Unidimensional Unfolding

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

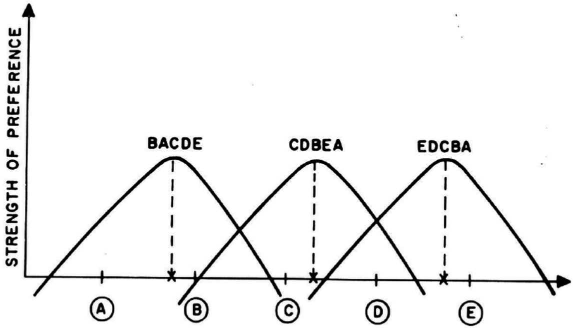
What is (Unidimensional) Unfolding?

- Scaling model that assumes "single-peaked" item response functions, rather than monotonic ones, like with the cumulative scaling model
- Represents **proximities** between the *n* rows and *k* columns of a **rectangular** data matrix as distances between points along a single continuum
- Objectives:
  - Represent row and column objects along a single latent continuum
  - Proximity between row and each column object should represent, to the best extent possible, the preferences/(dis)similarities from the original data matrix

Some Simple Examples

- "Do you like coffee with one lump of sugar?"
  1. "No, I like coffee without sugar"
  2. "No, I like coffee with more sugar"
- "Is voting the only way for people to have a say in government?"
  1. "No, voting is not a way"
  2. "No, there are more ways"
- In each case, the negative response could have one of two opposite meanings
- These are all proximity relationships, rather than dominance relationships
  - Imply a different IRF than the dominance/cumulative model does

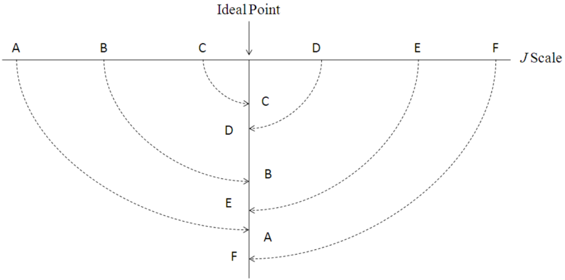
Subject "Utility Functions"



Each row object has it's own utility function that is also single-peaked

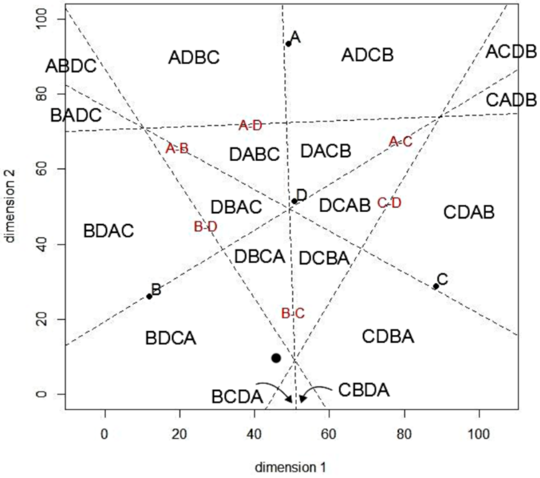
Preferences Can Be "Unfolded"

Just imagine a piece of string!



Generalization to Multiple Dimensions

Now imagine a napkin!



The Quintessential Example

- Spatial theories of voting, like the one popularized by Downs (1957)
  - The model itself (the unfolding model), devoid of substantive content, was first proposed Hotelling (1929) and completely developed by Coombs (1950)
- Commonality between most spatial theories of voting:
  1. Each voter can be represented by a point in some hypothetical space such that the point reflects the person's ideal set of policies
  2. The policy position of each candidate can be represented by a point in the same space
  3. A voter chooses the candidate whose policy position is closest to his or her own
- Note that "spatial" more or less implies "proximity," hence why proximity data is most appropriate for the unfolding model

