Use

~~Exploratory Factor Analysis~~

Oblique Principal Component Cluster Analysis

to uncover the underlying structure

of self-report instruments

Steve Gregorich

CAPS Methods Core Seminar

May 18, 2012

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)

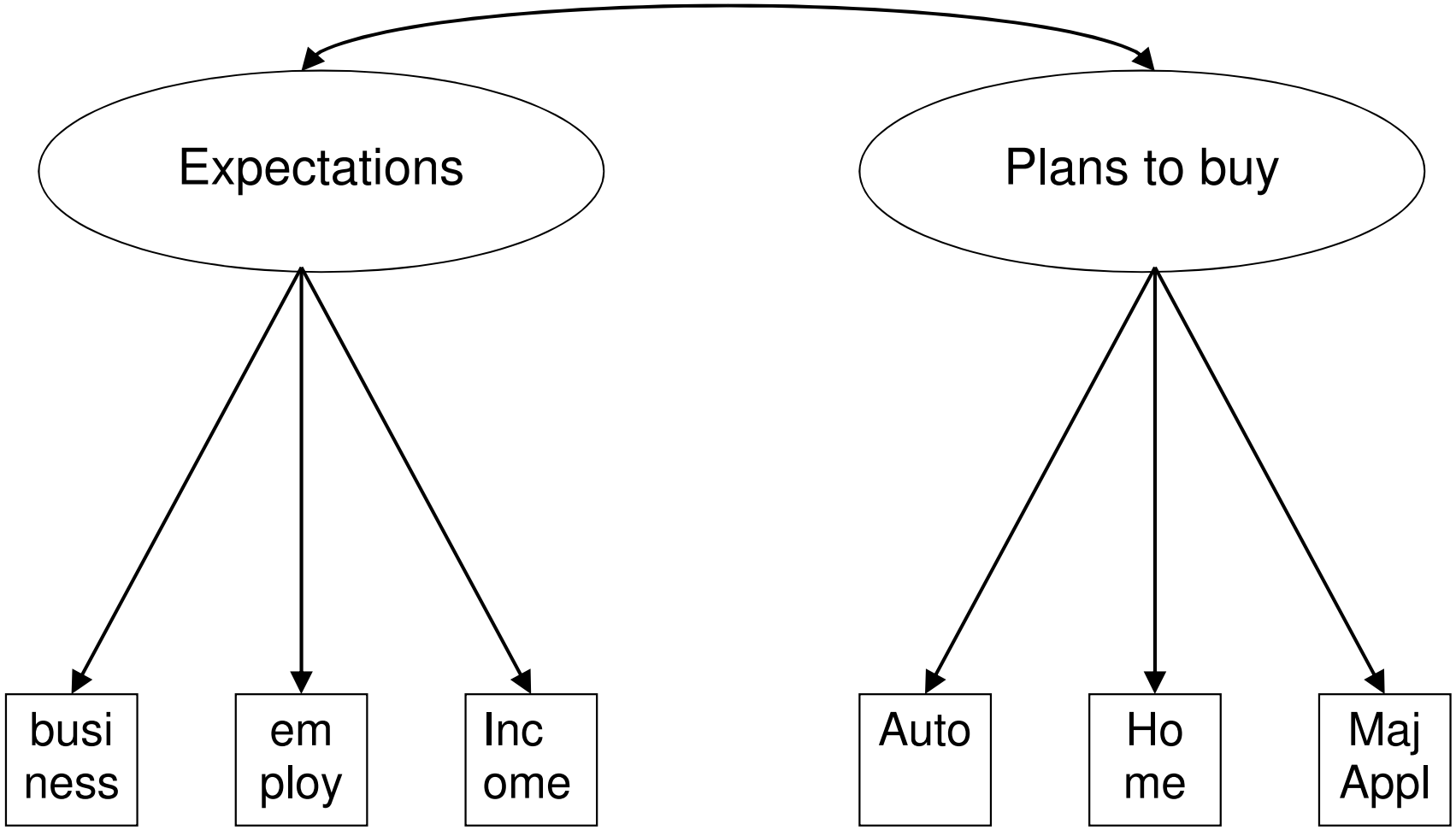
Personal purchase plans within 6-months

. 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)

