Use

~~Exploratory Factor Analysis~~

Oblique Principal Component Cluster Analysis

to uncover the underlying structure

# of self-report instruments

Steve Gregorich

Statistical/Methodological Seminar

December 11, 2014

Overview

. Measurement models

. Common factor analysis model

. Exploratory factor analysis (EFA)

. EFA pitfalls

. Introduction to VARCLUS

. Using VARCLUS with examples

. 'Confirmatory' factor analysis (CFA) of VARCLUS models, with examples

. Summary

The common factor model

Indirect measurement

Some constructs are not directly observable

. attitudes, intelligence, consumer confidence, top quark

Unobserved constructs are sometimes called *latent* variables

. Latent variables are 'everywhere' (physics, medicine, economics)

It is sometimes possible to assess latent variables indirectly,
via multiple, fallible, observed—or *manifest*—variables

A *measurement model* relates latent variables to manifest variables.
That is, the latent variables are hypothesized to directly cause
responses to corresponding manifest variables

With multiple manifest variables per latent variable, the measurement
model can be empirically evaluated, via *common factor analysis*

(define 'common')

Common factor model: Conceptual example

Suppose I want to measure two dimensions of consumer confidence

Expectations for 6-months hence

. Business conditions (1 = worse; 2 = same; 3 = better)

. Employment (1 = fewer jobs; 2 = same; 3 = more jobs)

. Income (1 = decrease; 2 = same; 3 = increase)

