Introduction to Likelihood Methods for SEM

Jarrett E. K. Byrnes
University of Massachusetts Boston

\[ \Sigma = \Sigma(\Theta) \]

Why the Likelihood Approach?

- Piecewise fitting is flexible, but, does not accommodate full range of SEM complexity
- No latent variables
- No non-recursive relationships
- Cannot compare wide range of competing models simply

A Likely Outline

1. What is different about fitting using likelihood and covariance matrices?
2. Identifiability
3. Introduction to lavaan

How does ML Estimation Work?

Hypothesized Model

\[ \begin{pmatrix} \hat{\theta} \end{pmatrix} \]

\[ S = \begin{pmatrix} 1.3 & .24 & .41 \\ .24 & .01 & 9.7 \\ .41 & 9.7 & 12.3 \end{pmatrix} \]

Corrected Covariance Matrix

Model Fit Evaluations

Parameter Estimates

Implied Covariance Matrix
The Maximum Likelihood Fitting Function

\[ F_{\text{ML}} = \log \left( \hat{\Sigma} \right) + tr \left( \hat{\Sigma}^{-1} \right) - log|S| - (p + q) \]

- \( S \) = Sample covariance matrix
- \( \Sigma \) = Fit covariance matrix
- \( p \) = endogenous variables
- \( q \) = exogenous variables

When \( S \) and \( \Sigma \) are equal, terms 1 and 3 = 0 and terms 2 and 4 = 0.
Thus, perfect model fit yields a value of \( F_{\text{ML}} \) of 0.

Assumptions Behind \( F_{\text{ML}} \)

- Multivariate normality
  - Test with multivariate Shapiro-Wilk’s Test (library mvnormtest)
  - In particular, no skew
  - Violations biases parameter error and tests of model fit
- No missing data in calculation of \( S \)
  - Will bias your estimates
- No redundant variables
  - \( S \) must be positive definite
- Sample size is large
  - Advice varies – take home, get as much data as possible
  - 10-20 samples per param, \( n=p^{3/2} \), etc...

Why FML? Alternatives?

\( F_{\text{FML}} \) is unbiased, scale invariant, best estimator

\( F_{\text{GLS}} = 0.5 \times tr[(S - \Sigma(\Theta))^2] \)
- A.K.A. the ULS criterion
- Least squares!
- Sensitive to scale of variables

\( F_{\text{WLS}} = 0.5 \times tr[(S - \Sigma(\Theta)W^{-1})^2] \)
- A.K.A. the ADF criterion – no normality assumption
- Weighted: flexible
- Scale free
- Asymptotically \( \chi^2 \) distributed
- Sensitive to fat or thin tailed data
- Sensitive to sample size (\( n>1000 \))

A Likely Outline

1. What is different about fitting using likelihood and covariance matrices?
2. Identifiability
3. Introduction to lavaan
Identifiability Revisited
1. For the model parameters to be estimated with unique values, they must be identified. We need as many known pieces of information as we do unknown parameters.

2. Several factors can prevent identification, including:
   a. too many paths
   b. certain kinds of model specifications can make parameters unidentified
   c. multicollinearity
   d. combination of a complex model and a small sample

3. Some software checks for identification (in something called the information matrix) and lets you know which parameters are not identified.

Whither the T-Rule
# of Parameters v. Covariance Matrix

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>y1</th>
<th>y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>∆x</td>
<td>y1</td>
<td>y2</td>
</tr>
<tr>
<td>y1</td>
<td>∆y</td>
<td>x1</td>
<td>0.5</td>
</tr>
<tr>
<td>y2</td>
<td>β1</td>
<td>0.7</td>
<td>0.5</td>
</tr>
<tr>
<td>Cov(x,y1,y2)=</td>
<td>y1</td>
<td>0.7</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>y2</td>
<td>0.2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

- # Parameters ≤ # Unique Entries in a Covariance Matrix
- Necessary condition for identification
- T-rule: t ≤ (p+q)(p+q+1)/2
- t=# params, p=# endogenous variables, q=# exogenous variables
- Model DF= min (p, q)

How Do I Count the Number of Parameters?

Yes, there is a variance here

If variance and covariances among exogenous variables is not shown
T-rule: t ≤ (p+q)(p+q+1)/2 - q(q+1)/2

You will see path diagrams drawn many ways...