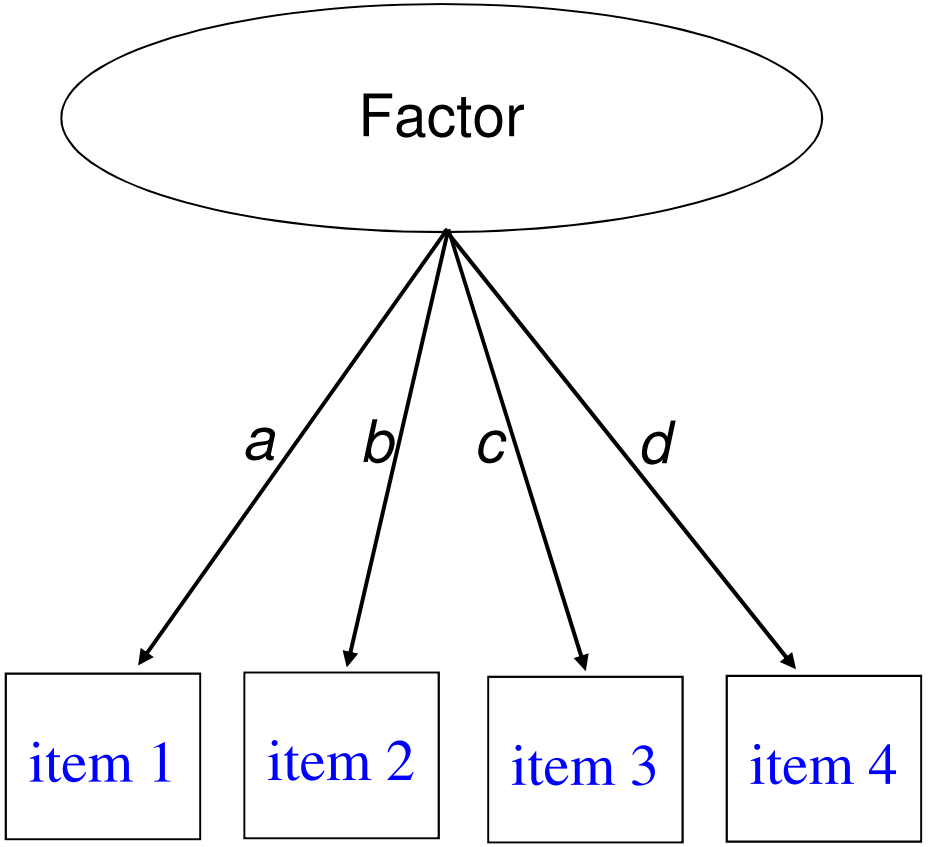
A well-fitting factor model and estimates can be used to
accurately reproduce the input data

