BLCFA: An R package for Bayesian Model Modification in Confirmatory Factor Analysis

Lijin Zhang
Sun Yat-sen University, China, zhanglj37@mail2.sysu.edu.cn

Joint work with Junhao Pan
Sun Yat-sen University, China
Edward Haksing Ip
Wake Forest University School of Medicine, United States

14 July 2020

Overview

1. Introduction
   - CFA
   - Post-hoc Model Modification
   - Bayesian CFA
2. Example
3. Simulation Study
   - Comparison Between Modification Index and BLCFA
4. Discussion

Introduction

Confirmatory Factor Analysis

Suppose $y_1, y_2, ..., y_n$ are independent random observations, and each $y_i = (y_{i1}, y_{i2}, ..., y_{ip})^T$ satisfies the following factor analysis model:

$$y_i = \mu + \Lambda\omega_i + \epsilon_i, i = 1, 2, ..., n, \tag{1}$$

- $\mu : p \times 1$ vector of intercepts.
- $\Lambda : p \times q$ factor loading matrix, reflects the relation of observed variables in y_i with the $q \times 1$ latent factors in ω_i.
- $\omega_i \sim N[0, \Phi]$.
- $\epsilon_i : p \times 1$ random vector of measurement errors, $\sim N[0, \Psi]$, independent of ω_i.

Post-hoc Model Modification

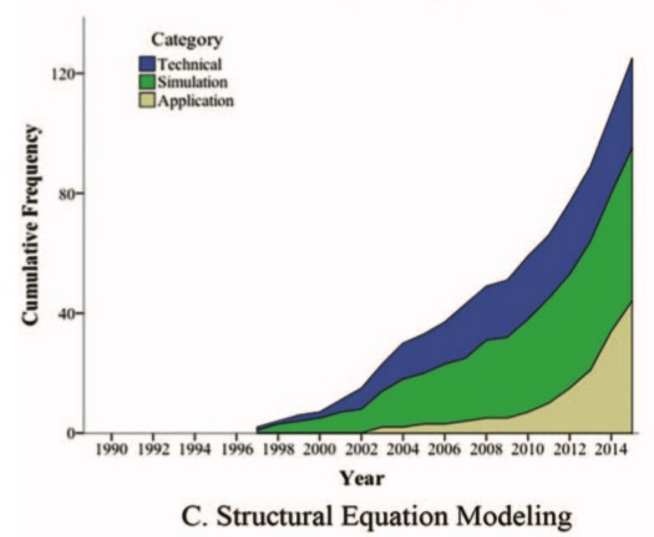
The theory being tested simply does not fit the data well:

- violation of local independence (residual correlations)
- missing cross loadings

Modification Index (Sörbom, 1989)

- the use of modification indexes can be easily influenced by the researchers' subjective choices.
- over-fitting problem.
- parameters must be modified sequentially, causes difficulties in finding the global optimal model (Chou & Bentler, 1990).
- there is no guarantee that the modified covariance matrix is positive definite.

Bayesian SEM (Van de Schoot et al., 2017)



Bayesian CFA (Muthén & Asparouhov, 2012)

Relax the strict constraints in traditional CFA using small variance priors

- Cross-loadings: zero mean, small variance prior (e.g., $N[0, 0.01]$).
- Residual covariances: inverse-Wishart prior ($IW(I, df)$ with $df = p + 6, p =$ number of items, gives a prior standard deviation of 0.1)

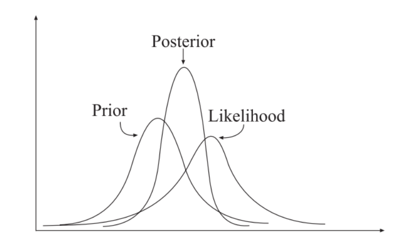


Figure 1. Prior, likelihood, and posterior for a parameter.

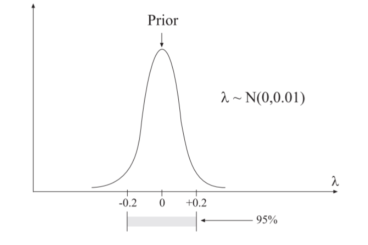


Figure 2. Informative prior for a factor loading parameter.

Bayesian CFA: Lasso Prior

Pan, Ip, and Dubé (2017) proposed a Bayesian Lasso method for deriving a sparse positive definite residual covariance matrix.

The method shrinks weak residual correlations toward zero and detects significant residual correlations simultaneously.

Independent exponential priors and double exponential priors can be assigned for the residual variances and covariances respectively to perform Lasso regularization.

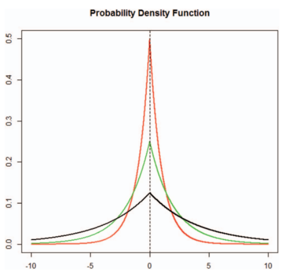


Figure 1. Probability density function of double exponential with three

Bayesian CFA: Lasso Prior

- Detects all the significant residual covariances in one estimation, thus, circumvents the problem of having to handle correlated residual terms sequentially.
- Achieves model parsimony as well as an identifiable model.
- The detection of significant residual covariances can reduce the bias in structural estimates.

