Principal Component Analysis
Factor Analysis

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Environmental Mixtures Workshop

08.23.2018

Outline

Introduction

- While clustering aims to identify distinct subgroups based on exposure experience
- Factor analytic techniques aim to describe variability among observed, correlated chemicals in terms of a potentially lower number of unobserved factors

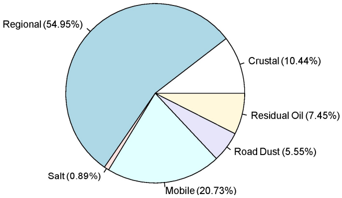
- These two unsupervised approaches are actually connected, but we will not get into this here
  - If interested in this please read: Ding C, He X. "K-means clustering via principal component analysis." ICLM 2004

Pattern Recognition & Dimensionality Reduction

- A potential topic of interest in assessing exposure to mixtures is to identify patterns of exposure
- This can be either specific *sources* of exposure
- Or *common behaviors* in the study population
- These patterns, to be meaningful, tend to have a *lower dimension* than the original data matrix, i.e. $k < p$
- The most common example arises in air pollution
    - Source apportionment

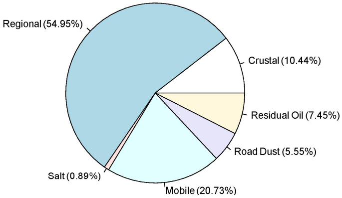
Source Apportionment

- Use factor analytic techniques to apportion $PM_{2.5}$ to its sources
  - E.g. Boston 2003 – 2010
  - BC and 17 other $PM_{2.5}$ XRF-measured components



Source Apportionment

- Use factor analytic techniques to apportion $PM_{2.5}$ to its sources
  - E.g. Boston 2003 – 2010
  - BC and 17 other $PM_{2.5}$ XRF-measured components



- Use $PM_{2.5}$ sources as the exposure of interest in the health models
  - Identification of most harmful emission sources
  - More effective air quality management and regulations

Pattern Recognition & Dimensionality Reduction (cont'd)

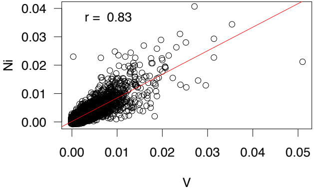
- There are many approaches for pattern recognition and dimensionality reduction
- Most common applications are in Computer Science and Machine Learning (e.g. face recognition etc)
- And many are being developed
- Here we will focus on the two most "traditional"
    1. Principal Component Analysis (PCA)
    2. Factor Analysis (FA)

What is PCA?

- Dimensionality reduction statistical tool
- Aims to explain as much of the total variance in the data as possible using a *smaller* number of variables
- The resulting variables (i.e. `components`) are linear combinations of the original variables
- Reducing the dimension of the data will somewhat reduce accuracy – tradeoff

2-D Example

- Ni and V are associated with residual oil combustion
- Commonly very highly correlated



- The above points are presented in two axes
- PCA will allow us to present the data along one axis (the red in this case), called *principal component*
- In reality, we don't use PCA for 2-D problems

How does PCA Work?

PCA aims to:

1. Identify a sequence of linear combinations of the p variables
2. That have maximal variance
3. And are mutually uncorrelated
   - I.e. orthogonal solution

Eigenvalues & Eigenvectors

$$A\nu = \lambda\nu$$

where A is the $p \times p$ var-covar or correlation matrix of X, ν is the eigenvector ($p \times 1$) and λ is a scalar (eigenvalue)

- Every eigenvector has an eigenvalue (p pairs)
- The eigenvector shows the direction (i.e. in the 2-D example above the direction of the red line)
- The eigenvalue is a scalar saying how much variance exists in the data along that direction
- The eigenvector with the highest eigenvalue is the first principal component

Singular Value Decomposition (SVD)

- SVD is a generalization of the eigen decomposition that also works on non-square matrices
  - I.e. on the data matrix directly
- Preferable to standardize (center and scale) variables
  - Especially if chemicals are on different scales and/or units
  - Most software has options to automatically do this
- Both SVD and eigen decomposition yield orthogonal solutions
  - I.e. the estimated components are not correlated