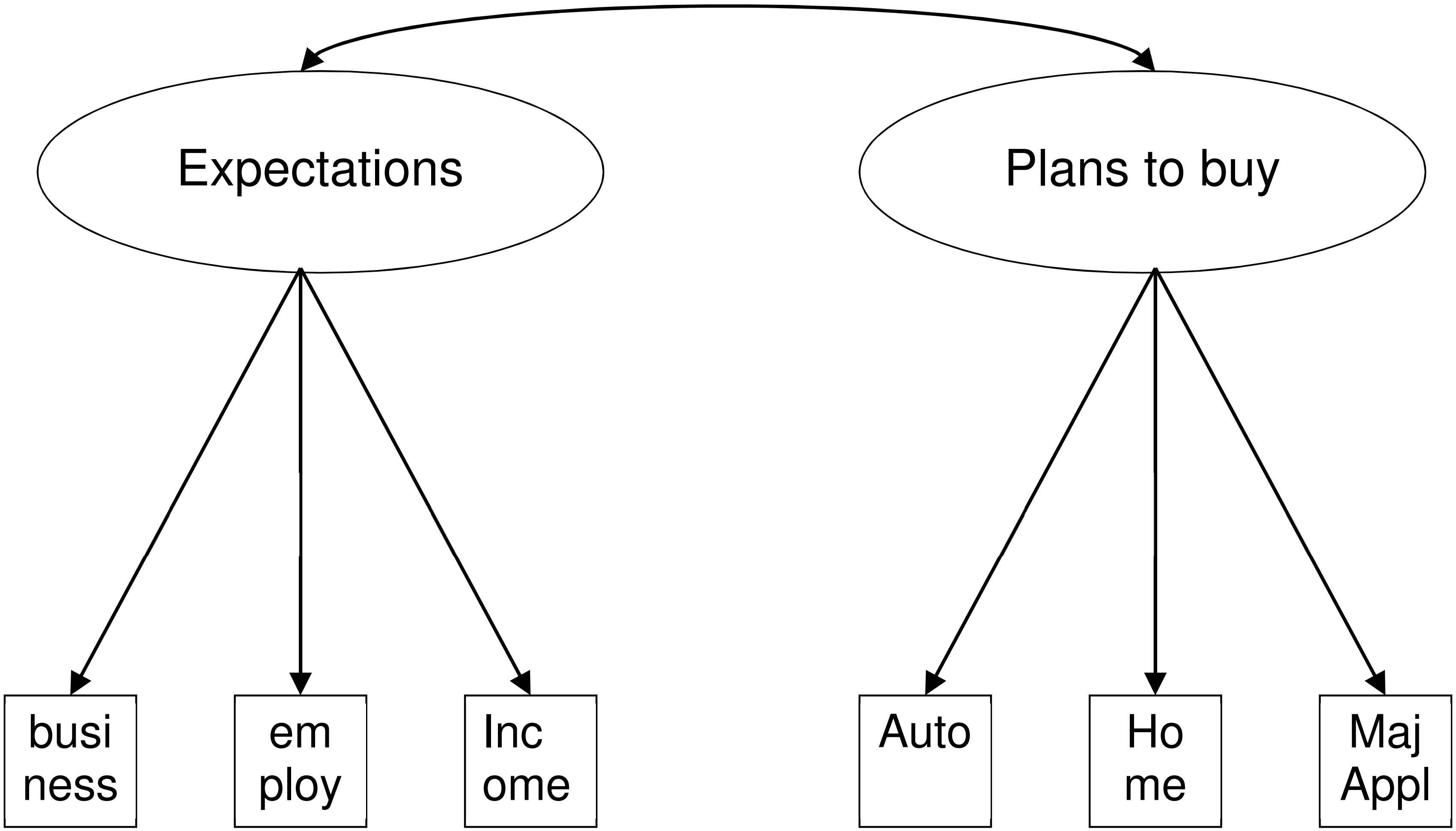
Personal purchase plans within 6-months

. Automobile

. Home

. Major appliances

Consumer confidence: Common factor model



(define single- and double-headed arrows)

Consumer confidence: *Made-up* common factor model

A generic representation of a factor pattern matrix
with 2 common factors and 6 manifest variables

	Expectations	Plans to buy
business	.67	.12
employment	.54	.11
income	.55	.07
auto	.05	.77
house	.09	.89
major appl.	.10	.57

The factor pattern matrix holds estimated correlations between latent and manifest variables

The latent variables are estimated from the observed data
. latent variables are unobserved, so their scalings are arbitrary

Correlations between latent and manifest variables aid interpretation

Question: Is the interpretation consistent with the motivating hypotheses?

Wait a minute…

How is it possible to estimate the relationship between
something measured (items) and something not measured (factors)?

Start with input data

The input data for a factor model are usually the
observed correlations or covariances among the observed items

Estimate factor loadings for your hypothesized model (an iterative search)

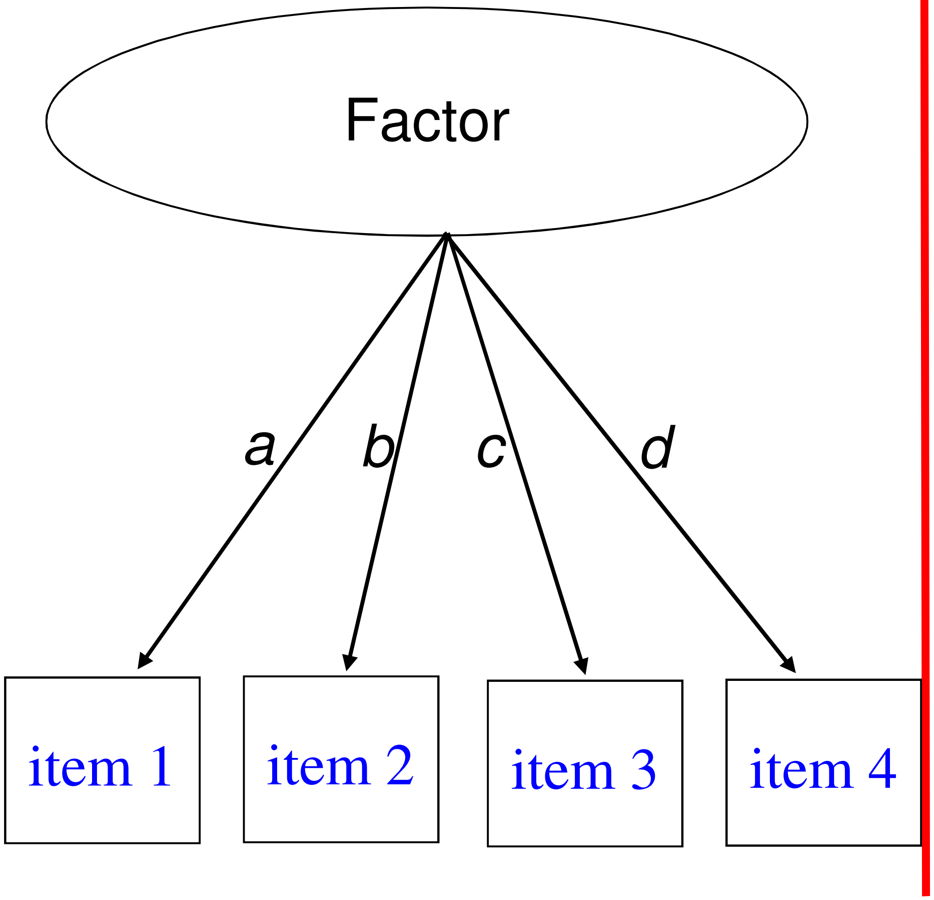
A well-fitting factor model and estimates can be used to
accurately reproduce the input data

