

Introduction to SAS PROC VARCLUS:

A (mostly) superior alternative
to Exploratory Factor Analysis

University of California San Francisco
Center for AIDS Prevention Studies
Methods Core & Visiting Professors Program

April 13, 2021

Steve Gregorich

Overview

- . 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

Conceptually, what is a **common factor model**?

Indirect observation

Some constructs are not directly observable
attitudes, intelligence, economic strength, top quark

Indirectly measured constructs are sometimes called *latent* variables
. Latent variables are 'everywhere' (physics, medicine, economics)

Latent variables often are identified 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**

What does 'common'
mean?

What are the goals of common factor model?

Assess a form of validity, i.e., **construct validity**.

Do the items measure the hypothesized constructs?

Represent a set of observed variables (or items)
by a more parsimonious set of related constructs
(AKA common factors, latent variables)

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

The common factor model is >115 years old (Spearman 1904)

Common factor model: Conceptual example

Suppose I want to measure two dimensions of consumer confidence

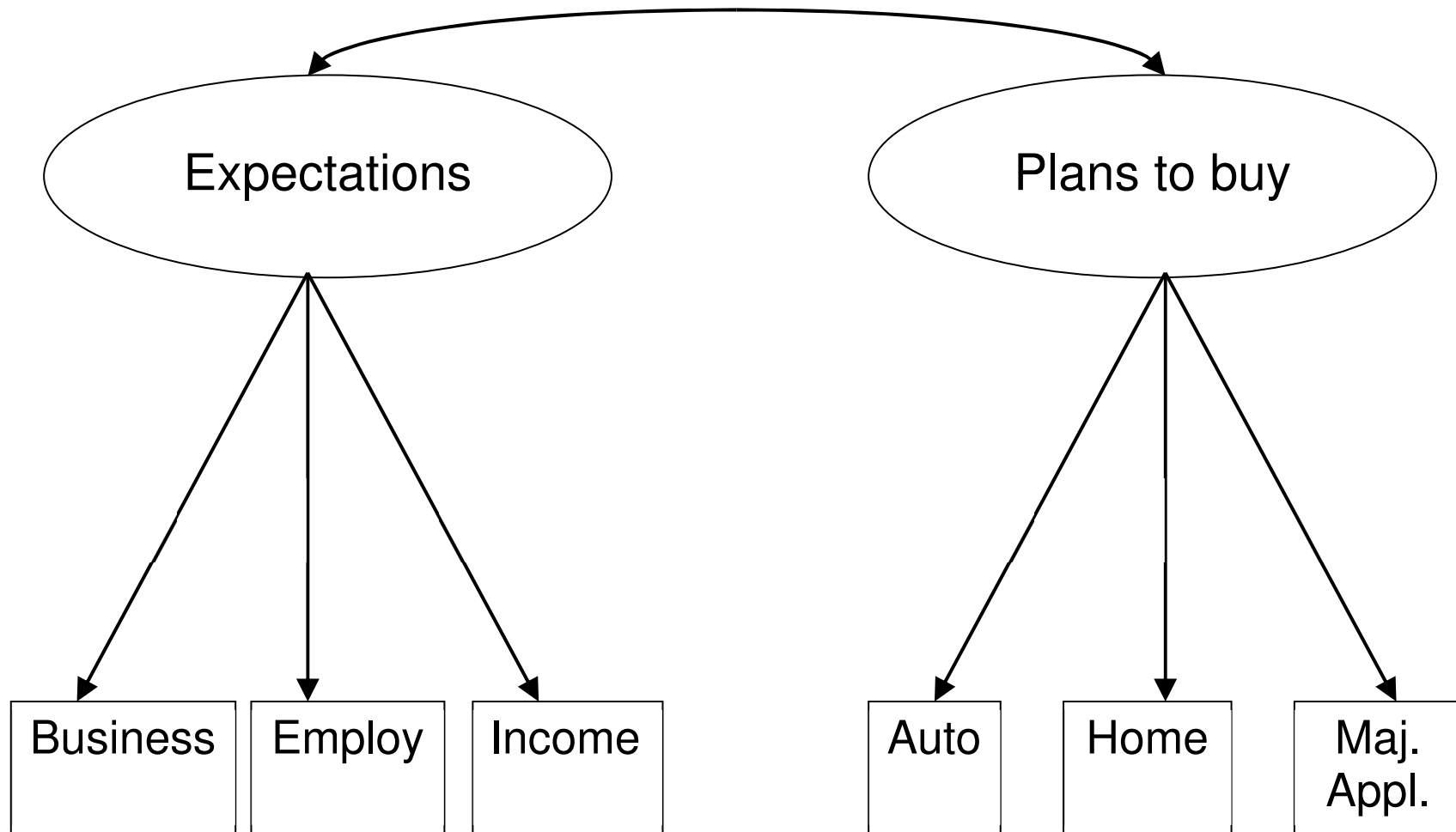
Expectations for 6-months hence

- . 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 configuration



(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 scaling is arbitrary

Correlations between latent and manifest variables aid interpretation

Q: Is the interpretation consistent with the motivating hypotheses?