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



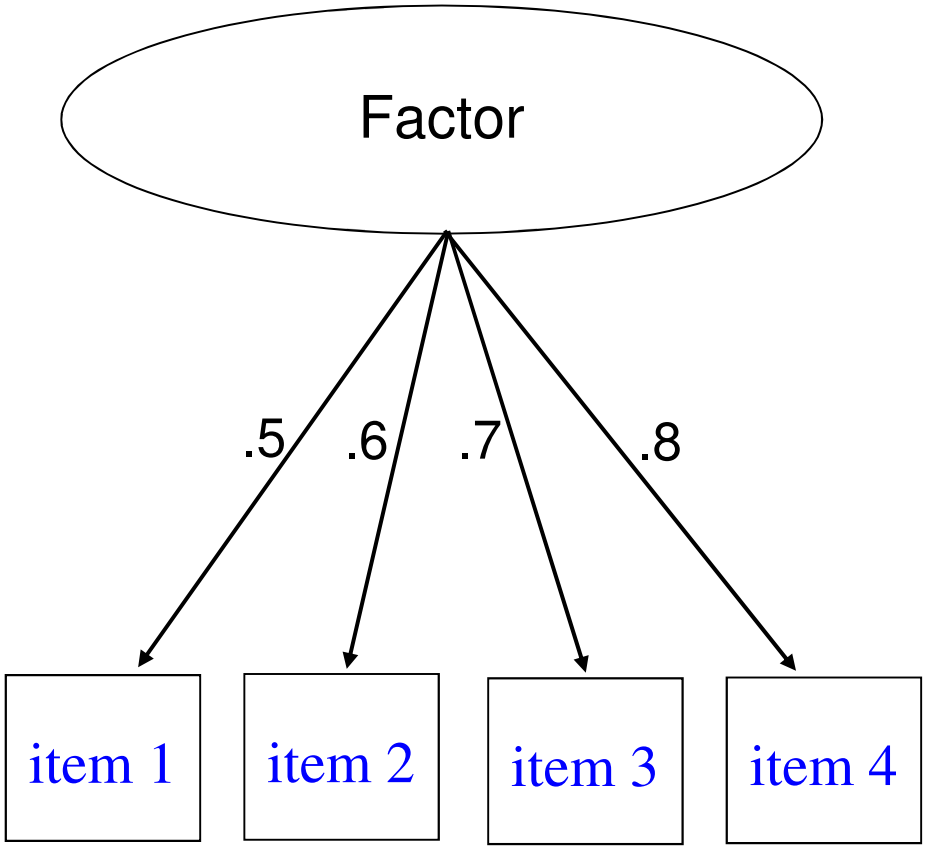
Model-implied item correlations

	item 1	item 2	item 3	item 4
item 1	1.0			
item 2	$a \times b$	1.0		
item 3	$a \times c$	$b \times c$	1.0	
item 4	$a \times d$	$b \times d$	$c \times d$	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,

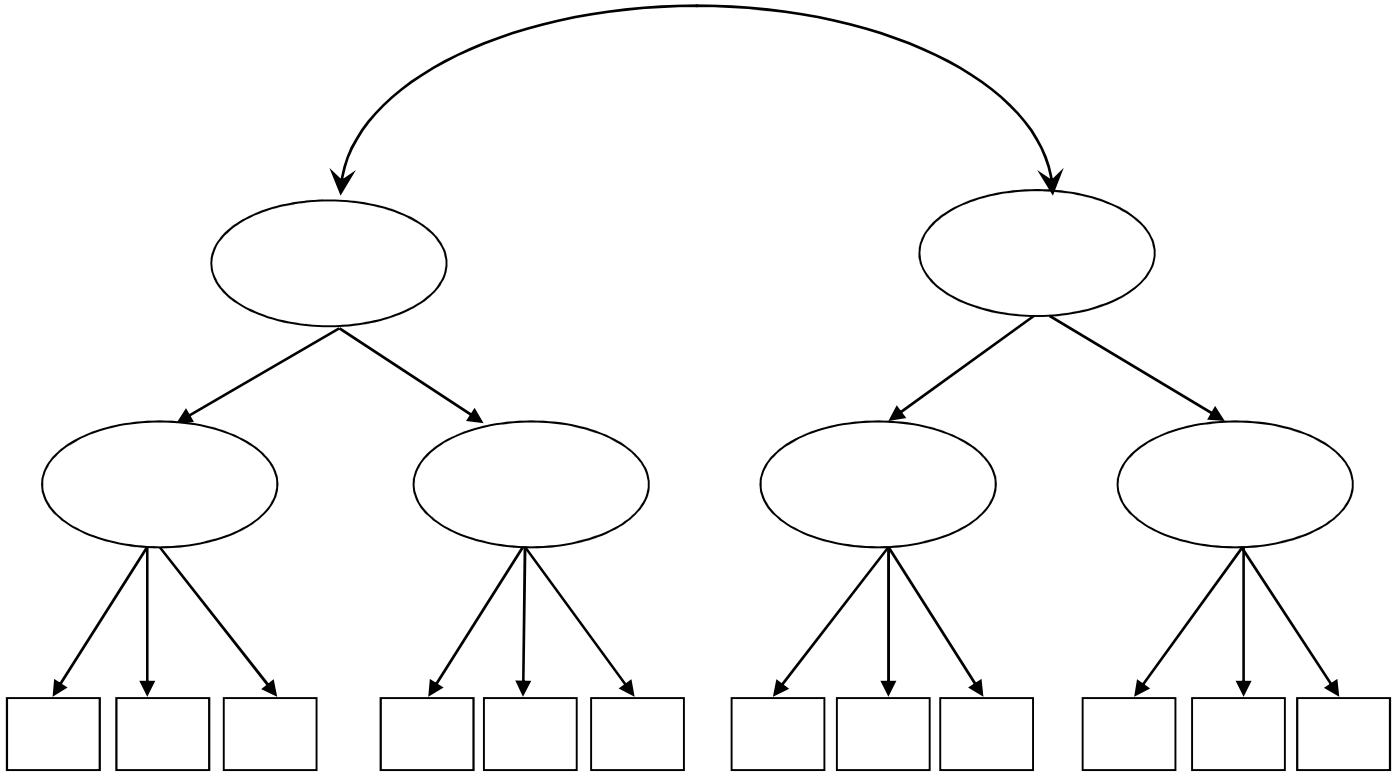
or 'scale scores,' which are more reliable than individual item scores

