Data Theory and Dimensionality

Measurement, Scaling, and Dimensional Analysis
2019 ICPSR Summer Program
Prof. Adam M. Enders

Overview of Course

- Goal:
  - Explore a family of techniques that will help us better measure the things we're interested in
- Scaling models:
  - Are all geometric representations of data
  - provide information about the substantive processes that produce the data (DGP)
- Why "scale" things?
  1. Data reduction
  2. Assessment of dimensionality
  3. Measurement creation/testing
  4. Statistical graphics

What is Scaling All About?

Hypothetical data about 11 states' preferences about 3 policy areas

	Policy		
	A	B	C
s_1	10	5	0
s_2	9	6	1
s_3	8	7	2
s_4	7	8	3
s_5	6	9	4
s_6	5	10	5
s_7	4	9	6
s_8	3	8	7
s_9	2	7	8
s_{10}	1	6	9
s_{11}	0	5	10

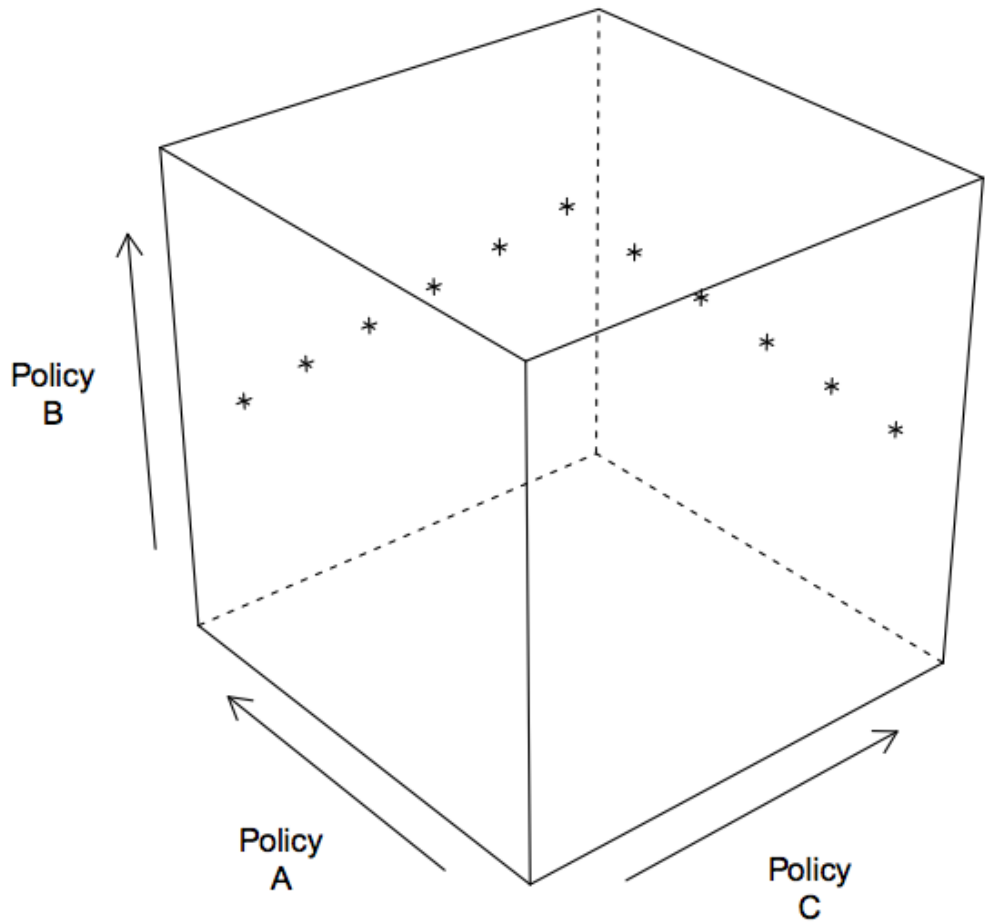
What is Scaling All About?, cont'd

Here's a matrix of correlations between the 3 variables

	A	B	C
A	1.00	0.00	-1.00
B	0.00	1.00	0.00
C	-1.00	0.00	1.00

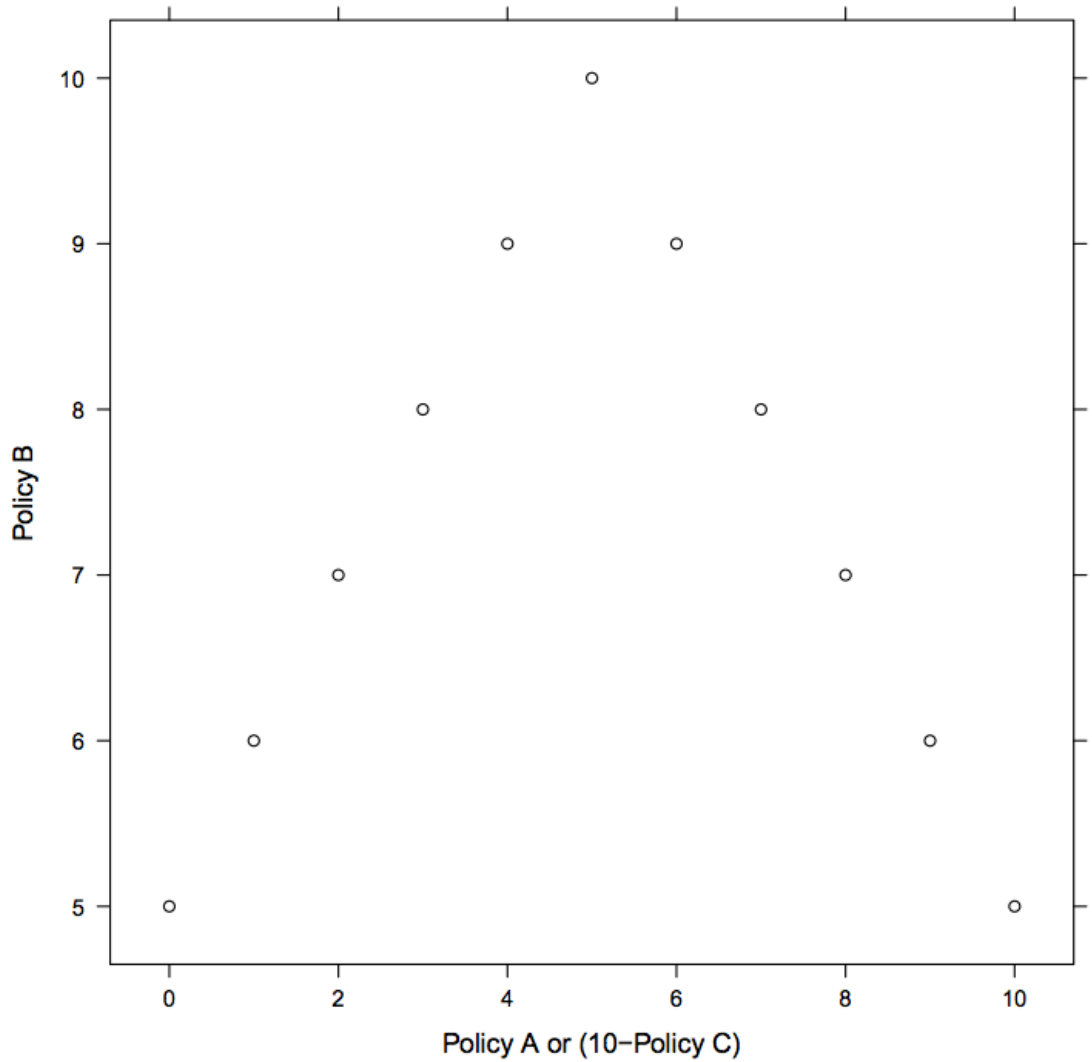
What is Scaling All About?, cont'd

Can represent this data graphically in 3 dimensions



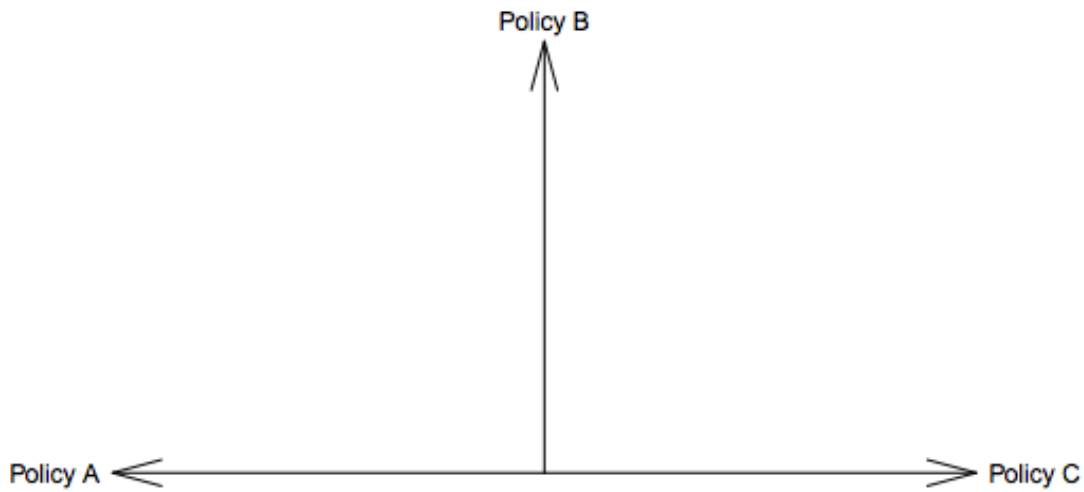
What is Scaling All About?, cont'd

Can represent this data graphically in a **more parsimonious** way in 2 dimensions