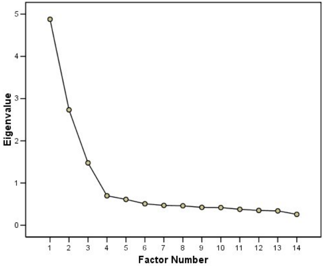
PCA – Results

$$
\begin{aligned}
z_{i1} &= \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip} \\
z_{i2} &= \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{p2}x_{ip} \\
&\vdots \\
z_{ip} &= \phi_{1p}x_{i1} + \phi_{2p}x_{i2} + \cdots + \phi_{pp}x_{ip}
\end{aligned}
$$

- $i \in 1, \cdots, N$
- $z_{11}, \cdots, z_{N1}$ the scores on the first PC
- $\phi_1 = (\phi_{11}, \phi_{21}, \cdots, \phi_{p1})^T \rightarrow$ loadings of the first PC
- $\phi_1, \cdots, \phi_p \rightarrow$ eigenvectors
- $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

PCA – Dimensionality Reduction

- PCA will yield a new data matrix of the same dimensions ($N \times p$)
- We need to choose how many components to keep for further analysis
- Look for "elbow" at scree plot



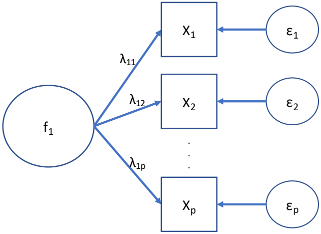
- Decide how much of the total variance explained we'd like to keep

Factor Analysis (FA)

- FA assumes that the measurements of the p pollutants arise from k underlying sources
  - That are not observed (i.e. latent)
  - With $k < p$, and k pre-specified
- Does not require orthogonality
- What we will discuss here is exploratory FA (EFA)
  - SEMs lie under confirmatory FA (CFA)

FA (cont'd)

- $\boldsymbol{x} = \boldsymbol{\mu} + \Lambda \boldsymbol{f} + \boldsymbol{\varepsilon}$, where
  - Λ a $p \times k$ matrix of factor loadings
  - $\boldsymbol{f}$ a matrix of k *common* sources of variation among X and accounts for their correlation structure
  - $\boldsymbol{\varepsilon}$ have mean zero and Ψ diagonal covariance matrix



FA (cont'd)

- To solve FA we can

1. Assume joint normality and use maximum likelihood
   - Ends up being a hard optimization problem
   - With often many local minima
2. SVD of the data matrix
   - If we assume that instead of diagonal Ψ the error matrix is Ξ and small (in sums of squares of its elements)
   - Now just a question of minimizing sums of squares $\rightarrow$ reduces to matrix decompositions

- Rotation to yield more interpretable factors (potentially correlated)
  - If the factor correlations are not driven by the data the solution will remain nearly orthogonal

PCA vs. FA

- PCA will always yield orthogonal solutions
- FA does not have to (orthogonal vs. oblique solutions)

PCA vs. FA

- PCA will always yield orthogonal solutions
- FA does not have to (orthogonal vs. oblique solutions)

- PCA aims to explain the total variance in the data
- FA aims to identify common sources of variation

PCA vs. FA

- PCA will always yield orthogonal solutions
- FA does not have to (orthogonal vs. oblique solutions)

- PCA aims to explain the total variance in the data
- FA aims to identify common sources of variation

- In PCA all p components are provided and the user selects k afterwards
- In FA k needs to be pre-specified

PCA vs. FA

- PCA will always yield orthogonal solutions
- FA does not have to (orthogonal vs. oblique solutions)

- PCA aims to explain the total variance in the data
- FA aims to identify common sources of variation

- In PCA all p components are provided and the user selects k afterwards
- In FA k needs to be pre-specified

- Usually very similar solutions

PCA vs. FA

- For both PCA & FA: no single correct answer
- And scores and factors are centered at zero $\cdots$
- Interpretability is key (and an issue)

PCA and FA in Health Models

- When the components/factors have been identified the resulting scores or source contributions can be included as the exposure(s) of interest in health models
- The PCA solution is orthogonal – no need to worry about co-component confounding
- This might not be the case with FA
- These are now continuous $\rightarrow$ non-linear exposure-response functions can also be explored
- Uncertainty propagation?
- Supervised extensions for PCA exist

Thank you!

Questions?

mk3961@cumc.columbia.edu