

- The cross-products matrix computed among the *p* principal axes has a simple form:
  - all off-diagonal values are zero (the principal axes are uncorrelated)
  - the diagonal values are the eigenvalues.

	PC_1	PC_2
PC_1	9.8783	0.0000
PC_2	0.0000	3.0308

Variance-covariance Matrix
of the PC axes

A more challenging example

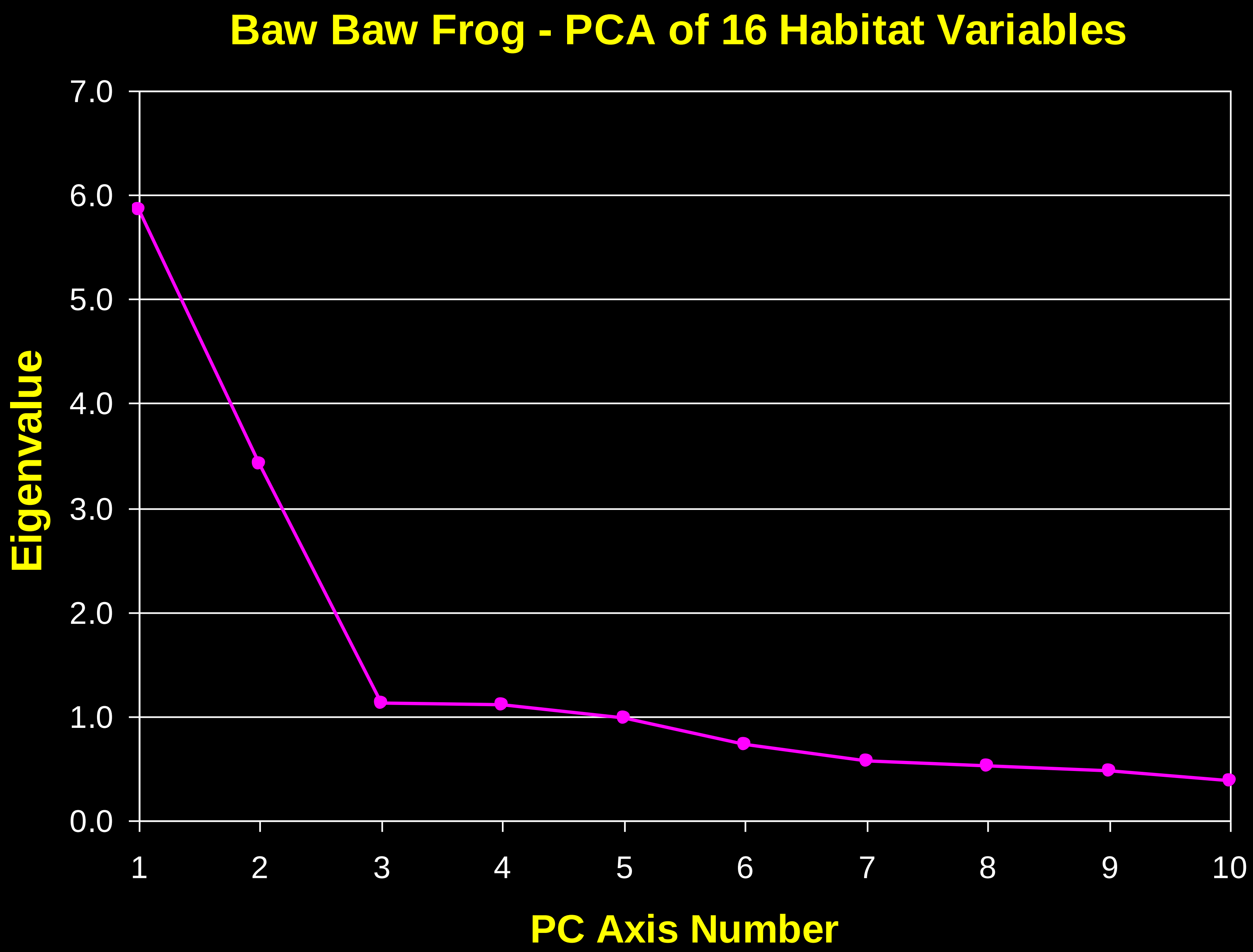
- data from research on habitat definition in the endangered Baw Baw frog
- 16 environmental and structural variables measured at each of 124 sites
- correlation matrix used because variables have different units