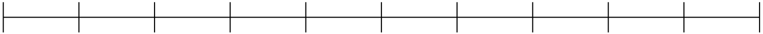
Basic Issue Space

Basic Issue Space

Abortion

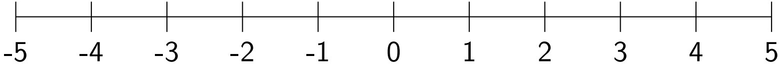
Basic Issue Space

Abortion



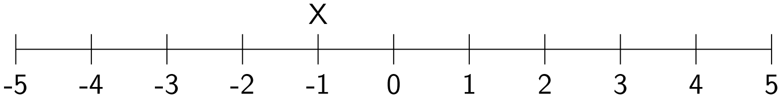
Basic Issue Space

Abortion

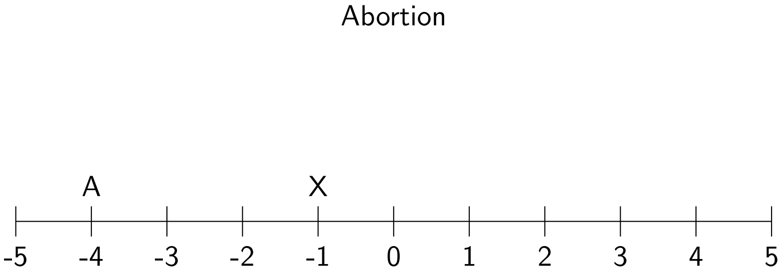


Basic Issue Space

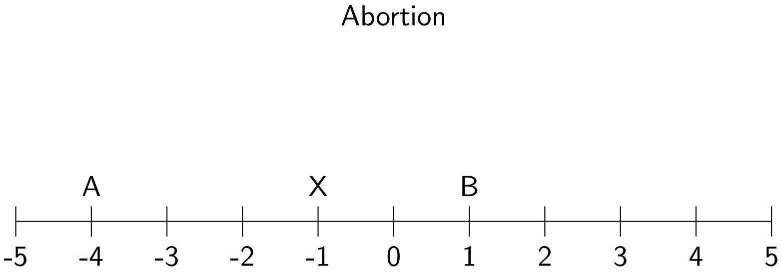
Abortion



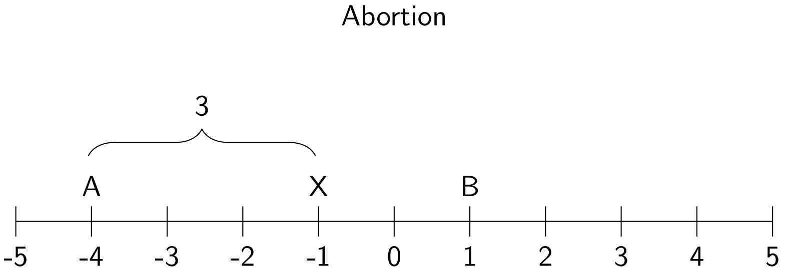
Basic Issue Space



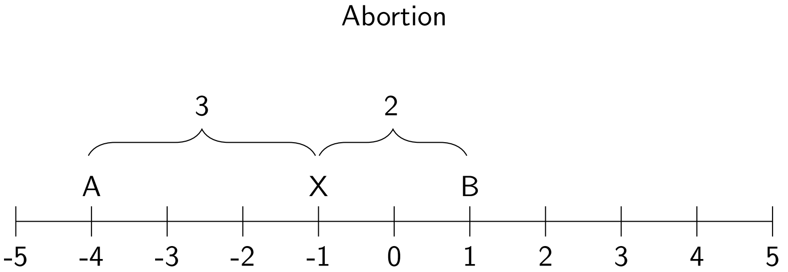
Basic Issue Space



Proximity Model

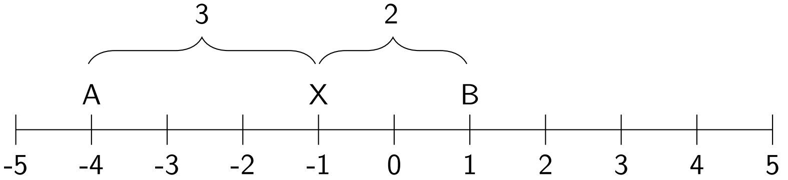


Proximity Model



Proximity Model

Abortion



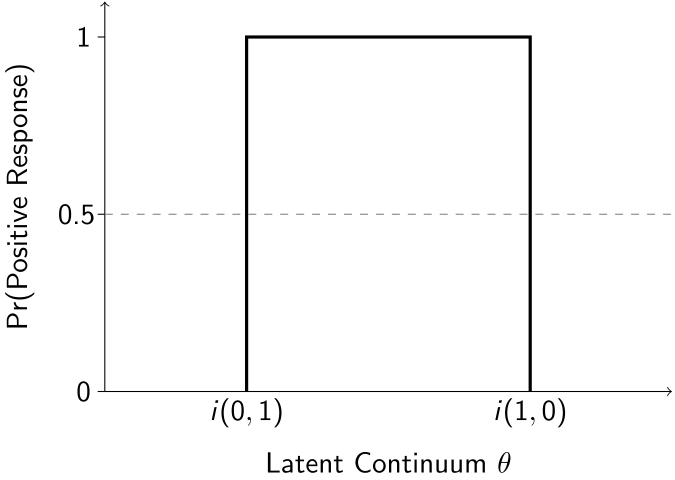
Since $2 < 3$, X votes for candidate B

The Deterministic Model

- Subjects only respond positively to items that represented close to their position on the latent continuum
- They respond negatively to items that are far away from that position
- Since they respond positively to all items close to their own position, they will respond positively only to items that are adjacent to each other
  - In other words, they have single-peaked preferences

The Deterministic Model

A "step function" with two steps: the left-sided step, in which the probability of a positive response increases from 0 to 1, and the right-sided step, in which the probability decreases from 1 to 0



Error-Free Response Pattern

Matrix of (rearranged) data that forms perfect (deterministic) unfolding scale:

Column Objects							
A	B	C	D	E	F	G	H
1	1	1	1	1	0	0	0
0	1	1	1	1	1	0	0
0	0	1	1	1	1	1	0
0	0	0	1	1	1	1	1

Unfolding originally called "parallelogram analysis" by Coombs (1964)

Assigning Scale Scores

1. First, column objects are given a rank score in the order in which they form an unfolding scale, using only the odd numbers
   - The rank score is assigned by rearranging rows and columns to construct (the closest approximation of) a parallelogram of 1's
2. Second, as their scale value, row objects are assigned the median value of the column scores of the column objects to which they responded positively

Error-Free Response Pattern

Matrix of (rearranged) data that forms perfect (deterministic) unfolding scale, with row and column scale scores:

	Column Objects								
Subjects	A	B	C	D	E	F	G	H	Row Score
1	1	1	1	1	1	0	0	0	5
2	0	1	1	1	1	1	0	0	7
3	0	0	1	1	1	1	1	0	9
4	0	0	0	1	1	1	1	1	11
Column Score	1	3	5	7	9	11	13	15	

A More Realistic Example

A close approximation of a parallelogram, but with obvious errors

Column Objects							
A	B	C	D	E	F	G	H
1	0	1	1	1	0	0	0
0	1	1	1	0	1	0	0
0	0	1	1	0	0	1	0
0	0	0	1	0	1	1	1

Problems: 1) how do we quantify error, and 2) how do we assign scale scores if we are comfortable that observed error is sufficiently negligible?

Assessing and Dealing with Error

- Like with the cumulative scaling model, we can use Loevinger's H coefficient
  - We compare the number of observed errors to the number of errors expected under statistical independence (i.e., the case where the data do not form unfolding scale)
  - $H = 1 - \frac{E(obs)}{E(exp)}$
  - Here, the observed errors are according to the unfolding model, rather than the cumulative scaling model
  - H still bound between 0 and 1, where 1 is perfect model fit
- This could help assess fit in the deterministic model, but couldn't help with dealing with erroneous response patterns
  - How could you assign scale values to response patterns with errors?
  - Most of the time, practitioners would simply drop rows with erroneous response patterns as long as a majority of the data was left

The Nonparametric Model

- Like with cumulative scaling, we're next going to consider a nonparametric version of the unfolding model
- Several reasons for doing so:
  - Like with Mokken Scaling, the nonparametric approach takes assumptions, and checking of those assumptions, very seriously
  - Assumes only nonmonotonic single-peaked preferences, rather than IRFs of a particular shape (e.g., Gaussian, quadratic, step function)
  - If you understand the process with the nonparametric formulation, you can easily understand the parametric one
- Important caveat:
  - Not many software packages can reliably estimate a A) unidimensional B) parametric unfolding model
  - "smacof" in R is unreliable...PROC MDS in SAS and ALSCAL in SPSS are best
  - Even though unidimensional parametric models are difficult, multidimensional unfolding is not

