Module 10
Principal Components and Factor Analysis

This module begins a discussion of a suite of methods related to classification of results in multivariate databases.

There are three of these major methods in total, of which we will deal in detail with two in this course:

- 1. Principal Components/Factor Analysis
- 2. Cluster Analysis
- 3. Multidimensional Scaling

Multidimensional Scaling

- First, a brief word on multidimensional scaling (MDS), even though this isn’t a main topic for our course:
  - **Definition**: show enough so you know what it is
- **MDS** deals explicitly with spatial problems
- Normally, we take a map and then construct a **distance matrix** from it

Multivariate Classification Methods

- **This module begins a discussion of a suite of methods related to classification of results in multivariate databases.**
- **There are three of these major methods in total, of which we will deal in detail with two in this course.**
  - 1. Principal Components/Factor Analysis
  - 2. Cluster Analysis
  - 3. Multidimensional Scaling
Multidimensional Scaling

- MDS does the reverse: takes a distance matrix and constructs a map out of it.

Possible inputs to MDS (distance matrix):
- Real, linear (geographic) distances
- Some other kind of “proximity” (not necessarily geographic)

Proximities: concept developed by psychologists to measure preferences or perceptions.

Example:
- Ask people to rate country pairs on a scale from 1 (different) to 9 (similar)

Rate 4 Countries, so 6 Country Pairs

<table>
<thead>
<tr>
<th>Countries</th>
<th>Country Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>USA-Canada</td>
</tr>
<tr>
<td>Canada</td>
<td>USA-Mexico</td>
</tr>
<tr>
<td>Mexico</td>
<td>USA-Spain</td>
</tr>
<tr>
<td>Spain</td>
<td>Canada-Mexico</td>
</tr>
<tr>
<td></td>
<td>Canada-Spain</td>
</tr>
<tr>
<td></td>
<td>Mexico-Spain</td>
</tr>
</tbody>
</table>

Survey people on their perceptions (country pair similarities), put the results of this survey into a “proximity matrix”.

<table>
<thead>
<tr>
<th></th>
<th>USA</th>
<th>Canada</th>
<th>Mexico</th>
<th>Spain</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>0</td>
<td>8.2</td>
<td>4.5</td>
<td>5.6</td>
</tr>
<tr>
<td>Canada</td>
<td>8.2</td>
<td>0</td>
<td>3.6</td>
<td>4.4</td>
</tr>
<tr>
<td>Mexico</td>
<td>4.5</td>
<td>3.6</td>
<td>0</td>
<td>4.8</td>
</tr>
<tr>
<td>Spain</td>
<td>5.6</td>
<td>4.4</td>
<td>4.8</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: this particular example uses a measure that is interpreted the opposite of a geographic distance (close similarity here = higher values).

Put the proximity matrix into the MDS routines of SPSS and get a “map” of people’s country pair perceptions.

“Cultural Space”

Map Output

Example: MDS of United Kingdom Members of Parliament, based on Twitter follower data.
Multidimensional Scaling

- **Bottom line:** MDS is a powerful, exploratory technique with many possible applications, some of which are of much interest to geographic research.

Principal Components/Factor Analysis

- **The basics on PC/FA**
  - Principal components analysis and factor analysis comprise yet another flexible exploratory multivariate tool.
  - We will begin by calling the whole thing "PC/FA" (many similarities between the two).
  - In a few minutes we will deal with the difference (relatively minor).
Principal Components/Factor Analysis

- The basics on PC/FA
  - Goal of PC/FA: simplify a complex situation without losing too much explanatory power
  - Many variables, many observations: what on earth is going on? This situation is too complex.
  - By its very nature, PC/FA is a multivariate technique

Many variables, many observations: what on earth is going on? This situation is too complex.

By its very nature, PC/FA is a multivariate technique.

Principal Components/Factor Analysis

Sample Database Format Analyzed by PC/FA

Human Example

<table>
<thead>
<tr>
<th>Income</th>
<th>Education</th>
<th>Marital Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
</tbody>
</table>

Observations:
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data

Variables:
- Income
- Education
- Marital Status

Observations:
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data
- data

In general, there are three approaches to the analysis of this kind of database

1. Use independent variables to generate a relationship with a dependent variable (multiple regression): hypothesis testing by examining possible predictive models.
2. Grouping of rows ("Q mode" PC/FA): hypothesis generation by examining similarities among observations.
3. Grouping of columns ("R mode" PC/FA): hypothesis generation by examining similar variables.

