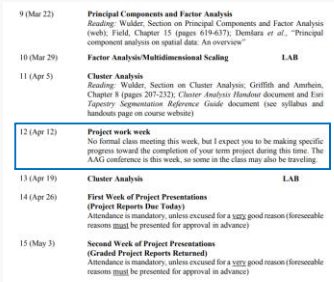
**Brief look at the rest of the semester**

Note that there is no class meeting the week of <u>April 12</u>: please use this as a project work week



| | | |
|---|---|---|
| 9 (Mar 22) | **Principal Components and Factor Analysis**<br>*Reading:* Wulder, Section on Principal Components and Factor Analysis (web); Field, Chapter 15 (pages 619-637); Demšara *et al.*, "Principal component analysis on spatial data: An overview" | |
| 10 (Mar 29) | **Factor Analysis/Multidimensional Scaling** | **LAB** |
| 11 (Apr 5) | **Cluster Analysis**<br>*Reading:* Wulder, Section on Cluster Analysis; Griffith and Amrhein, Chapter 8 (pages 207-232); *Cluster Analysis Handout* document and Esri *Tapestry Segmentation Reference Guide* document (see syllabus and handouts page on course website) | |
| 12 (Apr 12) | **Project work week**<br>No formal class meeting this week, but I expect you to be making specific progress toward the completion of your term project during this time. The AAG conference is this week, so some in the class may also be traveling. | |
| 13 (Apr 19) | **Cluster Analysis** | **LAB** |
| 14 (Apr 26) | **First Week of Project Presentations**<br>**(Project Reports Due Today)**<br>Attendance is mandatory, unless excused for a <u>very</u> good reason (foreseeable reasons <u>must</u> be presented for approval in advance) | |
| 15 (May 3) | **Second Week of Project Presentations**<br>**(Graded Project Reports Returned)**<br>Attendance is mandatory, unless excused for a <u>very</u> good reason (foreseeable reasons <u>must</u> be presented for approval in advance) | |

# MODULE 10

Principal Components and Factor Analysis

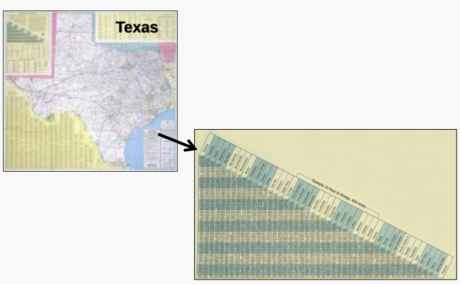
## Multivariate Classification Methods

- This module begins a discussion of a suite of methods related to <u>classification of results in multivariate databases</u>
- 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*

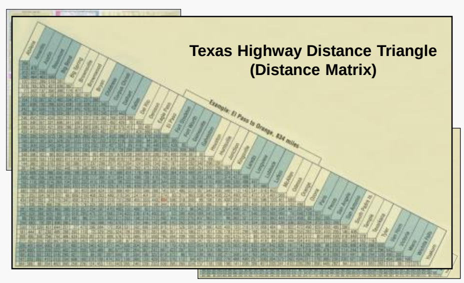
## Multidimensional Scaling

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

## Multidimensional Scaling

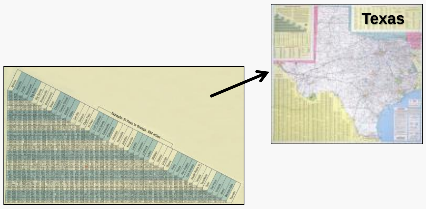



## Multidimensional Scaling



## Multidimensional Scaling

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




## Multidimensional Scaling

- 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

## Multidimensional Scaling

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

**Rate 4 Countries, so 6 Country Pairs**

| Countries | Country Pairs |
|---|---|
| USA | USA-Canada |
| Canada | USA-Mexico |
| Mexico | USA-Spain |
| Spain | Canada-Mexico |
| | Canada-Spain |
| | Mexico-Spain |