Compare the model-reproduced data to the original data

Good correspondence between the two suggests that
the model has 'good fit' and we have more confidence
in the model and estimates

Relationship between standardized factor loadings and item correlations

Factor model and loading estimates



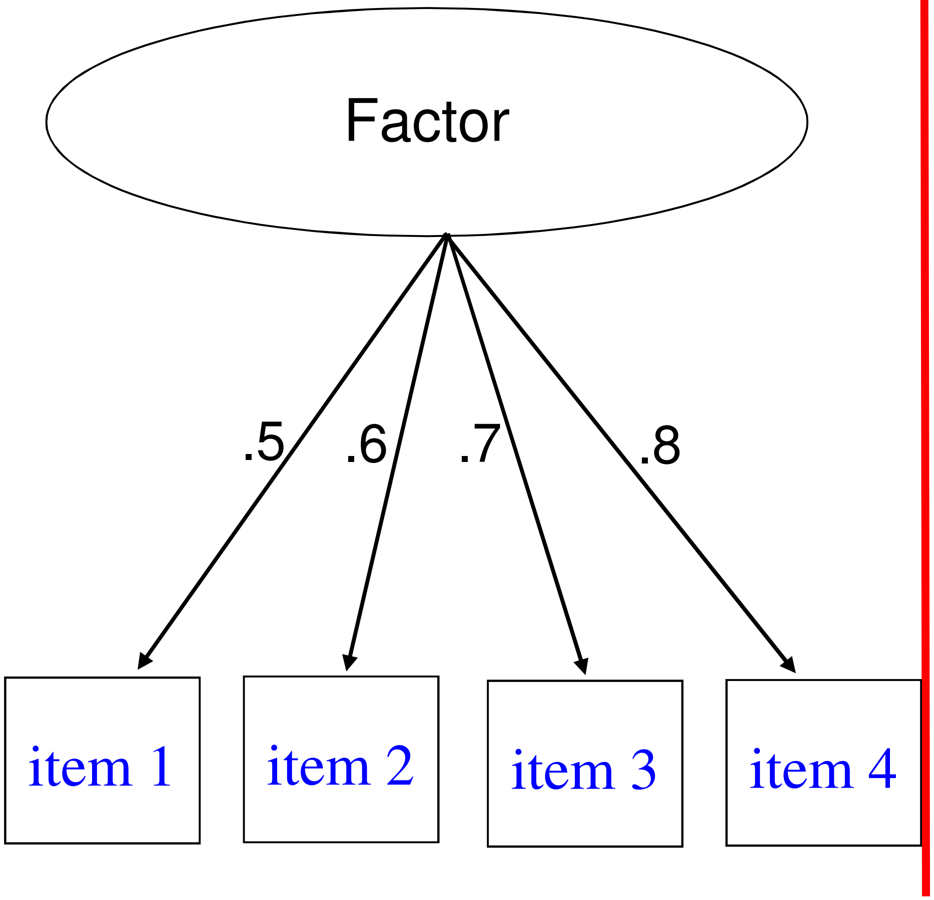
Model-implied item correlations

	item 1	item 2	item 3	item 4
item 1	1.0			
item 2	*a*×*b*	1.0		
item 3	*a*×*c*	*b*×*c*	1.0	
item 4	*a*×*d*	*b*×*d*	*c*×*d*	1.0

. 4 factor loadings (*a*, *b*, *c*, and *d*) attempt to explain 6 inter-item correlations

Relationship between standardized factor loadings and item correlations

Factor model and loading estimates



Model-implied item correlations

	item 1	item 2	item 3	item 4
item 1	1.0			
item 2	.30	1.0		
item 3	.35	.42	1.0	
item 4	.40	.48	.56	1.0

Empirical question

Do the model-implied correlations approximate the observed correlations?

