Week 7 Video 5

Factor Analysis

Factor Analysis

- You have a whole lot of variables
- Can you group them into “factors”?

Factor Analysis and Clustering

- Not the same
  - Clustering finds how data points group together
  - Factor analysis finds how data features/variables/items group together
- In many cases, one problem can be transformed into the other
- But conceptually still not the same thing

Goal 1 of Factor Analysis

- You have a lot of quantitative* variables
  - And since you have a lot of variables you have high dimensionality

- You want to reduce the dimensionality into a smaller number of factors

Goal 1 of Factor Analysis

* -- there is also a variant for categorical and binary data, Latent Class Factor Analysis (LCFA -- Magidson & Vermunt, 2001; Vermunt & Magidson, 2004), as well as a variant for mixed data types, Exponential Family Principal Component Analysis (EPCA – Collins et al., 2001)

Goal 2 of Factor Analysis

- You have a lot of quantitative* variables
  - And since you have a lot of variables you have high dimensionality

- You want to understand the structure that unifies these variables

Classic Example

- You have a questionnaire with 100 items
- Do the 100 items group into a smaller number of factors?
  - E.g. Do the 100 items actually tap only 6 deeper constructs?
  - Can the 100 items be divided into 6 scales?
  - Which items fit poorly in their scales?
- Common in attempting to design questionnaire with scales and sub-scales

Another Example

- You have a set of 600 features of student behavior
- You want to reduce the data space before running a classification algorithm
- Do the 600 features group into a smaller number of factors?
  - E.g. Do the 600 features actually tap only 15 deeper constructs?

Another Example

- You have a taxonomy of 120 design features that an e-learning lesson could possess
- You want to reduce the data space before studying the relationship between these features and student learning
- Do the 120 design features group into 8 factors?
  - E.g. Do the 120 features actually group into a set of 8 dimensions of tutor design?

Two types of Factor Analysis

- Experimental
  - Determine variable groupings in bottom-up fashion
  - More common in EDM
- Confirmatory
  - Take existing structure, verify its goodness
  - More common in Psychometrics

Mathematical Assumption in most Factor Analysis

- Each variable loads onto every factor, but with different strengths
  - Some strengths are infinitesimally small

Example

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

Computing a Factor Score
Can you write an equation for F1?

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

Can you write an equation for F1?

(It's just a straight-up linear equation, like in linear regression! Cazart!)

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

0.01V1-0.62V2+0.003V3+0.04V4+0.05V5-0.66V6
+0.04V7+0.02V8+0.32V9+0.01V10-0.03V11+0.55V12

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

Popup quiz
Can you write an equation for F2?

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

Can we do a fill-in-the-blank?

If so, the answer is

$$-0.7V1+0.1V2-0.14V3+0.03V4$$
$$+0.73V5+0.02V6-0.03V7-0.01V8$$
$$-0.34V9-0.02V10-0.02V11-0.32V12$$

Which variables load strongly on F1?

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

Wait… what's a "strong" loading?

- One common guideline: > 0.4 or < -0.4

- Comrey & Lee (1992)
  - 0.70 excellent (or -0.70)
  - 0.63 very good
  - 0.55 good
  - 0.45 fair
  - 0.32 poor

- One of those arbitrary things that people seem to take exceedingly seriously
  - Another approach is to look for a gap in the loadings in your actual data

Which variables load strongly on F1?

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	**-0.62**	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	**-0.66**	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	**0.55**	-0.32	0.02

Which variables load strongly on F2?

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

Which variables load strongly on F2?

	F1	F2	F3
V1	0.01	**-0.7**	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	**0.73**	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

Quiz:
Which variables load strongly on F3?