What is Scaling All About?, cont'd

More parsimoniously, yet (since all variables are independently represented)



What is Scaling All About?, cont'd

A unidimensional representation of the data – the most parsimonious possible depiction

This is what scaling seeks to do: minimize dimensionality, maximize parsimony (without sacrificing accuracy or utility)

Data and Data Theory

- What is data?
  - In the simplest sense, information to be analyzed
- What is data theory?
  - A theory that tries to impose a systematic structure on the data extraction process
  - Separate substantive and non-substantive content in data
  - Takes an abstract approach to data
  - Stripping away the trappings of the data and looking at the raw information reduces the myriad types of information to a small subset, allowing to see things we might otherwise miss

Example: Unfolding

- Clyde Coombs developed the unfolding model to answer his wife's (family psychologist) question about why people want X number of kids
- Fast forward 40 years – can be used on state spending data
- The substantive questions or data don't matter
- We're interested in the type of data, the structure of the data

Data Theory

- First comprehensive data theory developed by Coombs (1964) and reiterated by Jacoby (1991)
  - Conceptualized data as varying in two key ways:
    1. Whether rows and columns are from the *same* or *different* sets
    2. Whether relationship between rows and columns involves a *dominance* or *proximity* relation
- We're using one developed by Carroll, Arabie, and Young
  - Explicated nicely in MDS book by Young and Hamer (1987)
  - 2×2 design
  - Two types of classifications:
    1. Shape
    2. Relationship between rows/columns

1. Shape

- Data matrix can be **square** or **rectangular**
- Rectangular
  - The row and column objects are different entities
  - This is what most datasets look like, n rows and k columns
  - Example: people and variables
  - Usually more rows than columns, but really doesn't matter
  - If n and k are equal, it's still a rectangular matrix
    - Even though we're talking about "shape," we don't exactly mean *merely* physical shape
    - "Shape" is equally a question of (lack of) difference of row and column objects
- Square
  - Column and row objects that are identical
  - Example: correlation matrix – rows and columns are both variables

2. Relationship Between Rows/Columns

- How do we interpret the information in the cells of the data matrix?
- Regardless of the shape, the entry in each cell of the data matrix provides us info about the column and row
- Two central types of relationships: **dominance** and **proximity**
- Domniance
  - Refers to extremity of some object along a continuum
  - Example: A student possesses a certain level of arithmetic ability. A math problem possesses a certain degree of difficulty. If the student's ability exceed (i.e., "dominates") the math problem's difficulty, (s)he gets the answer correct.
- Proximity
  - Measures "matchingness" between row and column objects (admittedly clumsy language)
  - How proximal are they to each other?
  - Example: correlation – higher the correlation, the more proximal the variables are

Examples Combining Shape and Relationship

- Multidimensional scaling assumes a square, proximity data matrix
  - Rows and columns are equal (correspond to same objects)
  - Cell entries are measures of (dis)similarity between pairs of column/row objects
  - Substantive example: rating similarity between pairs of vehicles
- Unfolding assumes a rectangular, proximity data matrix
  - Rows and columns are unequal (correspond to different objects)
  - Row entries are measures of (dis)similarity between column objects
  - Substantive example: rank order of favorite fast food restaurants

Examples Combining Shape and Relationship, cont'd

- Cumulative scaling assumes a rectangular, dominance data matrix
  - Rows and columns are unequal (correspond to different objects)
  - Row and column entries can be rearranged to show extremity of objects along a latent continuum with respect to each other (if $o_i > o_j$ along latent dimension θ, then o_i "dominates" o_j)
  - Substantive example: participation in campaign activities
- Factor analysis models proximity relationships in a square matrix
  - Represents proximity as angles between variable and factor vectors in some m dimensional space
  - Row and column entries can be rearranged to show extremity of objects along a latent continuum with respect to each other (if $o_i > o_j$ along latent dimension θ, then o_i "dominates" o_j)
  - Substantive example: participation in campaign activities