Principal Components/Factor Analysis

Sample Database Format Analyzed by PC/FA

Physical Example

Variables

<table>
<thead>
<tr>
<th>Diseases</th>
<th>Moisture Content</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>data</td>
<td>data</td>
<td>data</td>
</tr>
</tbody>
</table>

Principal Components/Factor Analysis

Example: counties and farm products

Output by County

<table>
<thead>
<tr>
<th>County</th>
<th>Milk</th>
<th>Corn</th>
<th>Hay</th>
<th>Beef</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>25</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>9</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

Principal Components/Factor Analysis

- Approach #1 basically analyzes the database as it is.
- Approaches #2 and #3 focus on simplifying the situation in some way.
  - Goal: more understanding
- Factor analysis normally takes place in "R mode" (group variables or columns).
  - Rule: assume R mode unless told otherwise.
Principal Components/Factor Analysis

Potential R mode groupings for this dataset

<table>
<thead>
<tr>
<th>County</th>
<th>Milk</th>
<th>Corn</th>
<th>Hay</th>
<th>Beef</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>25</td>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>9</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

Milk } “Dairying” Corn } “Livestock” Beef }

Helpful to name the new factors if you can

Principal Components/Factor Analysis

Now we can replace the original data table with a simpler version:

<table>
<thead>
<tr>
<th>County</th>
<th>Livestock</th>
<th>Dairying</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>2</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>3</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>4</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>5</td>
<td>data</td>
<td>data</td>
</tr>
<tr>
<td>6</td>
<td>data</td>
<td>data</td>
</tr>
</tbody>
</table>

An advance (easier to interpret)

Principal Components/Factor Analysis

Group the counties (observations):

<table>
<thead>
<tr>
<th>County</th>
<th>Livestock</th>
<th>Dairying</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

“Livestock” “Dairying”

Alternative: Q mode groupings

1 and 6: Dairying counties
2 and 3: Livestock counties
4 and 5: Mixed (more balanced than the others)

One aside: where do these “combined” data fields come from?

We’ll conclude this evening with a brief mention of something called “factor scores”

Principal Components/Factor Analysis

Key question: how is this grouping done?
- Use correlation coefficients to determine degree of similarity (analyze a correlation matrix – more on this later)

Another key question: how many variables are appropriate for this method, versus multiple regression?
- Multiple regression: up to 10 variables
- Factor analysis: 10 or more

Example: Murdie’s classic analysis of census data for the city of Toronto
- 56 variables for 277 census tracts
- This is far too complex (specifically, way too many variables) for multiple regression
**Principal Components/Factor Analysis**

- **Principal Components versus Factor Analysis: The Difference**
  - Fundamentally the same
  - Both analyze correlation matrices the same way
  - Difference is in the interpretation
- **Principal Components: a closed analysis**
  - All variation is accounted for by the variables themselves
  - No outside influences

- **Variance of a variable**
  - Variance in common with other variables
  - Unique variance of that variable
  - Total Variance of the Variable

- **Factor Analysis**
  - Total Variance of the Factors
  - Unique Variance (unexplained)
  - Factor 1 Variance
  - Factor 2 Variance
  - Total Variance of the Database

- **Principal Components Analysis**
  - Total Variance of the Components (Equal to the Total Variance of Database)
  - Component 1 Variance
  - Component 2 Variance
  - Total Variance of the Database

- **With principal components analysis, keep adding components until you account for all of the variance in the database**
  - Factor analysis: limit the number of factors you use (use only a few)
- **Bottom line: the difference between principal components and factor analysis is not in the mathematics, but in the implementation and interpretation**
Given the understanding we’ve now developed, let’s come back to the purpose of PC/FA

- **Basic questions answered by PC/FA**
  1. What are the patterns of variable relationships represented in the correlation matrix?
  2. Can a given correlation matrix be parsimoniously described? (eliminate redundancy in columns/rows)
  3. Are certain “dimensions” latent in a given correlation matrix?

It is the “latent dimensions” idea that forms the basis for the next part of our discussion (this is also the one thing I mentioned about PC/FA in our very first class meeting)

- From here on in, we will start talking about “factor analysis” only, even though most of the concepts apply to principal components analysis and factor analysis equally

---

**Factor Analysis: A Geometric Approach**

- **Situation**: you are in a car going northeast at 100 miles per hour (you obviously have a radar detector)
  - **Q**: How fast are you going east?