## Multidimensional Scaling

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

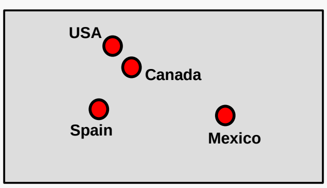
| | USA | Canada | Mexico | Spain |
|---|---|---|---|---|
| USA | 0 | | | |
| Canada | 8.2 | 0 | | |
| Mexico | 4.5 | 3.6 | 0 | |
| Spain | 5.6 | 4.4 | 4.8 | 0 |

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

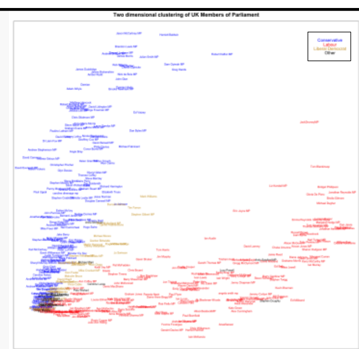
## Multidimensional Scaling

- Put the proximity matrix into the MDS routines of SPSS and get a "map" of people's country-pair perceptions

**"Cultural Space"**

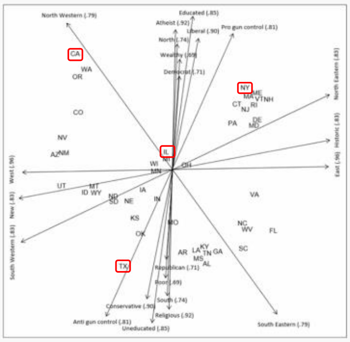



**Map Output Example:** MDS of United Kingdom Members of Parliament, based on Twitter follower data



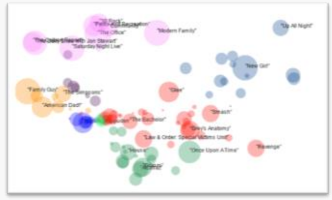
**Map Output Example:** MDS of people's perceptions of state similarities

Numbers on arrows are the result of follow-up survey-based opinion polling on respondents



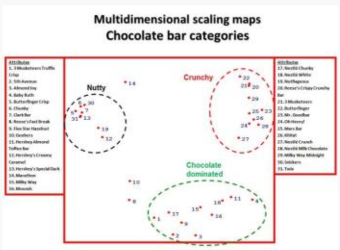
http://geoawesomeness.com/cognitive-map-of-the-united-states/

**Map Output Example:** MDS of "TV Show Similarity" from a Hulu Database



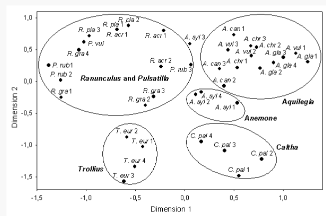
http://tech.hulu.com/blog/2012/04/

**Map Output Example:** MDS of Food Characteristics




https://www.quirks.com/articles/don-t-forget-the-graphic-multidimensional-scaling

**Map Output Example:** MDS of Pollen Odors, Demonstrating Patterns Related to Chemical Composition




http://www.pollination.de/SILENE_PROJECT/projects/pollen%20odors.html

## 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 technique*
  - *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*

## Principal Components/Factor Analysis

**Sample Database Format Analyzed by PC/FA**

**Human Example**

**Observations** (for census tracts, or perhaps observations for individual people)

| | Variables | | |
|---|---|---|---|
| | Income | Education | Marital Status |
| 1. | data | data | data |
| 2. | data | data | data |
| 3. | data | data | data |
| 4. | data | data | data |
| 5. | data | data | data |
| 6. | data | data | data |
| 7. | data | data | data |

## Principal Components/Factor Analysis

**Sample Database Format Analyzed by PC/FA**

**Physical Example**

**Observations** (by farm field, or perhaps observations for specific point locations)

| | Variables | | |
|---|---|---|---|
| | Biomass | Moisture Content | Temperature |
| 1. | data | data | data |
| 2. | data | data | data |
| 3. | data | data | data |
| 4. | data | data | data |
| 5. | data | data | data |
| 6. | data | data | data |
| 7. | data | data | data |

## Principal Components/Factor Analysis

- 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

- 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

- Example: counties and farm products

**Output by County**

| County | Milk | Corn | Hay | Beef |
|---|---|---|---|---|
| 1 | 4 | 0 | 4 | 0 |
| 2 | 3 | 9 | 2 | 10 |
| 3 | 6 | 25 | 6 | 30 |
| 4 | 4 | 9 | 5 | 6 |
| 5 | 1 | 2 | 1 | 2 |
| 6 | 9 | 1 | 10 | 0 |