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 Common Factor Analysis (EFA)

. Over 100 years old (Spearman 1904).

. EFA is a variance decomposition method

- . Collected data on items—the 'observed' variables
- . Estimate 'reduced,' item correlation or covariance matrix
  - E.g., communality estimates are the diagonal elements
- . Extract orthogonal principal components: AKA 'principal factors'
  - Principal factors are the eigenvectors of the reduced
    - item correlation or covariance matrix
- . rotate principal factors to aid interpretation
- . countless rotation criteria

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) however, 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

. 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 with the largest second eigenvalue

Extract 2 principal components from the items in this cluster and rotate via raw oblique QUARTIMAX

Step 2. Iteratively reassign items to clusters;

Attempt to maximize explained variance

Repeat until stopping rule satisfied

Note. Working with principal component0,s not principal factors

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 applications 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 #1: 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 #1: 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

Note.

Using VARCLUS requires detailed variable labels

VARCLUS example #2: Parental feeding practices

A 68-item self-report measure of Latino parents' child-feeding practices.

Jeanne Tschann (PI) R01 project in SF

Baseline data: *N*=174 mom/dad pairs

Example item

"How often do you let your child eat whatever he/she wants?"

1	2	3	4	5
never	sometimes	often	very often	always

VARCLUS example #2: Parental feeding practices

Analysis plan

. Stack moms' and dads' data. i.e., 174 × 2 = 348 cases

. Use VARCLUS to identify 1st-order factors

. Reshape data to represent 174 data records, one per couple

. Fit 1st-order factor CFA model and test mom/dad measurement invariance

. Explore 2nd-order factor structure via VARCLUS and EFA

. Fit 2nd-order factor CFA model and test mom/dad measurement invariance

VARCLUS example #2: Parental feeding practices

Look over the handout, pages 7 & 8

I chose the 14 cluster solution

2 items dropped based upon VARCLUS results

3 additional items dropped during preliminary CFA modeling

Findings

The 14 cluster 1st-order model was supported by subsequent CFA including invariance of model parameters across moms and dads

The 2nd-order factor structure was complex

I tried using VARCLUS and EFA to identify a 2nd-order structure

Eventually, I chose a second-order structure including

Four 2nd-order factors and

Twelve 1st-order factors (the other 1st-order clusters were dropped)

VARCLUS example #2: Parental feeding practices

. 63 of 68 items retained in the final model

. Originally, we tried EFA to identify a factor structure
    Many more items were dropped e.g., I could retain about 50 of 68 items

VARCLUS example 3: Interpersonal Processes of Care

A 79-item patient-report measure of MD/provider processes of care.

Anita Stewart (PI) R01 project in SF

Baseline data: *N*=1622 patients from 4 racial/ethnic/language groups

African American, Latino-English, Latino-Spanish, White

Example item

"How often did doctors speak too fast?"