Implications of empirical support for a measurement model

Demonstration of construct validity:
Do the items measure what they are hypothesized to measure?

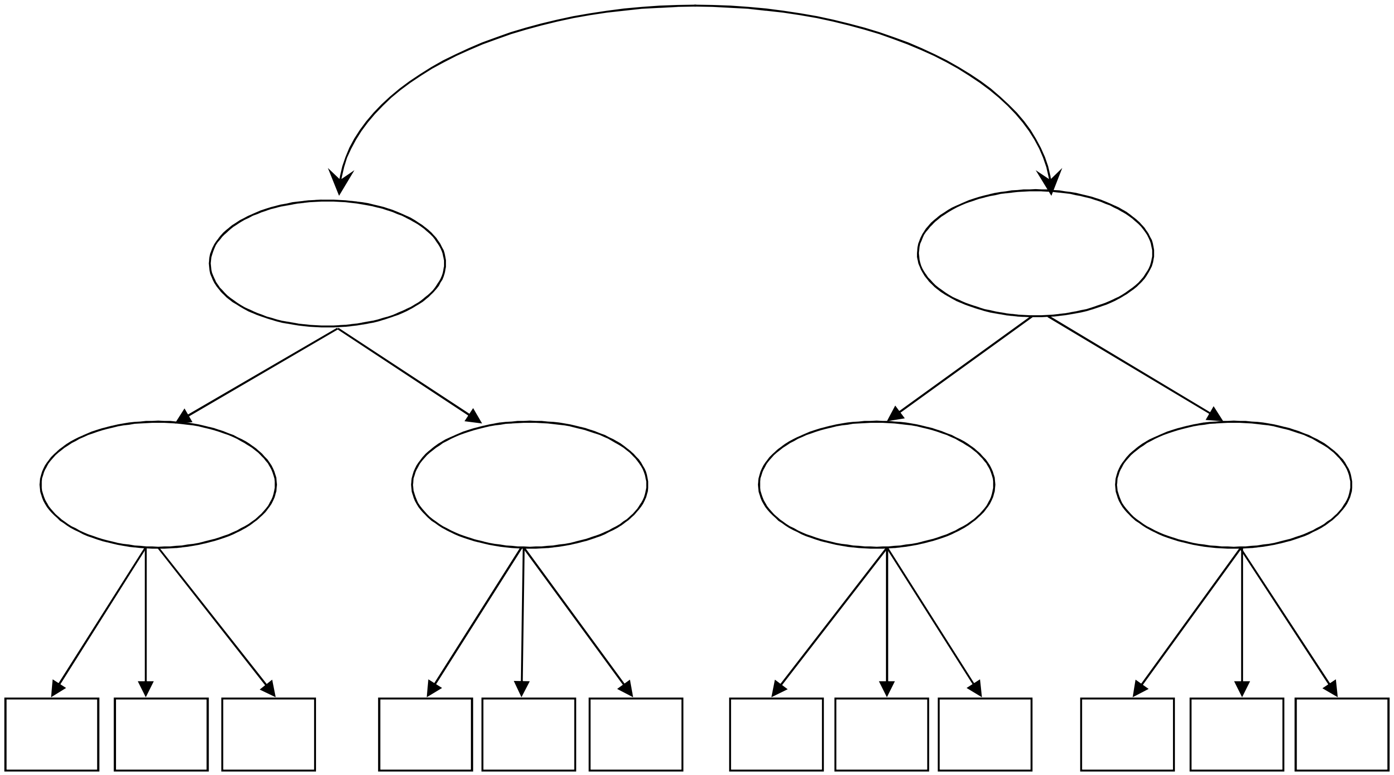
Provides empirical justification for creating summated composite scores, or 'scale scores,' which are more reliable than individual item scores

More parsimonious representation of information captured within item responses

Higher-order factor models

So far, we've discussed first-order factor models

Second- or higher-order factor models are possible



What if the hypothesized measurement model is not supported?

What if no-one proposes, a priori, a measurement model to test?

