Journey to efficient sampling in multivariate normal latent variable models

Ed Merkle

StanCon 2023

Introduction

Introduction

- Latent variable models overlap with item response models, mixed models, directed acyclic graphs, time series models, and more.
- Efficient estimation strategies are likely to transfer to many other models.

Model overview

- Multivariate models with random effects, where random effects can predict one another.

$$\boldsymbol{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i$$
$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{\eta}_i + \boldsymbol{u}_i$$

$\boldsymbol{\epsilon}_i$, $\boldsymbol{u}_i$ typically multivariate normal

length of $\boldsymbol{\eta}_i$ much smaller than length of $\boldsymbol{y}_i$

Introduction

- Historically, functionality for model estimation has existed in closed source software like LISREL and Mplus
- Around 2010: R packages (lavaan, OpenMx) come online, provide functionality similar to closed source software
- Around 2015: blavaan starts, combining model specification of lavaan with MCMC estimation

Estimation detail

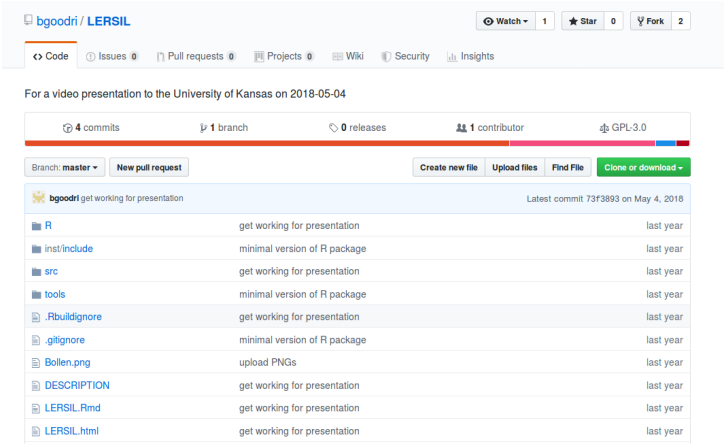
Initial implementation

- Initial steps of blavaan development: estimate the models the way that everyone else estimates the models
  - Sample latent variables ($\boldsymbol{\eta}_i$) as model parameters, so that the model becomes similar to multivariate regression
  - Benefits: univariate likelihoods instead of multivariate likelihoods; ability to model observed variables as non-normal; posterior distributions of latent variables
  - Start with JAGS, try to do a direct translation from JAGS to Stan

Initial implementation

- The initial implementation worked (and continues to work) well for some models. But:
  - Does not work well when we cannot condition away multivariate distributions (e.g., autocorrelated residuals)
  - Does not work well as the number of observations (people) increases (we keep adding more η_i)
  - Sometimes hours to usable results, which makes development a hassle
  - Stan and JAGS exhibit similar efficiency for many models

New implementation



New implementation

- Act more like a frequentist, avoid estimating latent variables as parameters (marginalize likelihood over latent variables)
- If you want posterior distribution of latent variables, sample them in generated quantities via rng functions
- Precompiled model

New implementation

Journal of Statistical Software

November 2021, Volume 100, Issue 6. *doi: 10.18637/jss.v100.i06*

Efficient Bayesian Structural Equation Modeling in **Stan**

Edgar C. Merkle
University of Missouri

Ellen Fitzsimmons
University of Missouri

James Uanhoro
Ohio State University

Ben Goodrich
Columbia University

Recent and ongoing developments

Multilevel SEM

- Multilevel SEM in blavaan (coming soon):
  - Like your usual "students in schools" multilevel model, except each student now has multiple variables, each of which may serve as both a predictor and response OR
  - Like your usual SEM (multiple variables within person), but each person is now clustered in a higher unit like school
  - These models result in multivariate normals of very high dimension, but psychometricians have developed efficient ways to compute the likelihoods

Multilevel SEM

- Example: Say that we observed 100 students in each of 20 schools. Each student is measured on 6 variables, with 1 student latent variable and 1 school latent variable.
  - Original approach: Sample 120 latent variables, so that we can evaluate 12k univariate normal likelihoods.
  - blavaan approach: Evaluate 20 multivariate normals, each of dimension 600. Using psychometrics results from the 1980s/90s, evaluate the 600-dimensional normal by computing inverses/determinants of 6×6 matrices.

Ordinal SEM

- SEM with ordinal variables has been available in blavaan for about 1 year.
  - Chib-Greenberg data augmentation approach.
  - This does not scale to large numbers of observations, because we need to augment more variables as we add extra people.
  - Frequentists have proposed many two-step approaches for handling these models, which may be merged into a single Bayesian model (ongoing work here).

Summary and conclusions

Summary

- Over time, blavaan has continued to improve in sampling efficiency due to the flexibility of Stan.
- This has allowed us to provide reliable estimation methods for relatively complex models.
- And the Stan models provide a starting point for psychometricians to develop new models.

General takeaways

- To improve sampling efficiency in Stan, it can be worthwhile to reconsider the old frequentist literature on estimating complicated models.
- But there is also a tradeoff between efficiency and flexibility: tuning estimation to a focal model, vs using an estimation approach that can be applied to many models.

Acknowledgments

- blavaan has been partially funded by Institute of Education Sciences Grant R305D210044. Contributors and collaborators include
  - Ellen Fitzsimmons, Missouri
  - Mauricio Garnier-Villareal, Amsterdam (Vrije U)
  - Ben Goodrich, Columbia
  - Terrence Jorgensen, Amsterdam (UvA)
  - Yves Rosseel, Ghent
  - James Uanhoro, North Texas

Thank you!

In R:

```
install.packages("blavaan")
```

blavaan website:

https://ecmerkle.github.io/blavaan/

lavaan

- Model specification and maximum likelihood estimation in *lavaan*:

```
library("lavaan")

HS.model <- ' visual  =~ x1 + x2 + x3
              verbal  =~ x4 + x5 + x6 '

fit <- cfa(HS.model, data = HolzingerSwineford1939)
```

blavaan

- Model specification and estimation in *blavaan*:

```
library("blavaan")

HS.model <- ' visual  =~ x1 + x2 + x3
              verbal  =~ x4 + x5 + x6 '

bfit <- bcfa(HS.model, data = HolzingerSwineford1939)
```