1	2	3	4	5
never	rarely	sometimes	usually	always

VARCLUS example 3: Interpersonal Processes of Care

Original analysis plan

. Stack racial/ethnic/language group data. i.e., *N*=1622 cases

. Use EFA to identify 1st- and 2nd-order factors

. Unstack data: separate 4 r/e/language groups

. Fit 2nd-order factor CFA model and test invariance across r/e/language

Result

This process took me well over 1 year of intermittent work

The final 2nd-order included only 29 of 79 items

Twelve 1st-order factors and seven 2nd-order factors

Stewart, A.L., Nápoles-Springer, A.M., Gregorich, S.E. and Santoyo-Olsson, J. (2007). Interpersonal processes of care survey: Patient-reported measures for diverse groups. *Health Services Research*, 42, Part I, 1235-1256.

VARCLUS example 3: Interpersonal Processes of Care

New analysis plan

. Stack racial/ethnic/language group data. i.e., *N*=1622 cases

. Use VARCLUS to identify 1st-order factors

. Fit 1st-order factor CFA model

. Explore 2nd-order factor structure via VARCLUS and EFA

. Fit 2nd-order factor CFA model

VARCLUS example 3: Interpersonal Processes of Care

Results: look at handout pages 9-11

I chose a 31-cluster VARCLUS solution, retaining 64 of 79 items

many singleton clusters and some 'rouge' items were dropped

The 1st- and 2nd-order CFA models suggested acceptable fit

The entire process took *one afternoon*: cf. EFA taking > 1 year, 29 items retained

→ Fitting the VARCLUS model

→ Selecting the 31 cluster solution—keeping 18 clusters w/ 64 items

→ Fitting a 1st-order CFA model

I just fit one, the fit was good—no model modifications via CFA

→ Creating cluster scale scores

→ Examining various EFA and VARCLUS models of 2nd-order structure

→ Choosing 2nd-order factor model—keeping 16 of the 18 1st-order factors

→ Fitting a 2nd-order CFA model

I made one modification to the model within CFA to deal with a

residual variance estimate with a small, negative value

VARCLUS example 3: Interpersonal Processes of Care

Still to do…

Unstack data & test measurement invariance across the 4 r/e/language groups

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;
```

VARCLUS: other considerations

2. Consider a transformation prior to VARCLUS
   Blom transformation provides a convenient option

```
proc rank normal=blom data=<indata> out=<outdata>;
  var <varlist>;
run;

data <outdata>;
 set <outdata>;

  label
  <var1> = '<var1 label>'
    …
  <var#> = 'var# label>;

run;
```

VARCLUS: other considerations

3. Just to make myself feel better,
   The following 2 sets of code will produce the same 2 'cluster' solution

(i)

```
proc varclus cov maxc=2 maxsearch=0 maxiter=1 data=<data> ;
  var <varlist>;
  run;
```

where
   MAXSEARCH=0 and MAXITER=1 omit Step 2 of VARCLUS, and

(ii)

```
proc factor cov  n=2 norm=cov m=prin r=oblimin(0) data=<data>;
  var <varlist>;
run;
```

where
   M=PRIN and COV request principal components of the covariance matrix, &
      NORM=COV and R=OBLIMIN(0) request 'raw' oblique QUARTIMAX

Summary: VARCLUS for 1st-order measurement models

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 v EFA

I didn't really provide a head-to-head comparison between VARCLUS and EFA
for the purpose of uncovering a measurement model, that included both
(i) deciding how many item-clusters or factors to retain &
(ii) deciding which items to retain

Unclear which model selection procedure to choose w/in the EFA framework
There are no established guidelines

If I choose one or more mechanized candidates, then someone will disagree

EFA with large item sets can be a real mess

This contention of mine is probably best demonstrated by
. a VARCLUS v EFA 'competition,' and
. reliability of subjective model selection via VARCLUS v EFA

Summary: VARCLUS & EFA: complementary

I prefer VARCLUS to identify 1st-order structure &
EFA to identify 2nd-order structure (somewhat to my surprise)

For exploration of the 2nd-order model,
I use VARCLUS results for the 1st-order structure
to guide creation of cluster scale scores and
use the scale scores as input data to EFA

Often 2nd-order structures are complex (cross-loadings)
EFA gives better insights into this than does VARCLUS

VARCLUS will output inter-cluster correlations, but not covariances.
which is why I resorted to using cluster scale scores as input to EFA

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 if you compare it to personal experiences with CFA after EFA

YMMV: subjective judgment of VARCLUS output plays a part

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. That takes less than an 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