One option…

Exploratory factor analysis (EFA)

The goal of EFA is to uncover the measurement model

Exploratory Factor Analysis (EFA): The Data

- Over 110 years old (Spearman 1904)
- EFA is a variance decomposition method
- Start with collected data on items—the 'observed' variables
- Estimate so-called 'reduced,' item correlation or covariance matrix

'Full' Correlation Matrix (diagonal=1)

	item 1	item 2	item 3	item 4
item 1	1.0			
item 2	.30	1.0		
item 3	.35	.42	1.0	
item 4	.40	.48	.56	1.0

'Reduced' Correlation Matrix

	item 1	item 2	item 3	item 4
item 1	R^2			
item 2	.30	R^2		
item 3	.35	.42	R^2	
item 4	.40	.48	.56	R^2

(diagonal entries of the reduced matrix are known as communality estimates)

Exploratory Factor Analysis (EFA): Factor Extraction

. Extract orthogonal principal factors (PF) from reduced item correlation matrix

. 1st PF is *the* weighted composite that explains the max. common variation

. 2nd PF is *the* weighted composite that explains the max. of the remaining common variation and is uncorrelated with the 1st PF

. 3rd PF is *the* weighted composite that explains the max. of the remaining common variation and is uncorrelated w/ both the 1st and 2nd PFs. Etc.

(the PFs are eigenvectors of the reduced item correlation or covariance matrix)

. Rotate principal factors to aid interpretation

  . countless rotation criteria

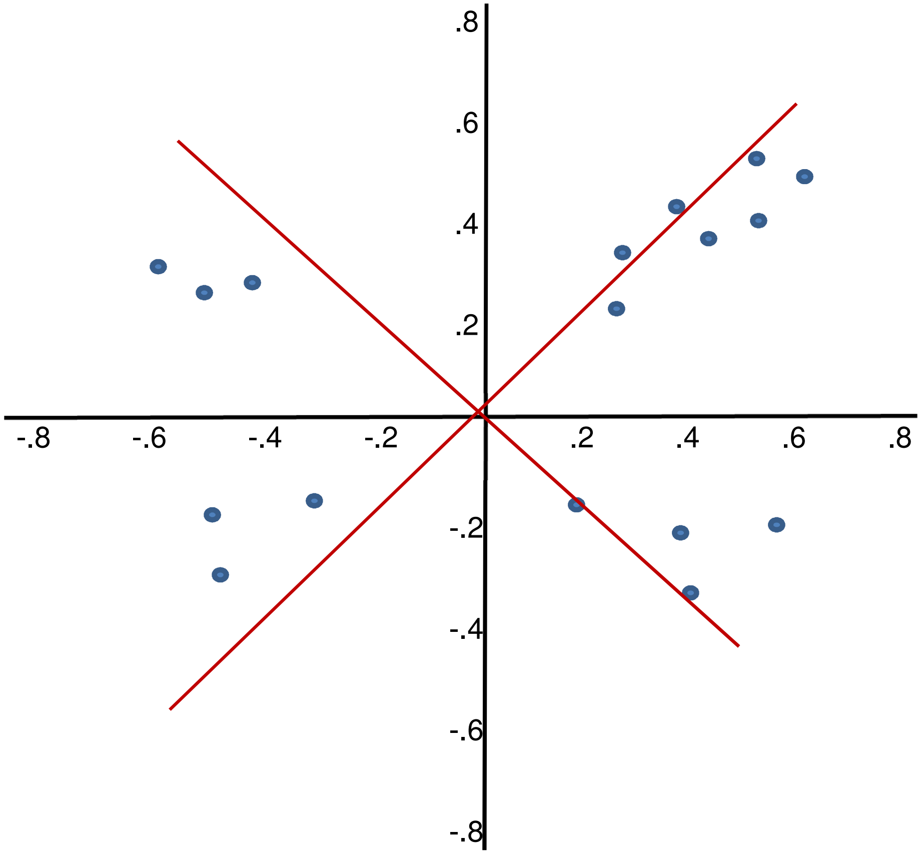
Note. If we started with the full item correlation matrix,
then we would have extracted so-called principal components (PC)