Eigenvalues

Axis	Eigenvalue	% of Variance	Cumulative % of Variance
1	5.855	36.60	36.60
2	3.420	21.38	57.97
3	1.122	7.01	64.98
4	1.116	6.97	71.95
5	0.982	6.14	78.09
6	0.725	4.53	82.62
7	0.563	3.52	86.14
8	0.529	3.31	89.45
9	0.476	2.98	92.42
10	0.375	2.35	94.77

How many axes are needed?

- does the $(k+1)^{th}$ principal axis represent more variance than would be expected by chance?
- several tests and rules have been proposed
- a common "rule of thumb" when PCA is based on correlations is that axes with eigenvalues > 1 are worth interpreting
- in our example 4 Eigenvectors fit this criterion (we shall keep 3 for simplicity)


Baw Baw Frog - PCA of 16 Habitat Variables
Eigenvalue
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
1
2
3
4
5
6
7
8
9
10
PC Axis Number

Interpreting Eigenvectors

- correlations between variables and the principal axes are known as loadings
- each element of the eigenvectors represents the contribution of a given variable to a component
- the loadings of variables on the first three PCs are shown here

	PC 1	PC 2	PC 3
Altitude	0.3842	0.0659	-0.1177
pH	-0.1159	0.1696	-0.5578
Cond	-0.2729	-0.1200	0.3636
TempSurf	0.0538	-0.2800	0.2621
Relief	-0.0765	0.3855	-0.1462
maxERht	0.0248	0.4879	0.2426
avERht	0.0599	0.4568	0.2497
%ER	0.0789	0.4223	0.2278
%VEG	0.3305	-0.2087	-0.0276
%LIT	-0.3053	0.1226	0.1145
%LOG	-0.3144	0.0402	-0.1067
%W	-0.0886	-0.0654	-0.1171
H1Moss	0.1364	-0.1262	0.4761
DistSWH	-0.3787	0.0101	0.0042
DistSW	-0.3494	-0.1283	0.1166
DistMF	0.3899	0.0586	-0.0175

Significance of Variables

- we can compute the significance of the variables as the sum of squared loadings on to the most significant Eigenvectors we selected (3 in our example)
- the next slide shows the table of the last slide expanded with these squared loadings
- we can then sort the table by the squared loadings and make a scree plot
- the most significant variables are those above some chosen cutoff, for example 0.4 (marked in yellow in the table)

Significance of Variables

	PC 1	PC 2	PC 3	sum of squared loadings
Altitude	0.3842	0.0659	-0.1177	0.41
pH	-0.1159	0.1696	-0.5578	0.59
Cond	-0.2729	-0.1200	0.3636	0.47
TempSurf	0.0538	-0.2800	0.2621	0.39
Relief	-0.0765	0.3855	-0.1462	0.42
maxERht	0.0248	0.4879	0.2426	0.55
avERht	0.0599	0.4568	0.2497	0.52
%ER	0.0789	0.4223	0.2278	0.49
%VEG	0.3305	-0.2087	-0.0276	0.39
%LIT	-0.3053	0.1226	0.1145	0.35
%LOG	-0.3144	0.0402	-0.1067	0.33
%W	-0.0886	-0.0654	-0.1171	0.16
H1Moss	0.1364	-0.1262	0.4761	0.51
DistSWH	-0.3787	0.0101	0.0042	0.38
DistSW	-0.3494	-0.1283	0.1166	0.39
DistMF	0.3899	0.0586	-0.0175	0.39

Significance of Variables

- Scree plot