Unbelievably fast estimation of multilevel structural equation models

Joshua N. Pritikin

Department of Psychology
University of Virginia

Spring 2016
Abstract

The challenge of quick optimization of multilevel structural equation models (SEM) will be introduced. To provide context, multilevel SEM will be compared with the mixed model. Rampart, a novel method to simplify the multilevel SEM likelihood will be introduced, inspired by the fact that the multivariate normal density is transparent to orthogonal rotation. Assumptions and limitation of Rampart will be discussed.
Acknowledgment

This research was aided by

- Timo von Oertzen & Steve Boker (University of Virginia)
- Tim Brick (Pennsylvania State University)
- OpenMx development team
Structural equation models
Multilevel structural equation models
A hypothetical example
Teacher

Diagram:

- Skill
- Years of education
- Self-rated friendliness
- Friendliness
- Academics
- Behavior

Arrows:
- s1
- s2: VAR_yoe
- s3
- VAR_yox
- VAR_srf
- f1
- f2

Numbers:
- 1
Estimation time?

How long will this model take to estimate?
Roadmap

- What is hard about multilevel?
- mixed model + SEM = Relational SEM
- Rampart (a novel method that favorably transforms the problem)
Direct sum

\[ B_1 \bigoplus B_2 = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} \]

\[ \bigoplus_{i=1}^{k} B_i = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & B_k \end{pmatrix} . \]

- \( \bigoplus \) stacks matrices in a block-diagonal arrangement
- Here we assume *nested multilevel* structure
Covariance

Suppose we build a covariance model $S$ for a particular student. A classroom of $s$ students will have covariance matrix

$$T = \begin{pmatrix} T_{1,1} & T_{1,2} \\ T_{2,1} & \bigoplus_{i=1}^{s} S_i \end{pmatrix}.$$
Sparseness pattern
Bottleneck in model evaluation

- Covariance matrix can become very large
- Inverting the covariance is expensive, $O(N^3)$
- Sparse matrix operations help, but not enough
- Conclusion: impractical without cleverness
Structural equation model

What is a SEM?

It’s similar to regression. But how?
Structural equation model, 1st moment

Regression is $y = X\beta + e$
SEM is $y = X\beta + e$

For $y$ observations, $X$ covariates/predictors, $\beta$ constant coefficients, $e$ residuals
Structural equation model, 2nd moment

Regression

\[ y = X\beta + e \]
\[ e \sim \mathcal{N}(., \sigma^2 I) \]

SEM

\[ y = X\beta + e \]
\[ (X, e) \sim \mathcal{N}(., \Sigma(\theta)) \]

For \( y \) observations, \( X \) covariates/predictors, \( \beta \) constant coefficients, \( e \) residuals, variance \( \sigma^2 \), parameters \( \theta \), covariance \( \Sigma \)
Mixed model

lme4 (Bates, Mächler, Bolker, & Walker, 2015)

Multilevel or crossed regression

Fast

How does it work?
Mixed model, 1st moment

\[ Y = \underbrace{X\beta}_{\text{constant}} + \underbrace{Zu + e}_{\text{varying}}. \]

- column vector of observations \( Y \)
- covariates \( X \) associated with constant coefficients \( \beta \)
- covariates \( Z \) associated with varying coefficients \( u \)
- column vector of residuals \( e \)

\[ \mathbb{E}\left( \begin{pmatrix} u \\ e \end{pmatrix} \right) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \]
Mixed model, 2nd moment

\[ Y = X\beta + Zu + e. \]

Constant varying

\[ \text{Cov} \begin{pmatrix} u \\ e \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \]
Mixed model, unconditional distribution

\[ Y = X\beta + e \]

where

\[ e \sim \mathcal{N}(0, ZGZ^T + R) \]

The formulation I showed earlier is actually conditional on a particular realization of the varying coefficients \( u \).
Mixed model, example

\texttt{lmer(gpa \sim 1 + (1 \mid school) + (1 \mid school:teacher), ...)}

- Intercept-only model with 3 levels
- The expression after the vertical bar indicates the partitioning design (like conditional probability)
- 4 free parameters are estimated: the grand intercept; 2 variances, one for each varying coefficient; and the residual variance
- Efficient
How does this look in SEM?
What is $\Join$?

Let $R$ and $S$ be tables (or data frames) that contain rows.

$$R \Join (F) S \equiv \{ r \cup s \land r \in R \land s \in S \land F(r \cup s) \}$$

where $F$ is a boolean valued function.