  ![Diagram](https://via.placeholder.com/150)

  **Answer**: 
  \[
  v_{east} = 100 \text{ mph} \times \cos(45^\circ) \\
  = 100 \text{ mph} \times 0.7071 \\
  = 70.71 \text{ mph}
  \]

  If \( h=100 \), \( a = 100 \times \cos \theta \) (\( a \) being the “east component”)

---

**Factor Analysis: A Geometric Approach**

- **Background from high school geometry**: 
  \[
  \cos \theta = \frac{a}{h}
  \]
  So, \( a = h \times \cos \theta \)

---

**Factor Analysis: A Geometric Approach**

- **Back to correlation**: 
  - Range of correlation values: -1.0 to +1.0
  - Range of cosine values: -1.0 to +1.0
  - **Same range**: so how about interpreting correlations as cosines of angles?
    - If a correlation = +1.0, this means that \( \cos(\text{angle}) = +1.0 \), so angle=0° (cos 0° = +1.0)
    - If a correlation = 0.0, this means that \( \cos(\text{angle}) = 0.0 \), so angle=90° (cos 90° = 0.0)

---

**Factor Analysis: A Geometric Approach**

- This cosine-correlation connection idea leads to a geometric visualization of correlations (variable relationships)
  - **Bottom line**: it is helpful to interpret correlation as an angle (an “angle” between the two variables in the correlation)
Factor Analysis: A Geometric Approach

- Correlation interpreted as being connected to an angle:

<table>
<thead>
<tr>
<th>θ</th>
<th>cos θ (correlation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>1.00</td>
</tr>
<tr>
<td>45°</td>
<td>0.70</td>
</tr>
<tr>
<td>90°</td>
<td>0.00</td>
</tr>
<tr>
<td>135°</td>
<td>-0.70</td>
</tr>
<tr>
<td>180°</td>
<td>-1.00</td>
</tr>
</tbody>
</table>

- So, we're saying that a correlation of 0.70 could be conceptualized as an angle of 45° between the two variables.

- Further discussion of this representation:

- Perfect correlation between X_2 and X_3 gives an angle of 0°.
- Perfect inverse correlation gives an angle of 180°.

- For example: FA on a 4x4 correlation matrix

<table>
<thead>
<tr>
<th></th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_1</td>
<td>1.00</td>
<td>0.866</td>
<td>0.385</td>
<td>0.230</td>
</tr>
<tr>
<td>X_2</td>
<td>1.00</td>
<td>0.866</td>
<td>0.530</td>
<td>0.300</td>
</tr>
<tr>
<td>X_3</td>
<td>1.00</td>
<td>0.866</td>
<td>0.530</td>
<td>1.000</td>
</tr>
<tr>
<td>X_4</td>
<td>1.00</td>
<td>0.866</td>
<td>1.000</td>
<td>0.230</td>
</tr>
</tbody>
</table>

- Convert to a matrix of angles:

<table>
<thead>
<tr>
<th></th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_1</td>
<td>0°</td>
<td>30°</td>
<td>60°</td>
<td>90°</td>
</tr>
<tr>
<td>X_2</td>
<td>0°</td>
<td>30°</td>
<td>60°</td>
<td>90°</td>
</tr>
<tr>
<td>X_3</td>
<td>0°</td>
<td>30°</td>
<td>60°</td>
<td>0°</td>
</tr>
<tr>
<td>X_4</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
<td>0°</td>
</tr>
</tbody>
</table>

- This means graphically:

Remember, the objective of factor analysis is to simplify this situation by substituting a smaller number of factors for a larger number of variables.

For example, we might simplify this situation by replacing the four variables with two factors:
- one where X_1 is (call it T_1)
- another where X_4 is (T_2)
Factor Analysis: A Geometric Approach

- **Result:** a new factor matrix (variable-factor link):

<table>
<thead>
<tr>
<th>Variables</th>
<th>$T_1$</th>
<th>$T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.000</td>
<td>0.500</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.500</td>
<td>0.000</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.500</td>
<td>0.000</td>
</tr>
<tr>
<td>$X_4$</td>
<td>0.866</td>
<td>0.500</td>
</tr>
</tbody>
</table>

- **Factors:**

  - $T_1$: 30°
  - $T_2$: 60°
  - $T_3$: 90°

Factor loadings show how well each variable relates to each factor.

- **Challenge:** in creating a small number of factors in place of a large number of variables:
  - Position the factors (e.g., $T_1$ and $T_2$) optimally so they represent as well as possible the variable vectors being replaced.