Check what researcher is doing with exogenous variables!
DF of all of these models = 4*5/2 - 8 = 2
Identification in SEM

*# of Parameters v. Covariance Matrix*

\[
\begin{array}{ccc}
\hat{x} & \hat{y} & y_1 \\
\hat{y} & \rho_{y2} & \hat{z}_2 \\
\end{array}
\]

Cov(x,y1,y2) =
\[
\begin{array}{ccc}
y_1 & 0.7 & 0.5 \\
y_2 & 0.2 & 0.8 & 0.3 \\
\end{array}
\]

Estimating 5 parameters from 6 variance/covariance relationships

**DF=1**
Model is Overidentified

Identification in SEM

*Is this model identified?*

Yes: There are no relationships between endogenous variables

**SUFFICIENT CONDITION**

Identification in SEM

*Is this model identified?*

Yes: Model is Recursive

**SUFFICIENT CONDITION**
Identification in SEM

**Is this model identified?**

YES: Model is Non-recursive, but y's have unique information

**NECESSARY CONDITION**

NO! Model is Non-recursive

AND not enough information for unique solution

Identification in SEM

**The Order Condition**

- G = # incoming paths
- H = # of exogenous vars + # indirectly connected endogenous vars
- G ≤ H
- **NECESSARY CONDITION**

Identification in SEM

**Is this model identified?**

NO! Everything that affects y2 affects y1 – Fails Rank Test

**SUFFICIENT CONDITION**
Rules of Identification

- $t \leq (p+q)(p+q+1)/2$
  - necessary
- No paths between endogenous variables
  - sufficient
- Model is recursive
  - sufficient
- Order: $G \leq H$
  - necessary
- Rank: see Kline 2005 or Bollen 1989
  - Endogenous vars must be affected uniquely
  - sufficient

Packages for Fitting an SEM in R

- lavaan – http://www.lavaan.org
  - Initially based on Mplus
  - Similar to R linear model syntax
- sem – http://socserv.socsci.mcmaster.ca/fox/Misc/sem/
  - RAM & equation syntax
  - good documentation
  - restricted, not always able to fit models
- openMx - http://openmx.psyc.virginia.edu/
  - Based off of Mx
  - Incredibly flexible, powerful
- RECS - http://www.mvsoft.com/
  - Interface for commercial EQS software
  - Flexible, developed by SEM statisticians

A Likely Outline

1. What is different about fitting using likelihood and covariance matrices?
2. Identifiability
3. Introduction to lavaan

Extensions off of lavaan

- lavaan.survey – http://www.lavaan.org
  - corrects tests based on complex survey design
  - Accounts for clustering/nesting using survey package
  - grab bag of additional tools
  - imputation and fit index tools
  - multivariate skew and kurtosis indices
  - Creation of new variables
  - and more! Constantly growing!
- semPlot - https://github.com/SachaEpskamp/semPlot/
  - plots fit models
  - can also be used to plot a piecewise SEM
  - works using ggplot2
**Lavaan**: A Package for Fitting SEMs in R Using Covariance Matrix Methods

1. *SOFTWARE IS A TOOL*

2. *IT IS NOT PERFECT*

3. *ALWAYS MAKE SURE IT IS DOING WHAT YOU THINK IT IS DOING!*

**Mediation in Analysis of Post-Fire Recovery of Plant Communities in California Shrublands***

*Five year study of wildfires in Southern California in 1993. 90 plots (20 x 50m), (data from Jon Keeley et al.)*

Analysis focus: understand post-fire recovery of plant species richness

- measured vegetation recovery:
  - plant cover
  - species richness

Examination of woody remains allowed for estimate of age of stand that burned as well as severity of the fires.

Other factors measured included:
- local abiotic conditions (aspect, soils)
- spatial heterogeneity
- landscape-level conditions (location, elevation)
Coding a Regression versus SEM

```r
#regression
aLM<-lm(cover ~ age, data=keeley)

#sem
library(lavaan)
aSEM<-sem('cover ~ age', data=keeley)
```

Compare to Regression

```r
summary(aLM)
```

```r
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.917395 0.071726 12.79 < 2e-16 ***
age -0.008846 0.002520 -3.51 0.00071 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2988 on 88 degrees of freedom

But what about the intercept?
```

Intercepts Estimated with Mean Structure

```r
> aMeanSEM<-sem('cover ~ age', data=keeley, meanstructure=T)
>
> summary(aMeanSEM)
```

```r
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.917395 0.071726 12.79 < 2e-16 ***
age -0.008846 0.002520 -3.51 0.00071 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2988 on 88 degrees of freedom
```

The model converged!