Without loss of generality, here $F$ is whether primary and foreign keys match. We will omit $F$ and write $\Join (k)$ where $k$ is the name of the key.
What is $\bowtie$? Example

<table>
<thead>
<tr>
<th>Employee</th>
<th>Dept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harry</td>
<td>Sales</td>
</tr>
<tr>
<td>Sally</td>
<td>Finance</td>
</tr>
<tr>
<td>George</td>
<td>Finance</td>
</tr>
<tr>
<td>Harriet</td>
<td>Sales</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dept</th>
<th>Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>George</td>
</tr>
<tr>
<td>Finance</td>
<td>Harriet</td>
</tr>
<tr>
<td>Production</td>
<td>Charles</td>
</tr>
</tbody>
</table>

Employee $\bowtie$(Dept) Manager

<table>
<thead>
<tr>
<th>Employee</th>
<th>Dept</th>
<th>Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harry</td>
<td>Sales</td>
<td>George</td>
</tr>
<tr>
<td>Sally</td>
<td>Finance</td>
<td>Harriet</td>
</tr>
<tr>
<td>George</td>
<td>Sales</td>
<td>George</td>
</tr>
<tr>
<td>Harriet</td>
<td>Finance</td>
<td>Harriet</td>
</tr>
</tbody>
</table>
Which is easier to understand?

Time point $t$, individual $i$, cluster $j$.

$y_{tij}$ : individual-level, outcome variable
$a_{1tij}$ : individual-level, time-related variable (age, grade)
$a_{2tij}$ : individual-level, time-varying covariate
$x_{ij}$ : individual-level, time-invariant covariate
$w_j$ : cluster-level covariate

Three-level analysis (Mplus considers Within and Between)

$$ Level 1 \ (Within) \ : \ y_{tij} = \pi_{0ij} + \pi_{1ij} a_{1tij} + \pi_{2tij} a_{2tij} + e_{tij}. \quad (1) $$

$$\begin{align*}
\pi_{0ij} &= \beta_{00j} + \beta_{01j} x_{ij} + r_{0ij}, \\
\pi_{1ij} &= \beta_{10j} + \beta_{11j} x_{ij} + r_{1ij}, \\
\pi_{2tij} &= \beta_{20j} + \beta_{21tij} x_{ij} + r_{2tij}.
\end{align*} \quad (2)$$

$$ Level 3 \ (Between) \ : \ \begin{align*}
\beta_{00j} &= \gamma_{000} + \gamma_{001} w_j + u_{00j}, \\
\beta_{10j} &= \gamma_{100} + \gamma_{101} w_j + u_{10j}, \\
\beta_{20j} &= \gamma_{200} + \gamma_{201} w_j + u_{20j}, \\
\beta_{01j} &= \gamma_{010} + \gamma_{011} w_j + u_{01j}, \\
\beta_{11j} &= \gamma_{110} + \gamma_{111} w_j + u_{11j}, \\
\beta_{21j} &= \gamma_{210} + \gamma_{211} w_j + u_{21j}.
\end{align*} \quad (3)$$
A relational schema
Autoregressive tree (pedigree)

- personID
- motherID
- fatherID

- height_VAR
- height
- m

- isMale
- male effect

- fromMother
- fromFather
Summary: □ vs conditional probability

Data is joined (□)

Conditional probability is an ingredient in multilevel models, not data

Join commutes, conditional probability doesn’t
Mixed model

Limitations:

▶ missing data $\rightarrow$ row-wise deletion
▶ only the lowest level unit has observations
▶ multivariate (more than one outcome) is very awkward
Sufficient statistic approach

Suppose we have data of $N$ independent observations of $K$-variate units. Let $\mu$ and $\Sigma$ be the model expected mean vector and covariance matrix, respectively. Let $m$ and $S$ be the mean vector and covariance matrix of the data, respectively.

$$-2 \log L(\text{data}|\theta) = N(K \log(2\pi) + \log(|\Sigma|) + \text{tr}(\Sigma^{-1}S) + \mu^T \Sigma^{-1}(\mu - 2m))$$

Maximum covariance dimension is $K$. 
Sparseness pattern
Uncompressed likelihood

Suppose we have $N$ observations consisting of data vector $x$. Let $\mu$ and $\Sigma$ be the model expected mean vector and covariance matrix, respectively.

$$-2 \log L(\text{data}|\theta) = N \log(2\pi) + \log(|\Sigma|) + (\mu - x)^T \Sigma^{-1} (\mu - x)$$

Maximum covariance dimension is $N$. 
Covariance becomes very large. What to do?

SEMs are specified using the RAM parameterization:

\[ \mu = F(I - A)^{-1}M \]
\[ \Sigma = F(I - A)^{-1}S(I - A)^{-T}F^T \]

\( A, S, F, \) and \( M \) are used for what?
RAM’s $A$ matrix

teacher model

student model

student model

student model

student model
RAM’s $A$ matrix

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Orthogonal or axis rotation
Call forth the **sublime** orthogonal rotation
Rampart transformed

- teacher model
- student model
- student model
- student model
- student model
Rampart transformed

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S1</td>
<td>$\sqrt{4}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Rampart transformed