- **Observation:** an infinite number of possible “factor configurations” exist:
  - There may be no “right” configuration: how to decide which one is best for your situation?

The angle visualization idea gives us a number of “tried and true” methods for determining factors:

- **1. Centroid method**
  - A straightforward method
  - Relatively clear, so it is helpful for introducing the idea of calculating factors geometrically
  - Basic idea: calculate an “average” of all factors to create an initial “most powerful” factor.

- **Methods of positioning or “rotating” the factors we are creating so they best fit the variable vectors in our datasets**

A second step would be to create another factor that is orthogonal to the first factors.

- By being orthogonal, this second factor is intended to capture something completely unique in comparison to the first.
Factor Analysis: A Geometric Approach

1. **Centroid method**
   - There are better ways to create factors, so this method is little used today. However, this method is simple in concept, so it illustrates well the basics of how factor creation works.

2. **Varimax method**
   - Emphasizes column simplification in the factor matrix
   - Meaning: each factor links highly with only a few variables
   - Advantage: easier interpretation

3. **Quartimax method**
   - Emphasizes row simplification in the factor matrix
   - Meaning: variables load highly onto one and only one factor
   - Often leads to one prominent, general factor along with many smaller factors that have only a few variables each

4. **Equimax method**
   - Compromise between #2 and #3: simultaneous row and column simplification

SPSS gives a number of factor analysis options (just need to select the one you want)

---

Evaluation of Results: What is Good?

1. Each row of your factor matrix has at least one "factor loading" close to zero
   - Q: what would a "factor loading of zero" mean?

---

**Sample Factor Matrix**

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>X2</td>
<td>0.866</td>
<td>0.500</td>
</tr>
<tr>
<td>X3</td>
<td>0.500</td>
<td>0.866</td>
</tr>
<tr>
<td>X4</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Two Factors

---

**Sample Factor Matrix**

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>X2</td>
<td>0.866</td>
<td>0.500</td>
</tr>
<tr>
<td>X3</td>
<td>0.500</td>
<td>0.866</td>
</tr>
<tr>
<td>X4</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Four Variables
Factor Analysis: A Geometric Approach

Evaluation of Results: What is Good?

- 2. For every pair of columns in the factor matrix, there should be several variables with high loadings on one column (factor) and low loadings on the other.
  - Q: what does that mean?

<table>
<thead>
<tr>
<th>X1</th>
<th>T1</th>
<th>T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>0.500</td>
<td>0.500</td>
<td></td>
</tr>
<tr>
<td>0.306</td>
<td>0.952</td>
<td></td>
</tr>
<tr>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Evaluation of Results: What is Good?

- 3. Only a small number of column pairs have high loadings for the same variable
  - Q: what does this mean?

<table>
<thead>
<tr>
<th>X1</th>
<th>T1</th>
<th>T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>0.500</td>
<td>0.500</td>
<td></td>
</tr>
<tr>
<td>-0.707</td>
<td>0.707</td>
<td></td>
</tr>
<tr>
<td>0.000</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Factor Analysis: A Geometric Approach

Interpretation of Results: How Many Factors Are Appropriate?

- A factor analysis of a dataset will often generate many factors – more than should be used.
  - Some strong factors that are worthwhile, and are helpful in describing and accounting for the key characteristics of the dataset
  - Some weak factors that don’t add much to your modeling of the dataset
  - How many factors should we take from a factor analysis?

To give a good basis for determining the number of factors to use, we need one more concept: the eigenvalue

- The term “eigenvalue” originates from the matrix models that are used to conceptualize and run the factor analysis calculations
  - In other words, the “eigenvalue” term has a meaning that goes beyond the realm of factor analysis, into pure matrix algebra

So what is an eigenvalue in factor analysis?

- Let’s go back to our geometric interpretation of factors.
  - Say we have created one factor that takes the place of two variables.
Factor Analysis: A Geometric Approach

■ Here is the factor matrix for this situation:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>0.70</td>
</tr>
<tr>
<td>X2</td>
<td>0.70</td>
</tr>
</tbody>
</table>

The matrix shows that each of the variables has a “loading onto” the factor of 0.70: a decent relationship between the factor and the variables.

Geometrically, this situation could be represented like this:

Q: what are the components of X1 and X2 that are “in the direction of” T1?

Subdivide the two variables into components:

Each variable has a component parallel to T1, and a component perpendicular to T1.