## Principal Components/Factor Analysis

| County | Milk | Corn | Hay | Beef |
|---|---|---|---|---|
| 1 | 4 | 0 | 4 | 0 |
| 2 | 3 | 9 | 2 | 10 |
| 3 | 6 | 25 | 6 | 30 |
| 4 | 4 | 9 | 5 | 6 |
| 5 | 1 | 2 | 1 | 2 |
| 6 | 9 | 1 | 10 | 0 |

- Potential R mode groupings for this dataset

Milk
Hay } "Dairying"

Corn
Beef } "Livestock"

**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:

| County | Livestock | Dairying |
|---|---|---|
| 1 | data | data |
| 2 | data | data |
| 3 | data | data |
| 4 | data | data |
| 5 | data | data |
| 6 | data | data |

**An advance (easier to interpret)**

## Principal Components/Factor Analysis

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

**One aside:**
**where do these "combined" data fields come from?**

**We'll conclude this evening with a brief mention of something called "factor scores"**

| County | Livestock | Dairying |
|---|---|---|
| 1 | data | data |
| 2 | data | data |
| 3 | data | data |
| 4 | data | data |
| 5 | data | data |
| 6 | data | data |

## Principal Components/Factor Analysis

**Group the counties (observations)**

| County | Milk | Corn | Hay | Beef | |
|---|---|---|---|---|---|
| 1 | 4 | 0 | 4 | 0 | "Dairying" |
| 2 | 3 | 9 | 2 | 10 | "Livestock" |
| 3 | 6 | 25 | 6 | 30 | "Livestock" |
| 4 | 4 | 9 | 5 | 6 | "Mixed" |
| 5 | 1 | 2 | 1 | 2 | "Mixed" |
| 6 | 9 | 1 | 10 | 0 | "Dairying" |

- Alternative: Q mode groupings

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

## 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*

} Rough guideline

## Principal Components/Factor Analysis

- 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

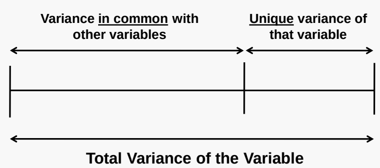
- **<u>Principal Components</u> versus <u>Factor Analysis</u>: The Difference**
  - *Fundamentally the same*
  - *Both analyze correlation matrices the same way*
  - *Difference is in the <u>interpretation</u>*
- Principal Components: a closed analysis
  - *All variation is accounted for by the variables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

## Principal Components/Factor Analysis

- **<u>Principal Components</u> versus <u>Factor Analysis</u>: The Difference**
  - *Fundamentally the same*
  - *Both analyze correlation matrices the same way*
  - *Difference is in the <u>interpretation</u>*
- Factor Analysis: an open analysis
  - *Allows for outside (unique) variance*
  - *Let's try to understand what this all means*

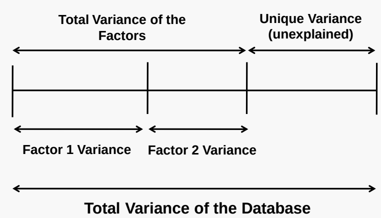
## Principal Components/Factor Analysis

- Variance of a variable



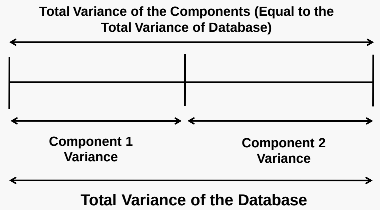

## Principal Components/Factor Analysis

- Factor Analysis




## Principal Components/Factor Analysis

- Principal Components Analysis




## Principal Components/Factor Analysis

- With principal components analysis, <u>keep adding components</u> until you account for all of the variance in the database
  - *<u>Factor analysis</u>: limit the number of factors you use (use only a few)*
- <u>Bottom line</u>: the difference between principal components and factor analysis is not in the mathematics, but in the implementation and interpretation

## Principal Components/Factor Analysis

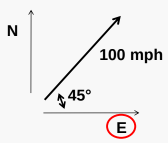
- Given the understanding we've now developed, let's come back to the purpose of PC/FA
  - ***Basic questions answered by PC/FA***
    - 1. What are the patterns of variable relationships represented in the correlation matrix?
    - 2. Can a given correlation matrix be parsimoniously described? (eliminate redundancy in columns/rows)
    - 3. Are certain "dimensions" latent in a given correlation matrix?