Exploratory Factor Analysis (EFA): Factor Rotation

. Black Axes are Extracted Principal Factors;   Blue Dots are Items

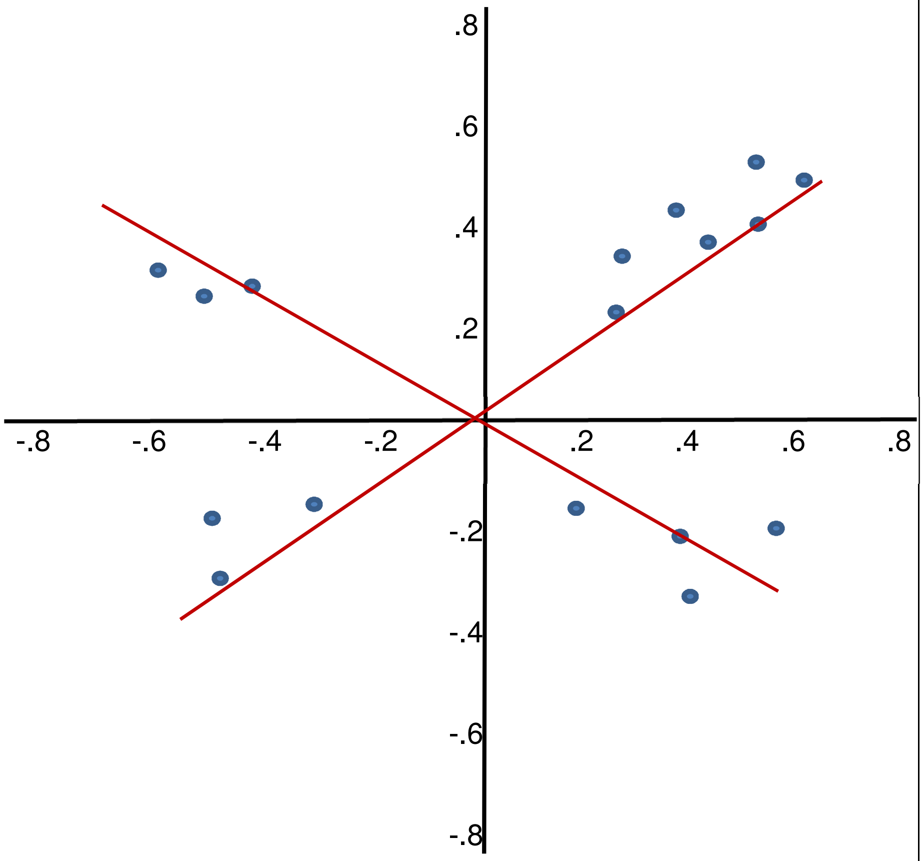
. Red Axes are Rotated Principal Factors

Orthogonal Rotation



(red axes are at 90-degree angles)

Oblique Rotation



(red axes are *not* orthogonal; better fit)

Exploratory Common Factor Analysis (EFA)

. The rotated principal factors are reported in a factor pattern matrix, e.g.,

	Expectations	Plans to buy
business	.67	.12
employment	.54	.11
income	.55	.07
auto	.05	.77
house	.09	.89
major appl.	.10	.57

Pitfalls of EFA

Can work well, if you are 'lucky'

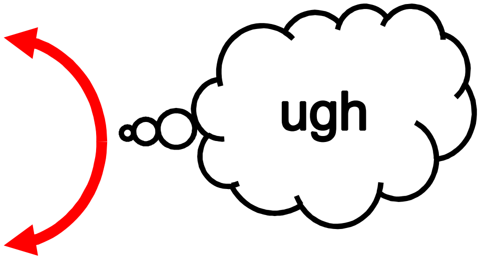
With large item sets (e.g., >30), difficulties often arise.

Simultaneous challenge

. (i) determine which items to drop from consideration…

extraneous items can obfuscate factor structure

. (ii) if the number of extracted factors is incorrect, then an important item can appear to be extraneous



Also

. Factor loading 'droop' (kudos to Ross Boylan)

. Many have sought a 'holy grail' rotation method—it doesn’t exist

. Personal example: IPC 79 items to 28 items in over 1 year

VARCLUS: Oblique Principal Components Cluster Analysis

A *'homespun hodgepodge'*

Divisive method
Start with all items in 1 cluster

Step 1. Identify the cluster that is most likely to benefit from splitting;
i.e., the cluster most likely to represent more than one underlying construct
(the cluster with the largest 2nd principal component)

Extract 2 principal components from the items in that cluster and rotate them
(via raw oblique QUARTIMAX)

Step 2. Iteratively reassign items to clusters;
Attempt to maximize explained variance