We will focus on the parallel components here.

Key idea #1: The parallel components are simply the factor loadings of X1 and X2 onto T1.

Eigenvalue $T_1 = X_{1Par} + X_{2Par}$

In our example, the eigenvalue of T1 is: $0.7 + 0.7 = 1.4$

Subdivide the two variables into components:

Key idea #2: the eigenvalue of T1 is simply the sum of the loadings of X1 and X2 onto T1.

Subdivide the two variables into components:

Key idea #3: the eigenvalue of T1 can be interpreted as the length (strength) of the factor.
Factor Analysis: A Geometric Approach

Subdivide the two variables into components

Key idea #4: each variable is conceived as having a "length" of 1, so knowing the "length" of each factor makes it possible to do some comparisons.

How long (powerful) is each factor compared to an individual variable?

Remember, factors are combinations of variables, so good ones should really be longer than 1.

Factor Analysis: A Geometric Approach

So, this is the basic interpretation of the "eigenvalue", in the context of factor analysis:

- Eigenvalue: a measure of the strength of each factor
  - A factor with an eigenvalue of 1 or less: not very powerful (we might as well keep using individual variables rather than weak factors like this)
  - A factor with an eigenvalue greater than 1: now we're talking, but of course, the bigger the better
  - Maximum eigenvalue = number of variables in the database (all variables load perfectly on one factor)

Back to our original purpose: how many factors should we use?

- The eigenvalue concept suggests an obvious rule: eliminate all factors with eigenvalues of 1 or less (weak)

Another idea: draw a scree plot of eigenvalues

Q: What conclusion should we draw from this graph?

Back to our original purpose: how many factors should we use?

- Lastly, you might also consider another common "rule of thumb", implemented from the FA "Total Variance Explained" table produced by SPSS:
  - Eliminate all factors with eigenvalues explaining less than a given cutoff (often 5%) of the total variance of the database
  - We will see this total variance table in the Factor Analysis lab we will do next time

One last issue: FA and geography

- Remember what we've done: FA gives us a nice summary of our complex dataset that derives a small number of factors from a large number of variables
  - Issue: it is great to have this "factor overview" of our database, but what can geographers do with the factors once we've generated them?
Factor Analysis: A Geometric Approach

- **One last issue: FA and geography**

  - **Concept:** “factor scores” (see pages 625-628 of your Field reading for details)
  - A factor score gives you the ability to determine how strongly each factor is operating in each observation in your database
  - We don’t have time to go into the details of factor score calculation here, but I want to at least tell you what we as geographers can do with these scores

- **Quick example:** you have a database with data for 1000 census tracts, and each tract has 50 variables (age, income, education, etc.)

  - Doing an FA on this database, let’s say you find 3 prominent factors that represent what is going on in this database
    - 1. a “social status” factor
    - 2. an “economic status” factor
    - 3. a “mobility” factor

- **Factor score application:** you can calculate a factor score for each factor in each census tract

  - This will give you a picture of what is happening with each of your census tracts, addressing questions like:
    - To what degree is “social status” the key factor at work in a given census tract?
    - In which census tracts is the “economic status” factor important? [idea: you could map this out]

Sample Factor Score Map: Economic Status

Sample Factor Score Map: Highest Score by Tract

Factor Analysis: A Geometric Approach

- **One last issue: FA and geography**

  - **Quick example:** you have a database with data for 1000 census tracts, and each tract has 50 variables (age, income, education, etc.)

  - Doing an FA on this database, let’s say you find 3 prominent factors that represent what is going on in this database
    - 1. a “social status” factor
    - 2. an “economic status” factor
    - 3. a “mobility” factor

- **Factor score application:** you can calculate a factor score for each factor in each census tract

  - This will give you a picture of what is happening with each of your census tracts, addressing questions like:
    - To what degree is “social status” the key factor at work in a given census tract?
    - In which census tracts is the “economic status” factor important? [idea: you could map this out]

Sample Factor Score Map: Economic Status

Sample Factor Score Map: Highest Score by Tract

Factor Analysis: A Geometric Approach

- **One last issue: FA and geography**

  - You will need to consult the Field reading on this topic for details on how to actually do this additional analysis, but factor scores can be a very powerful tool you should be aware of
    - **Note:** the Field reading explains this factor score concept by referring to a psychological example
    - Rather than a database of values by census tract, this example deals with a database of values by people taking part in a study (but this is the same basic idea as what we’re discussing with census tracts)