Model is saturated so, $\chi^2$ test has no df
Standardized Coefficients

| lhs op rhs est.std se z pvalue |
|------------------------|--------|--------|--------|------------------|
| 1 cover ~ age          | -0.350| 0.099  | -3.549| 0                |
| 2 cover ~ cover        | 0.877 | 0.131  | 6.708 | 0                |
| 3 age ~ age            | 1.000 | NA     | NA    | NA               |

Also: summary(aSEM, standardized=T, rsq=T)

Bringing Back Fire

```
partialMedModel<- 'firesev ~ age
cover ~ firesev + age'
partialMedSEM<-sem(partialMedModel, data=keeley)
```

```
summary(partialMedSEM, rsquare=T, standardized=T)
```

```
semPlot
```

```
library(semPlot)
semPaths(partialMedSEM, "std")
```
What if we know better?

```
zeroMedModel<- 'firesev ~ 0*age
cover ~ 0*firesev + age'
zeroMedFit<-sem(zeroMedModel, data=keeley)
```

standardizedSolution(zeroMedFit)

```
lhs op rhs est.std se z pvalue
1 firesev ~ age 0.000 NA NA NA
2 cover ~ firesev 0.000 NA NA NA
3 cover ~ age -0.350 0.099 -3.549 0
4 firesev ~ firesev 1.000 0.149 6.708 0
5 cover ~ cover 0.877 0.131 6.708 0
6 age ~ age 1.000 NA NA NA
```

What about Correlated Error?

```
#what about correlations
corModel <- 'firesev ~ age
cover ~ age
cover -- firesev'
corFit <- sem(corModel, data=keeley)
```

What about Correlated Error?

```
> standardizedSolution(corFit)
lhs op rhs est.std se z pvalue
1 firesev ~ age 0.454 0.094 4.832 0
2 cover ~ age -0.350 0.099 -3.549 0
3 firesev ~ cover -0.333 0.094 -3.556 0
4 firesev ~ firesev 0.794 0.118 6.708 0
5 cover ~ cover 0.877 0.131 6.708 0
6 age ~ age 1.000 NA NA NA
```
Morning Exercise

1. Fill in Standardized Coefficients and $R^2$ for this model
2. Refit is assuming that there is a 1:1 relationship between hetero and richness

Solution 1: The Model

```r
# The Richness Partial Mediation Model
distModel <- 'rich ~ distance + abiotic + hetero
               hetero ~ distance
               abiotic ~ distance'

distFit <- sem(distModel, data=keeley)
```

<table>
<thead>
<tr>
<th>lhs</th>
<th>op</th>
<th>rhs</th>
<th>est</th>
<th>std</th>
<th>se</th>
<th>z</th>
<th>pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>rich</td>
<td>~</td>
<td>distance</td>
<td>0.377</td>
<td>0.092</td>
<td>4.117</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>rich</td>
<td>~</td>
<td>abiotic</td>
<td>0.268</td>
<td>0.087</td>
<td>3.079</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>rich</td>
<td>~</td>
<td>hetero</td>
<td>0.256</td>
<td>0.082</td>
<td>3.104</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>hetero</td>
<td>~</td>
<td>distance</td>
<td>0.346</td>
<td>0.099</td>
<td>3.498</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>abiotic</td>
<td>~</td>
<td>distance</td>
<td>0.460</td>
<td>0.094</td>
<td>4.911</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>rich</td>
<td>~</td>
<td>rich</td>
<td>0.539</td>
<td>0.080</td>
<td>6.708</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>hetero</td>
<td>~</td>
<td>hetero</td>
<td>0.880</td>
<td>0.131</td>
<td>6.708</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>abiotic</td>
<td>~</td>
<td>abiotic</td>
<td>0.789</td>
<td>0.118</td>
<td>6.708</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>distance</td>
<td>~</td>
<td>distance</td>
<td>1.000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>
Solution 2: The Model

oneDistModel <- 'rich ~ distance + abiotic + 1*hetero
hetero ~ distance
abiotic ~ distance'

oneFit<-sem(oneDistModel, data=keeley)
summary(oneFit, standardized=T, rsquare=T)

Estimating Direct and Indirect Effects with Named Coefficients

totDistModel <- '
rich ~ a*distance + b*abiotic + c*hetero
hetero ~ b1*distance
abiotic ~ c1*distance
direct:= a
indirect:= b1*b + c1*c
total:= b1*b + c1*c + a
'

summary(totDistFit, standardized=T, rsquare=T)

standardizedSolution(totDistFit)

lhs op rhs est.std se z pvalue
1 rich ~ distance 0.377 0.092 4.117 0.000
2 rich ~ abiotic 0.268 0.087 3.079 0.002
3 rich ~ hetero 0.256 0.082 3.104 0.000
4 hetero ~ distance 0.346 0.099 3.498 0.000
5 abiotic ~ distance 0.460 0.094 4.911 0.000
6 rich ~ rich 0.539 0.080 6.708 0.000
7 hetero ~ hetero 0.880 0.131 6.708 0.000
8 abiotic ~ abiotic 0.789 0.118 6.708 0.000
9 distance ~ distance 1.000 NA NA NA
10 direct := a 0.377 NA NA NA
11 indirect := b1*b + c1*c 0.210 NA NA NA
12 total := b1*b + c1*c + a 0.588 NA NA NA