Repeat until stopping rule satisfied

VARCLUS

The SAS documentation is pretty Spartan

Only cites 3 references—none of them describe VARCLUS

I have no idea who invented VARCLUS

A literature search found <20 articles
—all but one is an application of VARCLUS

VARCLUS

VARCLUS code is simple

```
proc varclus data=<data> cov minclusters=<#min> maxclusters=<#max>;
  var <varlist>;
  run;
```

where

. <varlist> is the list of items to be clustered

. #min is the minimum number of clusters to extract
I suggest setting #min = 1; the default

. #max is the maximum number of clusters to extract
I suggest initially setting # to 1/3 the number of items in <varlist>

. COV requests analysis of the item covariance matrix (I always specify this!)
Analysis of the item correlation matrix is the default

VARCLUS example: MSM in China

A 42-item self-report measure of MSM's *Stigma Management* strategies.

Kyung-Hee Choi (PI) R01 project in China; Pilot data: *N*=150.

Kudos to Wayne Steward

Example item

. "To appear heterosexual, I sometimes talk about fictional dates with members of the opposite sex."

6-point response option:

1	2	3	4	5	6
Strongly Disagree	Moderately Disagree	Mildly Disagree	Mildly Agree	Moderately Agree	Strongly Agree

VARCLUS example: MSM in China

Look over the handout, pages 1-6

. Goal is to identify 'pure' first-order factors

'R-square with own cluster' and

'R-square with next closest'

I mostly rely on subjective judgment

I chose the 16 cluster solution (p. 5 of handout)

Note.

Using VARCLUS requires detailed variable labels

VARCLUS Example: MSM in China

. 16 clusters is a lot of clusters

. Perhaps there is a more parsimonious representation

. I used VARCLUS to identify 5, 2nd-order clusters (p. 6 of handout)

The approach

. Request VARCLUS to output the inter-cluster correlation matrix (ICCM)

Use the ICCM as the input data for a 2nd VARCLUS analysis.

By this method you obtain clusters-of-clusters, or 2nd-order clusters (see page 6 of handout)

VARCLUS: other considerations

1. The following two program runs likely will provide different results

A 'true' implementation of VARCLUS

```
proc varclus data=<data> cov minclusters=1 maxclusters=16;
  var <varlist>;
run;
```

. 16 principal components w/ raw oblique QUARTIMAX rotation

```
proc varclus data=<data> cov minclusters=16 maxclusters=16;
  var <varlist>;
run;
```

2. Consider transforming your data to better approximate normal distributions prior to fitting VARCLUS (or EFA) models.

Summary: VARCLUS

VARCLUS includes some helpful statistics such as
. R^2 with own cluster
. R^2 with next closest cluster
. Proportion of explained variation

Those are useful, but I consider them secondary
Subjective judgment about the conceptual 'purity' of candidate clusters is likely the best initial guide.

Singleton clusters are OK,
you can chose to ignore them

Don't be afraid to eliminate/ignore items that don't seem to be conceptually related to the other items within the same cluster

Summary: VARCLUS &VARCLUS/EFA followed by CFA

Subjective judgment is the best guide for choosing a VARCLUS solution

Still, I don't fully trust judgment

Therefore, CFA after VARCLUS is highly recommended

Not a confirmatory test, but does provide more stringent assessment

>3 times I have fit a CFA model to help defend a VARCLUS model and

>3 times the fit has been acceptable with little to no model modification

Given the size of the item sets (42, 68, 79), that success rate is surprising when compared to my personal experiences with CFA after EFA

YMMV: subjective judgment of VARCLUS output plays a part

Also, I tend to work with investigators who carefully craft items

Summary: Benefit to investigators

I used to 'run and hide' when investigators with large item sets
would ask about fitting factor analysis models

Even if they had an a priori measurement model, it likely required modification

So, after the confirmatory test of the measurement model failed, then what?
EFA? That process can take 'forever.'

So, unless there was a lot of salary support, I often 'ran'

With VARCLUS in the armamentarium, things have changed

Summary: Benefit to students/junior investigators

When students/post-docs/fellows/junior faculty would ask for a consult on EFA
I would tell them the topic is too 'deep' to be covered via consultation

I would suggest they take a class, but none are offered locally

This usually meant that the person seeking help was not going to get very far

Now, I can tell them about PROC VARCLUS. Explaining it takes < one hour.

The rub is that the consultee won't know anything about CFA—
That topic is also too 'deep' to be covered via consultation

Still, a junior investigator armed with VARCLUS and a little guidance
is much better off than with any available alternative

END