Multivariate Data Analysis

Lindsey Gao, mentored by Sarah Teichman

Background & Motivation

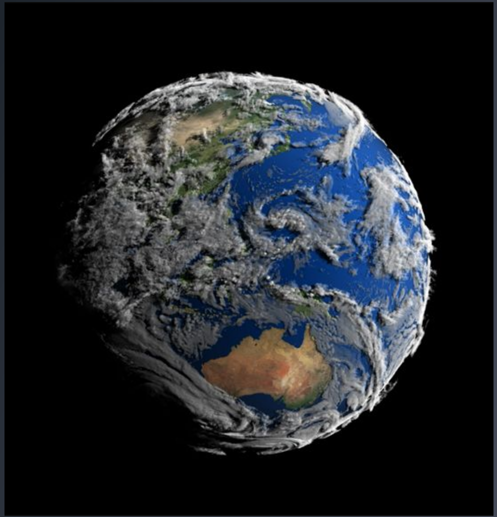
- Multivariate data: large data matrix => difficult to interpret
- Techniques for extracting what large data sets show us
- Simultaneous statistical analysis of a collection of variables, by using information about the relationships between the variables.
- Analysis of each variable separately is likely to miss key features and interesting patterns in the multivariate data

Chosen Dataset

-Retrieved NASA's Earth Science Data: The Compendium of Environmental Sustainability Indicator Collections

- 426 environmental sustainability indicators for 239 countries from 5 major data collection efforts between 2004-2006

-Collection compiled and distributed by the Columbia University Center for International Earth Science Information Network



Overview

1. Principal Component Analysis (PCA)
2. Multidimensional Scaling (MDS)
   a. Classical Multidimensional Scaling (cMDS)
   b. Non-Metric Multidimensional Scaling (nMDS)
3. Exploratory Factor Analysis (EFA)
4. Confirmatory Factor Analysis (CFA)
5. Cluster Analysis
   a. K-Means Clustering
   b. Model-Based Clustering

1.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- Goal: Reduce the dimensionality of data set while accounting for as much of the original variation as possible
- Transform original variables: $x^T = (x_1, \ldots, x_q)$ to new set of uncorrelated variables (principal components): $y^T = (y_1, \ldots, y_q)$ where

$$
\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1q}x_q \\
y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2q}x_q \\
&\ldots \\
y_q &= a_{q1}x_1 + a_{q2}x_2 + \ldots + a_{qq}x_q
\end{aligned}
$$

- Covariance/ Correlation Matrix S:

$$
\begin{aligned}
S &= A\Lambda A^T \\
A &= [\vec{a_1}, \vec{a_2}, \ldots, \vec{a_q}]
\end{aligned}
$$

- Transform original data points in terms of eigenvectors that capture most of the variance $(\vec{a_1}, \vec{a_2})$ and plot on new axes with principal component 1 on x, and principal component 2 on y

2.

Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS)

- Class of methods with similar goals as PCA to produce low dimensional visualizations of data
    - Operates on distance matrices instead of data matrix
    - Goal: Find a set of points in low dimension that approximate the high dimensional distance matrix

Classical MDS

- Inner product matrix of data:

$$\mathrm{B} = \mathrm{XX}^{\mathrm{T}}$$

- Find B in terms of the of distances
- SVD matrix B => Coordinate axes are the 1st k eigenvectors multiplied scaled by corresponding eigenvalues

Non-Metric MDS

- Uses the rank order of the distances
- Find disparities $\hat{d_{ij}}$

$$s.t.\ d_{ij} = \hat{d}_{ij} + \epsilon_{ij}$$

- Minimize stress function:

$$\mathrm{S}(\hat{X}) = \min\left(\frac{\sum_{i<j}\left(\hat{d}_{ij} - d_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}\right)$$

3.

Exploratory Factor Analysis

Exploratory Factor Analysis

- Factor analysis: method used to uncover the relationship between assumed latent variables (factors) and manifest variables
- EFA: used to investigate the relationship between manifest variables & factors without making assumptions about which manifest variables relate to which factors
- Assume we have a set of observed/manifest variables: $x^T = (x_1, x_2, \ldots, x_q)$ linked to k factors: $f_1, \ldots, f_k \; s.t. \; k < q$ by a regression model:

$$x_1 = \lambda_{11} f_1 + \ldots + \lambda_{1k} f_k + u_1$$
$$\ldots$$
$$x_q = \lambda_{q1} f_1 + \ldots + \lambda_{qk} f_k + u_q$$

Matrix Notation:

$$x = \Lambda f + u$$

Random disturbance terms u_i **specific to** x_i & uncorrelated with each other & factors

4.

Confirmatory Factor Analysis (CFA)

Confirmatory Factor Analysis

- Postulate a specific factor model on data where you hypothesize that particular manifest variables are allowed to relate to particular factors while other manifest variables are constrained to have 0 loadings on some factors
- Usually perform EFA on part of data to form hypothesis & CFA on other portion to test hypothesis
  - *CFA must be performed on new data not used in EFA*
- CFA model parameters: covariances/variances of residuals & latent variables $\theta = (\theta_1, \ldots, \theta_t)^T$
  - Determines covariance matrix implied by the model: $\Sigma(\theta)$
- Estimate parameters by minimizing discrepancy function
  - Ordinary least squares: $\mathrm{FLS}(\mathrm{S}, \Sigma(\theta)) = \sum_{i<j} \sum_j (s_{ij} - \sigma_{ij}(\theta))^2$
  - Maximum likelihood*: $\mathrm{FML}(\mathrm{S}, \Sigma(\theta)) = \log(|\Sigma(\theta)|) - \log|S| + \mathrm{trace}(\mathrm{S}\Sigma(\theta)^{-1}) - q$

5.

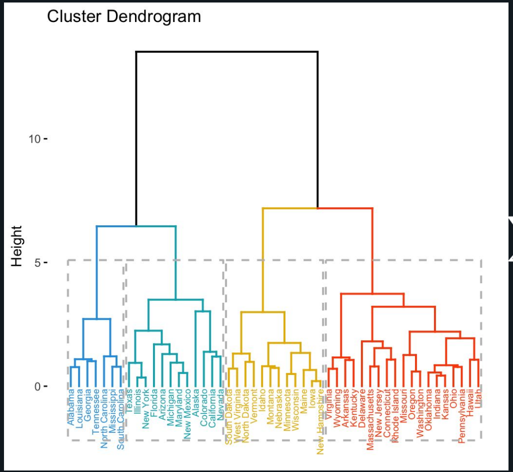
Cluster Analysis

Cluster Analysis

- Cluster analysis: generic term for many numerical techniques with the goal of uncovering groups of observations that are homogeneous & separated from other groups

Agglomerative hierarchical clustering

(AGC)



K-Means

Goal: find partition of n individuals into k groups that minimizes the within group sum of squares (WGSS)

$$\sum_{j=1}^{q}\sum_{l=1}^{k}\sum_{i\in G_l}\left(x_{ij}-\bar{x}^{l}\right)^{2}$$

Model-Based Clustering

Postulate a formal statistical model on population => results in overall population with finite mixture density => use maximum likelihood estimation to estimate parameters in finite mixture model

Credits

Special thanks to Sarah for mentoring me on this project!

- Textbook: An Introduction to Applied Multivariate Analysis with R
- Data: https://sedac.ciesin.columbia.edu/data/set/cesic-complete-collection-v1-1
- Shiny App: https://lindseygao.shinyapps.io/exploringmvdata/