Again: Implications of empirical support for a common factor model

Demonstration of construct validity:

Do the items measure what they are hypothesized to measure?

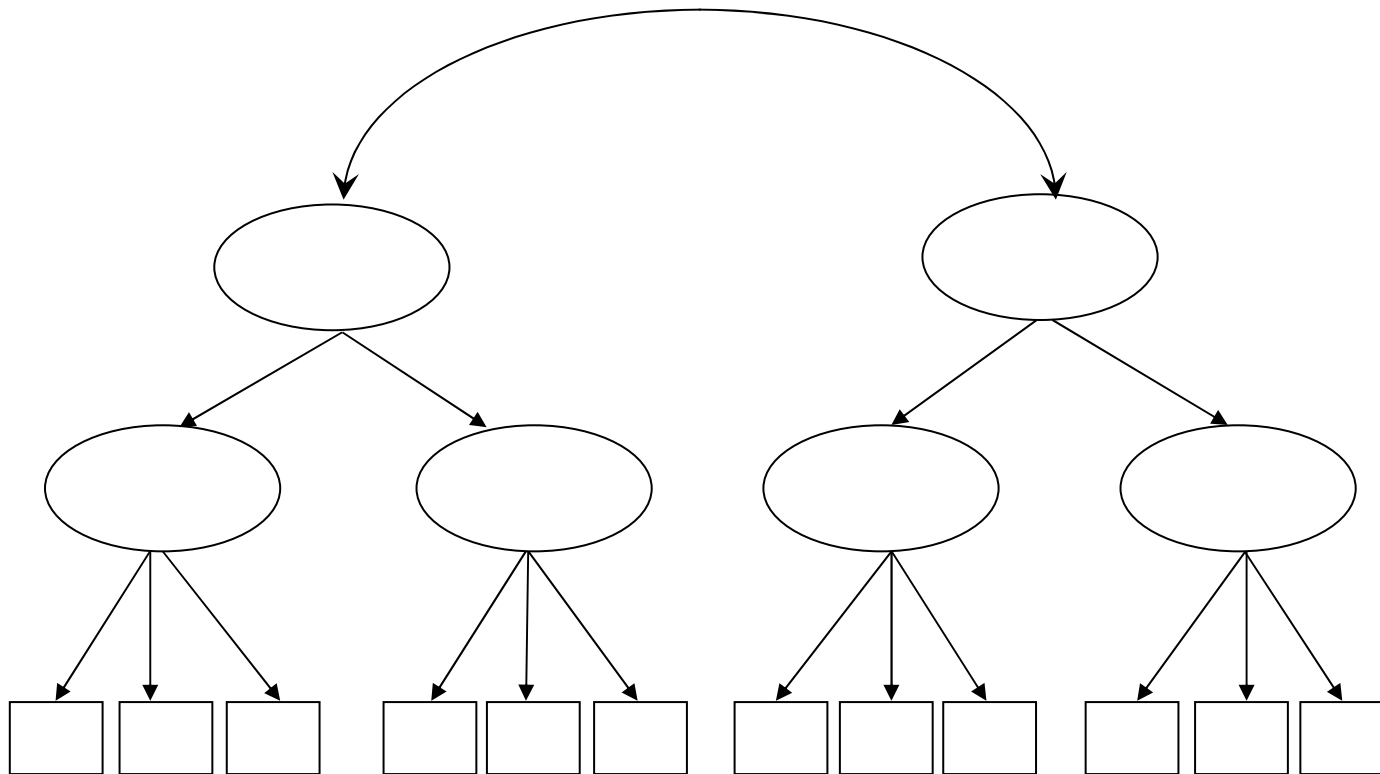
More parsimonious representation of information captured by items

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

Higher-order factor models

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

Second- or higher-order factor models are possible



Optimally, the configuration of a common factor (AKA measurement) model is specified a priori

The configuration specifies a set of common factors (latent variables). Each is hypothesized to cause responses to specific subsets of items (manifest variables).

It should be based upon theory, or previous empirical findings

But...

- . What if the hypothesized measurement model isn't supported?
- . Or, what if there is no a priori measurement model to test?

One option...

- . Exploratory factor analysis (EFA).