## Principal Components/Factor Analysis

- 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?*



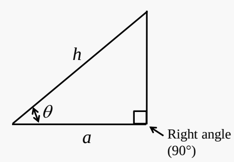

Answer:

$$V_{east} = 100mph \times \cos(45°)$$
$$= 100mph \times 0.7071$$
$$= 70.71mph$$

## Factor Analysis: A Geometric Approach

- Background from high school geometry:

$$\cos\theta = \frac{a}{h}$$




So, $a = h \times \cos\theta$

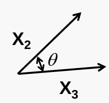
If h=100, $a = 100 \times \cos\theta$ ($a$ being the "east component")

## 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(angle) = +1.0, so angle=0° (cos 0°=+1.0)*
  - *If a correlation = 0.0, this means that cos(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:

| $\theta$ | cos $\theta$ (correlation) |
|---|---|
| 0° | 1.00 |
| 45° | 0.70 |
| 90° | 0.00 |
| 135° | -0.70 |
| 180° | -1.00 |

## Factor Analysis: A Geometric Approach

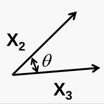
- Correlation interpreted as being connected to an angle:

| $\theta$ | cos $\theta$ (correlation) |
|---|---|
| 0° | 1.00 |
| 45° | 0.70 |
| 90° | 0.00 |
| 135° | -0.70 |

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

## Factor Analysis: A Geometric Approach

- Further discussion of this representation




Each "vector" represents a variable

$r_{23}$ (correlation of $X_2$ & $X_3$) = cos $\theta$
$X_2$, $X_3$ = standardized (unit length) vectors

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

## Factor Analysis: A Geometric Approach

- Example: FA on a 4x4 correlation matrix

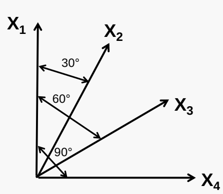
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| $X_1$ | 1.000 | 0.866 | 0.500 | 0.000 |
| $X_2$ | | 1.000 | 0.866 | 0.500 |
| $X_3$ | | | 1.000 | 0.866 |
| $X_4$ | | | | 1.000 |

- Convert to a matrix of angles

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| $X_1$ | 0° | 30° | 60° | 90° |
| $X_2$ | | 0° | 30° | 60° |
| $X_3$ | | | 0° | 30° |
| $X_4$ | | | | 0° |

## Factor Analysis: A Geometric Approach

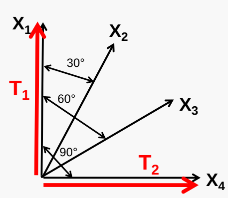
- This means graphically:




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

## Factor Analysis: A Geometric Approach

- This means graphically:

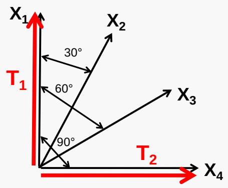



Remember, the objective of factor analysis is to simplify this situation by substituting a smaller number of factors for a larger number of variables.

**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):




Factors

| | $T_1$ | $T_2$ |
|---|---|---|
| $X_1$ | 1.000 | 0.000 |
| $X_2$ | 0.866 | 0.500 |
| $X_3$ | 0.500 | 0.866 |
| $X_4$ | 0.000 | 1.000 |

Variables

"Factor Loadings"

Factor loadings show how well each variable relates to each factor

## Factor Analysis: A Geometric Approach

- 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?*

## Factor Analysis: A Geometric Approach

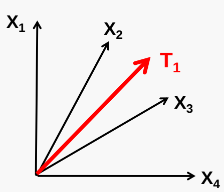
- The angle visualization idea gives us a number of "tried and true" methods for determining factors
- Each of the following methods are what we would call "rotations"
  - *Methods of positioning or "rotating" the factors we are creating so they best fit the variable vectors in our datasets*

## Factor Analysis: A Geometric Approach

- **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*

## Factor Analysis: A Geometric Approach

- **1. Centroid method**

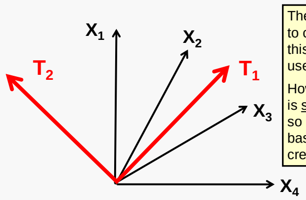



## Factor Analysis: A Geometric Approach

- **1. Centroid method**
  - *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.