The Nonparametric Model

1. We will consider the ordinal nonparametric unidimensional unfolding model developed by van Schuur (1984)
   - Frequently referred to as the MUDFOLD model
2. Very recently (December 2017) written into an R package called "mudfold"
   - Does not *yet* have ability to deal with non-dichotomous data (though it does include a function for dichotomizing data if that seems worthwhile/appropriate)
   - However, this means that not many people have had a chance to implement it in published research yet
   - So, lots of opportunities to do something totally unique in your field

Parametric Unfolding

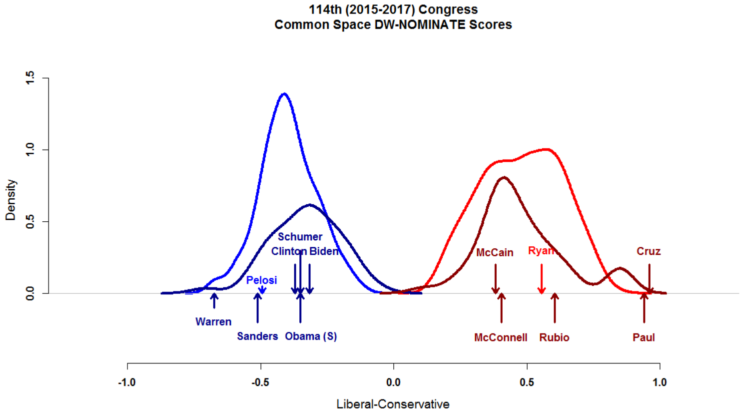
- Just like the cumulative model (MSA) has a family of parametric analogues, so too does the unfolding model
- Note: some think of the unfolding model as a particular form of IRT model
  - In a way that makes sense...the IRF is just a different shape
  - But, (cumulative) IRT models are **dominance** models, and unfolding is a **proximity** model
  - Thus, many psychometricians think of them as different
- The "mirt" package includes two IRT-based formulations of the unfolding model: the "dichotomous ideal point model" and the "generalized graded unfolding model" (GGUM)
- Can estimate these just like other IRT models, and use same person and item fit statistics to assess model fit

Other Options for Estimation

1. Optimal Classification, developed by Poole (2000, 2005)
   - R package called "oc"
   - Pros: can estimate unidimensional model, nonparametric
   - Cons: only dichotomous data, programmed in language of legislators/votes
2. Ordinal Optimal Classification, Hare et al. (2018)
   - Pros: nonparametric, ordinal data, can estimate unidimensional model
   - Con: not fully implemented in package yet
3. Smacof, developed by de Leeuw (lots of papers)
   - Pros: R package, ordinal and interval input data
   - Cons: won't fit unidimensional models, no dichotomous data
   - In a pinch could fit 2-dimensional model to unidimensional data and just use first dimension coordinates

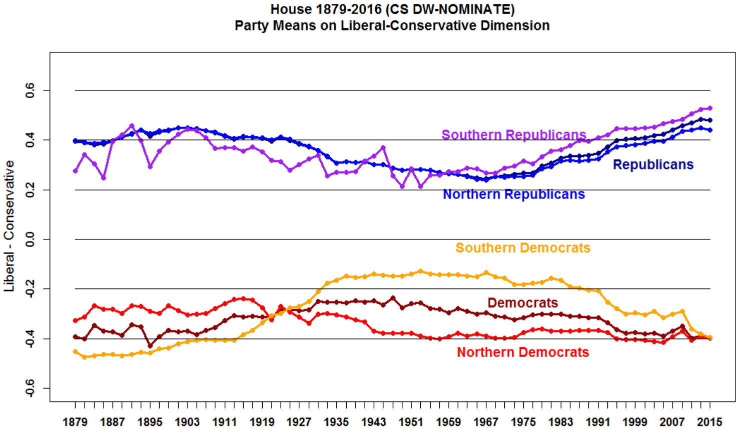
Example: Congressional Ideology

Distribution of ideal points along first dimension in 2015-2017 House of Representatives:



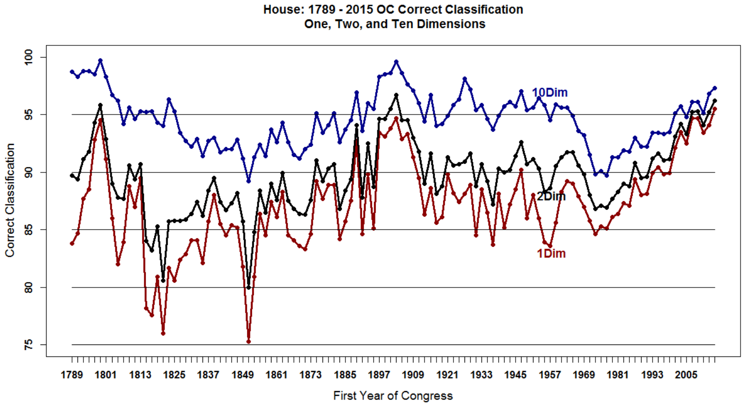
Example: Congressional Ideology

Average ideal point of (sub-)party along single dimension over time



Example: Congressional Ideology

Though the model is frequently estimated in multiple dimensions, only the first really matters much:



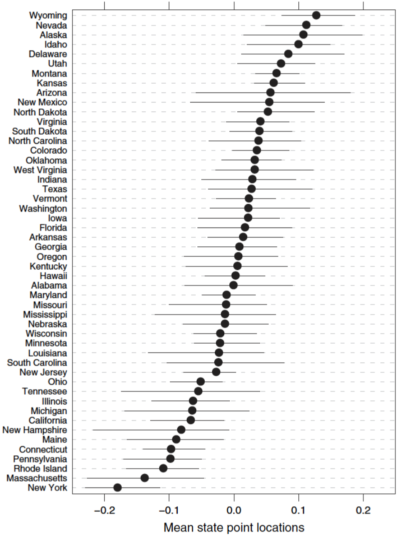
Example: State Spending Priorities

Locations of policy spending areas along latent dimension:

Interpretation: spending on particularized benefits vs. spending on collective goods

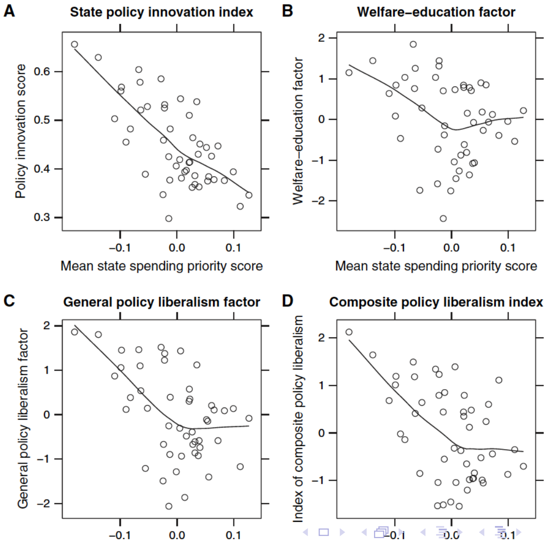
Example: State Spending Priorities

Average locations of states on latent continuum over many years:



Example: State Spending Priorities

(Criterion) validation by examining relationships with other indicators of state spending and ideology:



Unfolding vs. Factor Analysis

Hypothetical dataset

	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

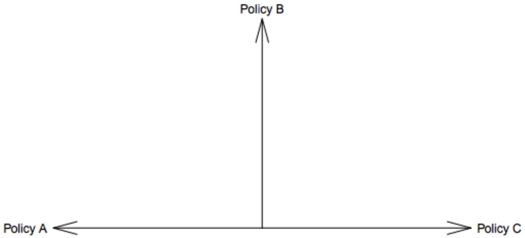
Unfolding vs. Factor Analysis

Correlation matrix

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

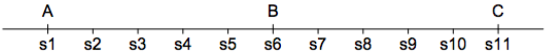
Unfolding vs. Factor Analysis

Results of factor analysis (which models correlations)



Unfolding vs. Factor Analysis

Results of unfolding analysis



Unfolding vs. Factor Analysis

- Bipolar constructs, in particular, can be "difficult" for factor analysis
- Oftentimes the two "halves" of a bipolar continuum end up being represented by two distinct latent factors
- Why?
  - Unfolding is a model of distance, or proximity
  - Factor analysis is a model of correlations, which correspond to the angular separation between pairs of variable vectors
- My point: dimensionality is much more theoretical and flexible than we tend to think
  - The most powerful measurement of a latent construct will start with some serious thinking about the DGP, and then finding a model that corresponds to it (and the data "type")