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

A) **V3, V7, V8, V11**
B) V3, V7, V11
C) V8
D) V1, V2, V4, V5, V6, V9, V10, V12

Which variables don't fit this scheme? (e.g. don't load strongly on any factor)

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	0.04	0.03	-0.02
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	0.32	-0.34	0.02
V10	0.01	-0.02	-0.07
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

Which variables don't fit this scheme? (e.g. don't load strongly on any factor)

	F1	F2	F3
V1	0.01	-0.7	-0.03
V2	-0.62	0.1	-0.05
V3	0.003	-0.14	0.82
V4	**0.04**	**0.03**	**-0.02**
V5	0.05	0.73	-0.11
V6	-0.66	0.02	0.07
V7	0.04	-0.03	0.59
V8	0.02	-0.01	-0.56
V9	**0.32**	**-0.34**	**0.02**
V10	**0.01**	**-0.02**	**-0.07**
V11	-0.03	-0.02	0.64
V12	0.55	-0.32	0.02

But note that if the magic number was lower, V9 would be fine →

Assigning items to factors to create scales

- After loading is created, you can create one-factor-per-variable models ("scales") by iteratively
  - assigning each item to one factor
  - dropping the one item that loads most poorly in its factor, if it has no strong loading
  - re-fitting factors

Item Selection

- Some researchers recommend conducting item selection based on face validity – e.g. if it doesn't look like it should fit, don't include it

- Depends on how theory-driven you want to be
  - And how much of a theory you actually have!

How does it work mathematically?

- Two algorithms (Ferguson, 1971)
  - Principal axis factoring (PAF)
    - Fits to shared variance between variables
  - Principal components analysis (PCA)
    - Fits to all variance between variables, including variance unique to specific variables
- PCA is more common these days
- Similar, especially as number of variables increases

How does it work mathematically?

- First factor tries to find a combination of variable-weightings that gets the best fit to the data
- Second factor tries to find a combination of variable-weightings that best fits the remaining unexplained variance
- Third factor tries to find a combination of variable-weightings that best fits the remaining unexplained variance…

How does it work mathematically?

- Factors are then made orthogonal (e.g. uncorrelated to each other)
  - Uses statistical process called factor rotation, which takes a set of factors and re-fits to maintain equal fit while minimizing factor correlation
  - Essentially, there is a large equivalence class of possible solutions; factor rotation tries to find the solution that minimizes between-factor correlation

Looking at this another way…

- This approach tries to find lines, planes, and hyperplanes in the K-dimensional space (K variables)
- Which best fit the data
- This may remind you of support vector machines…

Goodness

- What proportion of the variance in the original variables is explained by the factoring?
  (e.g. r^2 – called in Factor Analysis land the estimate of the *communality*)

- Better to use cross-validated r^2
  - Still not standard

How many factors?

- Best approach: decide using cross-validated r^2
- Alternate approach: drop any factor with fewer than 3 strong loadings
- Alternate approach: add factors until you get an incomprehensible factor
  - But one person's incomprehensible factor is another person's research finding!

Desired Amount of Data

- At least 5 data points per variable (Gorsuch, 1983)
- At least 3-6 data points per variable (Cattell, 1978)

- At least 100 total data points (Gorsuch, 1983)
- Comrey and Lee (1992) guidelines for total sample size
  - 100= poor
  - 200 = fair
  - 300 = good
  - 500 = very good
  - 1,000 or more = excellent

- My opinion: use cross-validation and see empirically

OK you've done a factor analysis, and you've got scales

- One more thing to do before you publish
- Check internal reliability of scales
- Cronbach's α

Cronbach's α

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$$

- N = number of items
- C = average inter-item covariance (averaged at subject level)
- V = average variance (averaged at subject level)

Cronbach's α: magic numbers (George & Mallory, 2003)

- > 0.9 Excellent
- 0.8-0.9 Good
- 0.7-0.8 Acceptable
- 0.6-0.7 Questionable
- 0.5-0.6 Poor
- < 0.5 Unacceptable

Factor Analysis

- A powerful tool for discovering unknown structure in data
- Conceptually similar to clustering
- Finds an orthogonal type of structure

Next week

- Discovery with Models and Other Topics