## Factor Analysis: A Geometric Approach

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

## Factor Analysis: A Geometric Approach

- **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*

## Factor Analysis: A Geometric Approach

- **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)

## Factor Analysis: A Geometric Approach

- **Evaluation of Results: What is Good?**
  - *Regardless of which method you use, a few general principles should characterize a good factor matrix solution*

**Sample Factor Matrix**

Two Factors

Four Variables

| | $T_1$ | $T_2$ |
|---|---|---|
| $X_1$ | 1.000 | 0.000 |
| $X_2$ | 0.866 | 0.500 |
| $X_3$ | 0.500 | 0.866 |
| $X_4$ | 0.000 | 1.000 |

## Factor Analysis: A Geometric Approach

- **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**

| | $T_1$ | $T_2$ |
|---|---|---|
| $X_1$ | 1.000 | 0.000 |
| $X_2$ | 0.866 | 0.500 |
| $X_3$ | 0.500 | 0.866 |
| $X_4$ | 0.000 | 1.000 |

## Factor Analysis: A Geometric Approach

- **Evaluation of Results: What is Good?**
  - ***2. For every <u>pair of columns</u> in the factor matrix, there should be several variables with <u>high loadings on one column</u> (factor) and <u>low loadings on the other</u>***
    - Q: what does that mean?

| | $T_1$ | $T_2$ |
|---|---|---|
| $X_1$ | 1.000 | 0.000 |
| $X_2$ | 0.866 | 0.500 |
| $X_3$ | 0.500 | 0.866 |
| $X_4$ | 0.000 | 1.000 |

## Factor Analysis: A Geometric Approach

- **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?

**Sample Factor Matrix**

| | $T_1$ | $T_2$ |
|---|---|---|
| $X_1$ | 1.000 | 0.000 |
| $X_2$ | 0.866 | 0.500 |
| $X_3$ | 0.500 | 0.866 |
| $X_4$ | 0.000 | 1.000 |

## 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*
    - <u>Some strong factors</u> that are worthwhile, and are helpful in describing and accounting for the key characteristics of the dataset
    - <u>Some weak factors</u> that don't add much to your modeling of the dataset
  - *How many factors should we take from a factor analysis?*

## Factor Analysis: A Geometric Approach

- **To give a good basis for determining the number of factors to use, we need one more concept: the <u>eigenvalue</u>**
  - *The term "eigenvalue" originates from the <u>matrix models</u> that are used to conceptualize and run the factor analysis calculations*
  - *In other words, the "eigenvalue" term has a meaning that <u>goes beyond</u> the realm of factor analysis, into pure matrix algebra*

## Factor Analysis: A Geometric Approach

- **To give a good basis for determining the number of factors to use, we need one more concept: the <u>eigenvalue</u>**
  - *For our purposes here, we will make a couple of simple connections:*
    - 1. <u>Each factor</u> has an associated eigenvalue
    - 2. The <u>magnitude of the eigenvalue for each factor</u> is a measure of the "strength" of the factor (how *useful* or *powerful* the factor is)

## Factor Analysis: A Geometric Approach

- **So <u>what is an eigenvalue</u> in factor analysis?**
  - *Let's go back to our geometric interpretation of factors*
  - *Say we have created <u>one factor</u> that takes the place of <u>two variables</u>*

## Factor Analysis: A Geometric Approach

- Here is the factor matrix for this situation:

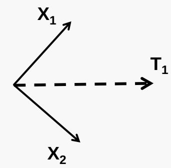
Two Variables → rows; One Factor → column

| | $T_1$ |
|---|---|
| $X_1$ | 0.70 |
| $X_2$ | 0.70 |

- The matrix shows that each of the variables has a "loading onto" the factor of 0.70: a <u>decent relationship</u> between the factor and the variables