- Teacher model
  - Student model
  - Student model
Rampart algebraically

- Upper level has $K$ outgoing regressions
- Lower level has $M$ incoming regressions with data $D$
- Find an orthogonal matrix $Q \in \mathbb{R}^{M \times M}$ such that the lower $M - K$ rows of $QA$ are zero.
- Define new model $A'$ as the first $K$ rows of $Q^TA$
- Define new lower level dataset $D' = Q^TD$
- Proceed with optimization as usual
Rampart geometry

\[
\begin{bmatrix}
1.00 & 6.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
1.00 & -1.00 & 5.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
1.00 & -1.00 & -1.00 & 4.00 & 0.00 & 0.00 & 0.00 \\
1.00 & -1.00 & -1.00 & -1.00 & 3.00 & 0.00 & 0.00 \\
1.00 & -1.00 & -1.00 & -1.00 & -1.00 & 2.00 & 0.00 \\
1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & 1.00 \\
1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00 & -1.00
\end{bmatrix}
\]

Use QR decomposition to scale to an orthogonal rotation
Rampart, apply recursively

```
<table>
<thead>
<tr>
<th>district model</th>
</tr>
</thead>
<tbody>
<tr>
<td>school model</td>
</tr>
<tr>
<td>teacher model</td>
</tr>
<tr>
<td>student model</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>school model</th>
</tr>
</thead>
<tbody>
<tr>
<td>teacher model</td>
</tr>
<tr>
<td>student model</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>student model</th>
</tr>
</thead>
<tbody>
<tr>
<td>teacher model</td>
</tr>
<tr>
<td>student model</td>
</tr>
</tbody>
</table>
```

Joshua N. Pritikin
University of Virginia
Rampart historical lineage

Who done it?

- summer of 2012: idea conceived by Timo von Oertzen, Steven M. Boker, and Timothy R. Brick, inspired by von Oertzen and Hackett (submitted)
- spring 2013: prototyped in OpenMx (by me)
- completed predissertation on a different topic
- spring 2016: optimized implementation in OpenMx (again by me)
Does it work?
Conditions

- 11 parameters per level, 2 between level regressions (total 35)
- 1st student indicator was set to missing with 20% probability
- 2 sets of true parameters ($\theta_1$ and $\theta_2$)
- Parameter $\theta_1$: 7 schools, 38 teachers, and 293 students
- Parameter $\theta_2$: 7 schools, 37 teachers, and 296 students
- 200 Monte Carlo replications for each condition (Algorithm $\times \theta$)
Monte Carlo bias and variance

| $\theta$ | replications | method  | $||\text{bias}||$ | $||\sigma^2||$ |
|----------|--------------|---------|------------------|----------------|
| 1        | 174          | rampart | 1.686            | 0.769          |
|          |              | regular | 1.702            | 0.780          |
| 2        | 171          | rampart | 2.336            | 0.557          |
|          |              | regular | 2.335            | 0.560          |
Scatterplot of deviance at the maximum likelihood
Seconds required per replication

![Histograms showing seconds required per replication for rampart and regular]
OpenMx is a free and open source extension to the R statistical environment.

Software and support available at http://openmx.psyc.virginia.edu/

Questions?
Appendix

Some extra slides follow
Terminology

Historically, coefficients that help predict all observations are called \textit{fixed effects} whereas the other type of coefficient has been called a \textit{random effect}. These are unfortunate terminology. In the statistical literature, there are at least five definitions of these phrases, all of which differ from each other (Gelman, 2005). Moreover, in computer science, the term \textit{random} is usually associated with draws from a uniform random number generator, not synonymous with \textit{stochastic} that does not suppose a particular distribution. Here we follow Gelman (2005) and use the terms \textit{constant} and \textit{varying}. 
Example, lme4

\texttt{lmer(Reaction} ~ Days + (Days | Subject), sleepstudy)
Example, OpenMx (part 1)

```r
bySubj <- mxModel(
  model="bySubj", type="RAM",
  latentVars=c("slope", "intercept"),
  mxData(data.frame(Subject=unique(sleepstudy$Subject)),
    type="raw", primaryKey = "Subject"),
  mxPath(c("intercept", "slope"), arrows=2, values=1),
  mxPath("intercept", "slope", arrows=2,
    values=.25, labels="cov1"))
```
Example, OpenMx (part 2)

```r
ss <- mxModel(
  model="sleep", type="RAM", bySubj,
  manifestVars="Reaction", latentVars = "Days",
  mxData(sleepstudy, type="raw", sort=FALSE),
  mxPath("one", "Reaction", arrows=1, free=TRUE),
  mxPath("one", "Days", arrows=1, free=FALSE,
    labels="data.Days"),
  mxPath("Days", "Reaction", arrows=1, free=TRUE),
  mxPath("Reaction", arrows=2, values=1),
  mxPath(paste0('bySubj.', c('intercept','slope')),
    'Reaction', arrows=1, free=FALSE, values=c(1,NA),
    labels=c(NA, "data.Days"), joinKey="Subject"))
```