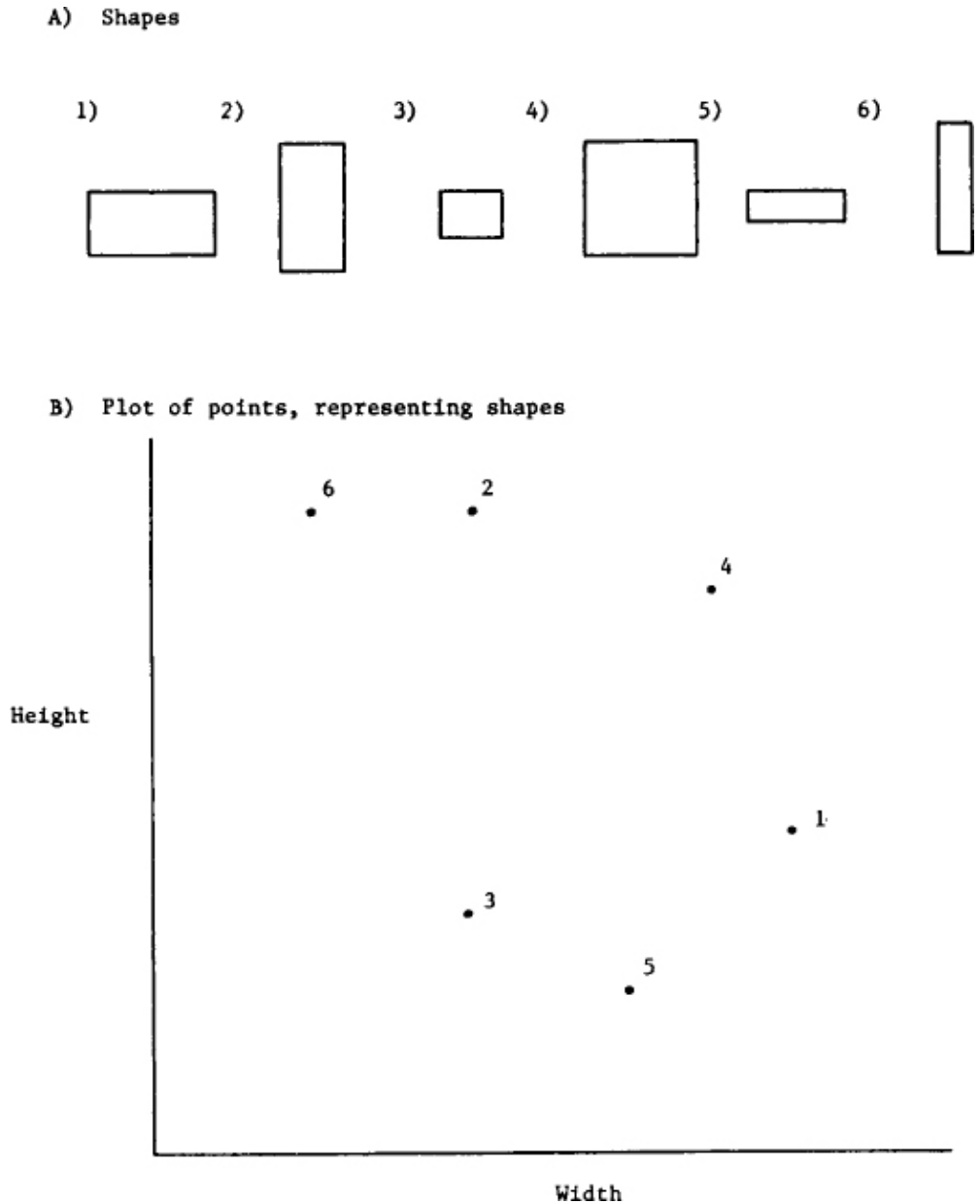
Dimensionality

- Refers to the number of **distinct, important**, or otherwise **interesting** sources of variability in a dataset (objects, variables, individuals, whatever)
  - A dataset with 500 will probably have 500 dimensions, but not all will be of *interest* to us
- Put differently, it refers to the number of coordinate axes required to map the objects with respect to each other
- Dimensionality is conceptual, rather than physical
  - We can try to represent conceptual dimensions physically, like constructing a graph, but it's not necessary
  - *Flat Land* by Edwin Abbot – we're limiting ourselves by confining dimensionality to the physical
  - Weisberg (1974), "Dimensionland: An Excursion into Spaces"

Dimensionality, cont'd

- We only want to focus on the important dimensions required to map objects
  - The objective of scaling methods is to find the important dimensions
  - Importantly, we want to find the **minimum** number of axes required to map a set of objects with respect to each other
- It follows, then, that there is no such thing as the "true" number of dimensions of some data/datum
- "Curse of dimensionality"
  - It's easier to understand what you can see, but we can only really "see" 3-4 dimensions
  - Scaling techniques help:
    1. reduce dimensionality (thus, diminishing the effects of the "COD"), and
    2. provide geometric interpretations of data that are particularly amenable to graphical visualization

Dimensionality: An Example (Jacoby 1991)



Dimensionality: An Example (Jacoby 1991)