## Factor Analysis: A Geometric Approach

- Geometrically, this situation could be represented like this:



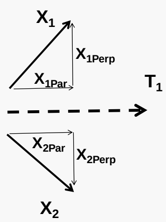

**Q:** what are the <u>components</u> of $X_1$ and $X_2$ that are "in the direction of" $T_1$?

## Factor Analysis: A Geometric Approach

- Subdivide the two variables into components

**Each variable has a component <u>parallel</u> to $T_1$, and a component <u>perpendicular</u> to $T_1$**

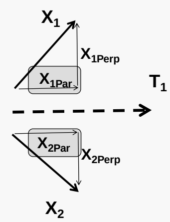
**We will focus on the <u>parallel</u> components here**




## Factor Analysis: A Geometric Approach

- Subdivide the two variables into components

**Key idea #1:** The parallel components are simply the <u>factor loadings</u> of $X_1$ and $X_2$ onto the factor $T_1$

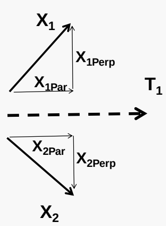



## Factor Analysis: A Geometric Approach

- Subdivide the two variables into components

**Key idea #2:** the eigenvalue of $T_1$ is simply the <u>sum of the loadings</u> of $X_1$ and $X_2$ onto $T_1$

**Eigenvalue$_{T1}$ = $X_{1Par}$ + $X_{2Par}$**

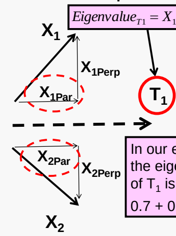



## Factor Analysis: A Geometric Approach

- Subdivide the two variables into components

**Key idea #3:** the eigenvalue of $T_1$ can be interpreted as the length (strength) of the factor

$Eigenvalue_{T1} = X_{1Par} + X_{2Par}$




In our example, the eigenvalue of $T_1$ is:
0.7 + 0.7 = <u>1.4</u>

## 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)***

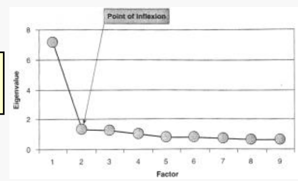
## Factor Analysis: A Geometric Approach

- **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)*

## Factor Analysis: A Geometric Approach

- **Back to our original purpose: how many factors should we use?**
  - *Another idea: draw a scree plot of eigenvalues*

**Q:** What conclusion should we draw from this graph?




## Factor Analysis: A Geometric Approach

- **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*

## Factor Analysis: A Geometric Approach

- **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**
  - *<u>Concept</u>: "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 <u>in each observation</u> 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

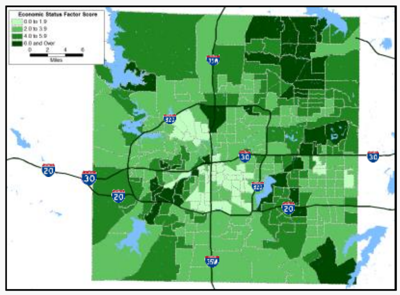
## Factor Analysis: A Geometric Approach

- **One last issue: FA and geography**
  - *<u>Quick example</u>: 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 <u>3 prominent factors</u> that represent what is going on in this database*
    - 1. a "social status" factor
    - 2. an "economic status" factor
    - 3. a "mobility" factor

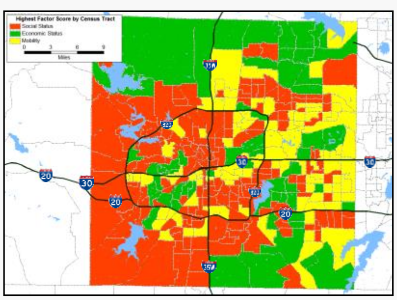
## Factor Analysis: A Geometric Approach

- **One last issue: FA and geography**
  - *<u>Factor score application</u>: you can calculate a factor score for <u>each factor</u> in <u>each census tract</u>*
  - *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 <u>values by census tract</u>, this example deals with a database of <u>values by people taking part in a study</u> (but this is the same basic idea as what we're discussing with census tracts)*