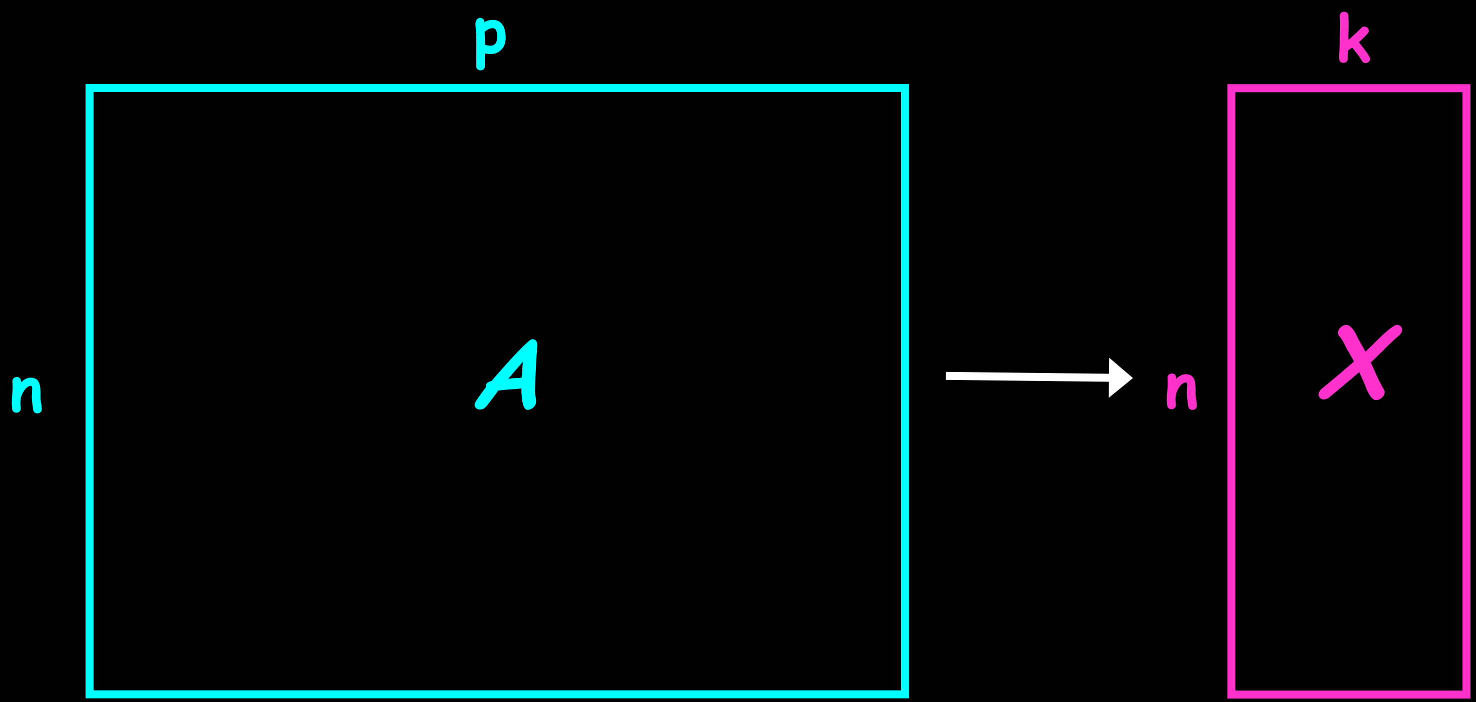
Principal Component Analysis (PCA)

Utkarsh Kulshrestha
Data Scientist - TCS
LearnBay

Data Reduction

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



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.

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).

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.

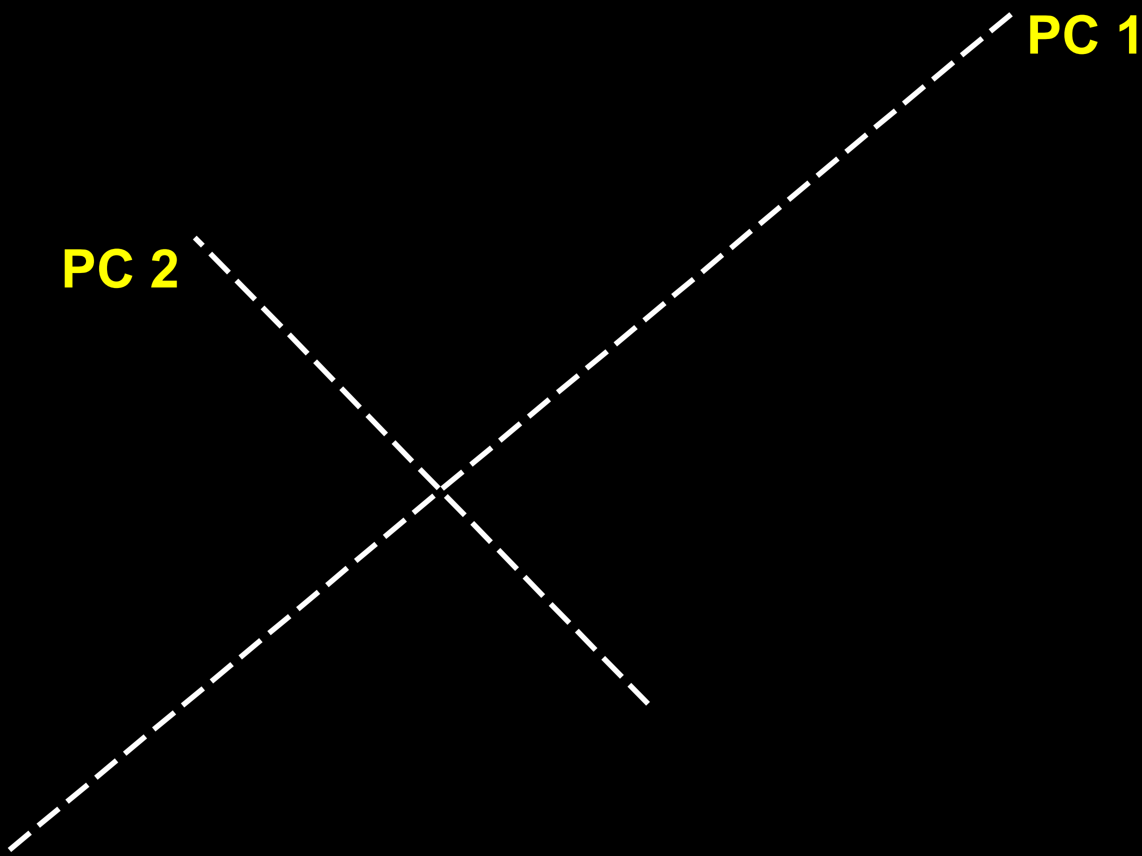
Generalization to *p*-dimensions

- 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
- $PC_j = a_{i1}Y_1 + a_{i2}Y_2 + \ldots a_{in}Y_n$
- a_{ij}'s are the coefficients for factor i, multiplied by the measured value for variable j



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

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

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

	1	2	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

When should PCA be used?

- In community ecology, PCA is useful for summarizing variables whose relationships are approximately linear or at least monotonic
  - *e.g.* A PCA of many soil properties might be used to extract a few components that summarize main dimensions of soil variation
- PCA is generally NOT useful for ordinating community data
- Why? Because relationships among species are highly nonlinear.

THANK YOU !!!