Principal Components Analysis

LECTURE 03

Objectives

- At the end of this series of lectures you should be able to:
  - Define terms.
  - Understand the basics of ordination.
  - Describe PCA.
  - Describe the use and limitations of PCA.
  - Explain the assumptions of PCA.
  - Interpret the results of PCA.
  - Perform PCA using R

Review

- Biologists often analyze data using a methodological triad:
  - Direct gradient analysis
    - Examines how one variables (dependent variable) changes with another variable (independent variable or gradient)
    - Most univariate techniques, MANOVA, and Multiple regression
  - Classification
    - Groups cases based on their relationships indicated by their different variables.
    - Cluster analysis and related techniques.
  - Ordination
    - Reduce the dimensionality of the data matrix such that similar entities are close together and dissimilar entities are far apart.
    - Principal Components Analysis (PCA) and Correspondence Analysis (CA) et al.

Review

- Ordination
  - Ordination is a collective term for multivariate techniques that that arrange cases along axes based on variable values.
  - The most common applications within ecology are the arrangement of sites based on their species composition.
  - However, by noting the arrangement of the sites and knowing something about the sites' environmental characteristics you can infer the importance of those variables on species occurrence or species composition. -- Indirect gradient analysis.

Review

- Advantages of indirect gradient analysis
  - Species composition data is relatively easy to collect while it is difficult to characterize completely the environment. Ordination analysis can point to previously unrecognized environmental variables.
  - The actual occurrence of a species may be too unpredictable to discover the relation to environmental conditions by direct means.
  - Ecosystem or landscape level analysis or planning may be more interested in species composition than the occurrence of individual species. Incorporating more than a few results of direct gradient analysis may be quite difficult.
- There are many different ordination techniques. (PCA, PCO, CCA, CCO, DCA).

Review

- Principal Components Analysis is a good introduction to ordination techniques because:
  - It is a very general technique with wide applicability.
  - It based on linear correlation (similar to Pearson's Correlation), so it is easy to understand.
  - It is available in a wide variety of statistical software packages.
- PCA and PCO are types of factor analysis.

Overview

- Intent -- Statistical
  - Reduce the dimesionality of the data set (number of variables) such that the relationships between the cases can be better assessed.
  - To construct new axes (variables) in which the most variation in the original dataset can be represented in as few dimensions as possible.
  - The new axes minimize the amount of shared variance.

Overview

- Intent -- Uses
  - Reduce the number of variables to better visualize the data.
  - Confirm classification schemes.
  - Exploratory data analysis.

Overview

- Data requirements
  - Independent variable – None
  - Dependent variable – None

Overview

- Data requirement
  - Factors are usually ratio/interval, however other types of variables (ordinal and dichotomous) can be used but it tends to increase the difficulty of interpreting the results.
  - Irrelevant variables will have a substantive detrimental effect on the results.
  - Sensitive to outliers and care should be taken to remove them from the data set.
  - The number of cases should be 100 or greater or 5 times the number of variables. (A harsh recommendation)
  - The number of variables should be 3 times number of axes extracted from the dataset. (A harsh recommendation)

Overview

- Rationale
  - By example.

Assumptions

- Assumes that the data is truly representative of the population as opposed to a sample of the population.
- Inclusion of all pertinent variables and exclusion of extraneous variables.
- Linearity.
- Multivariate normality (but only if using significance tests).
- There is structure to the dataset (underlying dimensions that can be interpreted).
- Some level of multicollinearity in the data set.

Issues

- Outliers
- Arch effect
- Horseshoe effect

Issues

- Determining the Number of Eigenvectors
  - The objective of PCA is to identify the smallest number of factors that together account for all of the total variance of the correlation matrix of the original variables.

Issues

- Determining the Number of Eigenvectors
  - How does one determine the number of factors to extract (i.e., to retain) in a given analysis? Several different types of stopping rules have been developed as an aid in answering this question.
    - Percentage of variance criterion
    - A priori criterion
    - Kaiser's stopping rule
    - Scree test

Issues

- Rotation
  - Simple structure
  - Unfortunately, most eigenvectors do not have simple structure and that increases the difficulty of interpretation.
    - The eigenvectors can be rotated so that simple structure is obtained.
    - The types of rotations are best distinguished are in terms of whether they are orthogonal (uncorrelated) or obliqne (correlated).
      - Varimax
      - Quartimax

Follow-up Analyses

- Hypothesis testing techniques (Direct gradient analysis)
- Confirmatory Classification (either direction – SAHN or K-Means)
- Construction of confidence intervals.