The goal of EFA is to uncover the measurement model


For an introduction to EFA, see my March 2, 2021 talk in this series

Pitfalls of EFA

Can work well if you are 'lucky'

Difficulties often arise w/ large item sets and large numbers of factors

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
 - . Many have sought a 'holy grail' EFA rotation method—it doesn't exist
 - . Personal example: The Interpersonal Processes of Care (IPC) measure
EFA modeling of 79 initial items
winnowed down to 28 items in over 1 year
 - . SAS PROC VARCLUS can help circumvent EFA frustrations
- 

VARCLUS:

Oblique Principal Components Cluster Analysis

Where did it come from? SAS isn't saying...

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 that cluster* and
rotate via raw oblique QUARTIMAX
Thus, the targeted cluster is split to form two clusters

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

Repeat Steps 1 and 2 until stopping rule satisfied

Notes

- . Working with principal components not principal factors
- . Even so, the process produces correlated item clusters
- . SAS, R, and Python implementations are available

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 (PubMed) found 15 articles—
all applications of VARCLUS

There are more...VARCLUS isn't always mentioned in the abstract.

Even so...

VARCLUS

VARCLUS code is simple

```
proc varclus data=<data> cov minclusters=1 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 =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 use 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 ordinal response options

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 clusters

 - Mostly, I rely on subjective judgment based upon item wording

 - I also consider 'R-square with own cluster.'

 - I like to see values ≥ 0.50

 - I don't pay much attention to

 - 'R-square with Next Closest (cluster)' or '1=R**2 Ratio'

VARCLUS also reports 'Proportion of Variance Explained by Clusters'

 - I like to see a value ≥ 0.70

I chose the 16-cluster solution, output the inter-cluster correlation matrix, input that matrix into VARCLUS, and extracted five 2nd-order clusters

Note

- Using VARCLUS requires detailed variable labels

- All 42 items retained in final solution

- The five 2nd-order clusters differ from the five 1st-order clusters (p. 4)

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., $174 \times 2 = 348$ cases
- . Use VARCLUS to identify 1st-order factors
- . Reshape data to represent 174 data records, one per couple
- . Fit 1st-order CFA model and
 - . Further probe the adequacy (fit) of the VARCLUS model
 - . Test invariance of model parameters across moms/dads
- . Explore 2nd-order factor structure via VARCLUS and EFA
- . Fit 2nd-order CFA model and
 - . Further probe the adequacy (fit) of the 2nd-order model
 - . Test invariance of model parameters across moms/dads

VARCLUS example #2: Parental feeding practices

Look over the handout, pages 7 & 8

I chose the 14-cluster VARCLUS solution

2 items dropped based upon VARCLUS results (#23 and #29)

3 more items dropped during preliminary CFA modeling (#58, 67, 68)

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 (I dropped the other 1st-order clusters)

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., via EFA, I was able to retain about 50 of 68 items

Tschann, J.M., Gregorich, S.E., Penilla, C., Pasch, L.A., de Groat, C.L., Flores, E., Deardorff, J., Greenspan, L.C., and Butte, N.F (2013). Parental feeding practices in Mexican American families: Initial test of an expanded measure. *International Journal of Behavioral Nutrition and Physical Activity*, 10:6

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 San Francisco

$N=1622$ 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 CFA model and test invariance across r/e/language

Result

The EFA 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.

Much later, I learned about VARCLUS and re-analyzed the data

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: 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;
```

Versus

. 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 transforming item responses prior to VARCLUS

Blom transformation provides a convenient option (PROC RANK)

PROC RANK will strip variable labels. You'll need to add them back.

```
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

```
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

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

where

`m=prin` and `cov` request principal components of the covar matrix, &
`norm=cov` and `r=oblmin(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 re. the conceptual 'purity' of candidate clusters is likely the best initial guide.

Singleton clusters are OK,

you can choose to drop them from consideration

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 versus EFA

With 'large' item sets, I prefer VARCLUS to identify 1st-order factors

VARCLUS can work well to identify 2nd-order factors, but only if the 2nd-order factors have a simple structure

2nd-order factor structures can be complex (cross-loadings)

In those cases, EFA is more flexible than VARCLUS @ 2nd-order

VARCLUS will output inter-cluster correlations, but not covariances.

You can choose to calculate 1st-order cluster scale scores

& input them into EFA, allowing a covariance-based 2nd-order 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

Many times, I have fit a CFA model to help defend a VARCLUS model.

Each time, the fit has been good w/ little to no model modification

Given the size of the item sets described here (i.e., 42, 68, 79),
that success rate is surprising (compared w/ CFA after EFA)

YMMV: subjective judgment of VARCLUS output is key

Summary

Before I used VARCLUS, the prospect of exploring a large item set with EFA was daunting

There is no way to estimate the amount of time, effort, and frustration required to complete such an EFA modeling task

With VARCLUS, I can consistently complete initial analyses through completed CFA (with good model fit!) in an afternoon

And, in my experience, VARCLUS tends to retain more items than EFA

EFA remains a useful tool, but I believe that most applications of EFA would be better served by instead applying VARCLUS.

Thank you.