Extension of Bayesian Covariance Lasso CFA

Chen, Guo, Zhang and Pan (2020) recently extended the lasso regularization to the loading matrix Λ.

With at least one specified loading per item, a one-step procedure can be applied to figure out both structures Λ, Ψ simultaneously.

The performance of the proposed Bayesian Lasso in estimating the loadings was better than that of ridge priors in terms of BIAS and RMSE.

R Package: BLCFA

To make use of the advantages of Bayesian Lasso CFA in detecting residual covariances and cross-loadings, we propose a two-steps method for model modifications:

- (1) detect significant cross-loadings and/or residual covariances different from zero by Bayesian Lasso CFA;
- (2.1) free the identified significant parameters;
- (2.2) automatically feed the output from (2.1) into Mplus to obtain an appropriately modified CFA model using Maximum likelihood (ML) estimator or Bayesian estimation.

We built an R package named 'blcfa' to facilitate the application of this method.

Example

Example

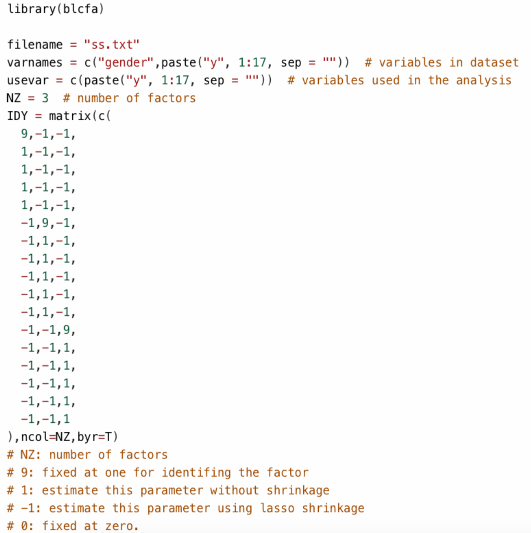
Detailed Illustration: https://github.com/zhanglj37/blcfa

Installation

```
install.packages(”devtools”)
library(devtools)
install_github(”zhanglj37/blcfa”)
```

Detect Cross-loadings and Residual Covariances

Social Support Scale, 5-points Likert scale, 17 items, three factors



```
library(blcfa)

filename = "ss.txt"
varnames = c("gender",paste("y", 1:17, sep = ""))  # variables in dataset
usevar = c(paste("y", 1:17, sep = ""))  # variables used in the analysis
NZ = 3  # number of factors
IDY = matrix(c(
  9,-1,-1,
  1,-1,-1,
  1,-1,-1,
  1,-1,-1,
  1,-1,-1,
  -1,9,-1,
  -1,1,-1,
  -1,1,-1,
  -1,1,-1,
  -1,1,-1,
  -1,1,-1,
  -1,-1,9,
  -1,-1,1,
  -1,-1,1,
  -1,-1,1,
  -1,-1,1,
  -1,-1,1
),ncol=NZ,byr=T)
# NZ: number of factors
# 9: fixed at one for identifing the factor
# 1: estimate this parameter without shrinkage
# -1: estimate this parameter using lasso shrinkage
# 0: fixed at zero.
```

Example

Function:

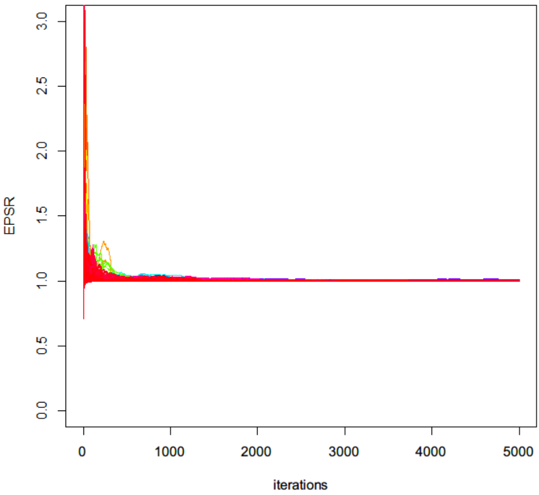
```
blcfa(filename, varnames, usevar, IDY, estimation = 'Bayes', ms = -9)
# estimation ( = 'ML' / 'Bayes', the default value is 'Bayes')
# ms represents missing value
```

After running this function:

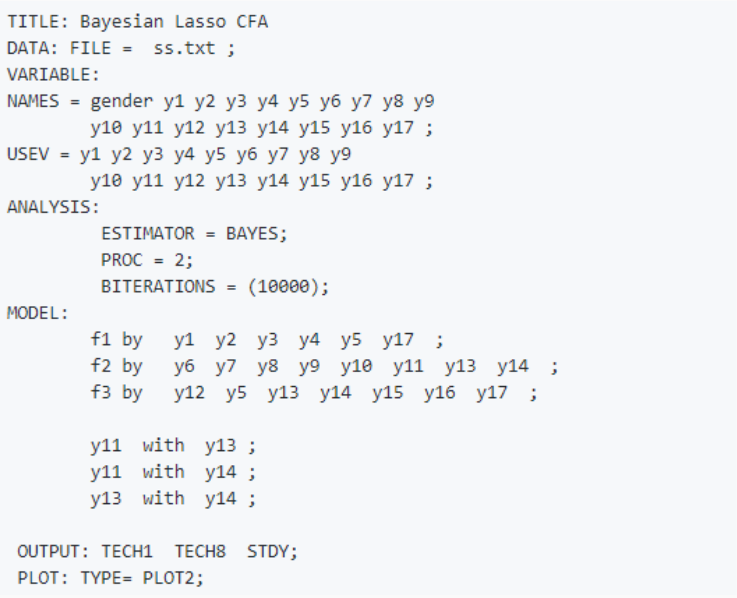
```
The program is running. See 'log.txt' for details.
Gibbs sampling ended up, specific results are being calculated.
```

('log.txt' records the process of parallel computing of two MCMC chains)

Convergence: Estimated Potential Scale Reduction Value



Results



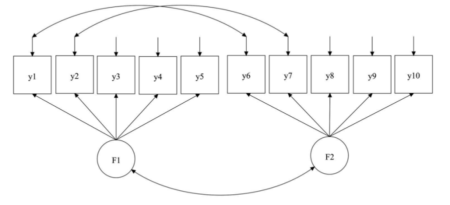
```
TITLE: Bayesian Lasso CFA
DATA: FILE =  ss.txt ;
VARIABLE:
NAMES = gender y1 y2 y3 y4 y5 y6 y7 y8 y9
        y10 y11 y12 y13 y14 y15 y16 y17 ;
USEV = y1 y2 y3 y4 y5 y6 y7 y8 y9
        y10 y11 y12 y13 y14 y15 y16 y17 ;
ANALYSIS:
         ESTIMATOR = BAYES;
         PROC = 2;
         BITERATIONS = (10000);
MODEL:
        f1 by   y1  y2  y3  y4  y5  y17  ;
        f2 by   y6  y7  y8  y9  y10  y11  y13  y14  ;
        f3 by   y12  y5  y13  y14  y15  y16  y17  ;

        y11  with  y13 ;
        y11  with  y14 ;
        y13  with  y14 ;

 OUTPUT: TECH1  TECH8  STDY;
 PLOT: TYPE= PLOT2;
```

Simulation Study

Simulation Study

To demonstrate the validity of the proposed method and compare it with the post-hoc model modification method.

- factor loadings: 0.5, 0.8
- factor correlation: 0.3, 0.7
- residual correlations (ψ_{16}, ψ_{27}): 0.0, 0.3 and 0.7
- sample size: 200, 500
- Estimator: Bayesian Lasso + ML **VS** ML + Post-hoc model modification



Results: Post-hoc model modification

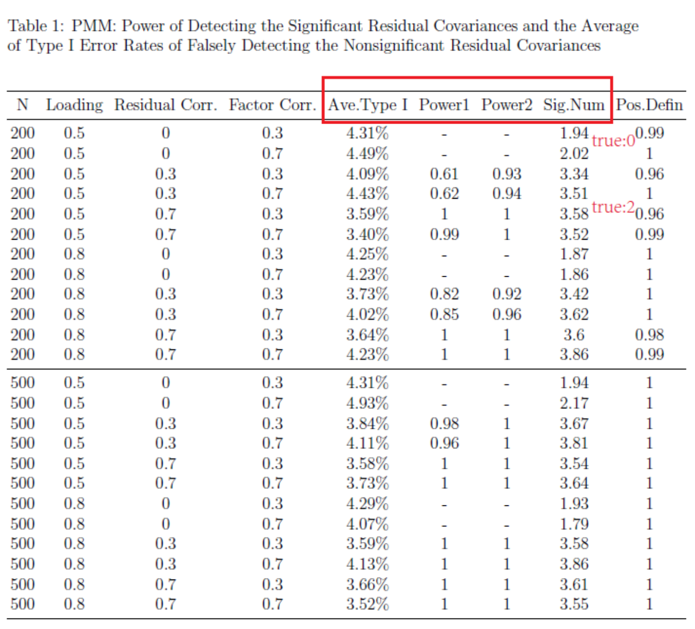


Table 1: PMM: Power of Detecting the Significant Residual Covariances and the Average of Type I Error Rates of Falsely Detecting the Nonsignificant Residual Covariances

N	Loading	Residual Corr.	Factor Corr.	Ave.Type I	Power1	Power2	Sig.Num	Pos.Defin
200	0.5	0	0.3	4.31%	-	-	1.94	0.99
200	0.5	0	0.7	4.49%	-	-	2.02	1
200	0.5	0.3	0.3	4.09%	0.61	0.93	3.34	0.96
200	0.5	0.3	0.7	4.43%	0.62	0.94	3.51	1
200	0.5	0.7	0.3	3.59%	1	1	3.58	0.96
200	0.5	0.7	0.7	3.40%	0.99	1	3.52	0.99
200	0.8	0	0.3	4.25%	-	-	1.87	1
200	0.8	0	0.7	4.23%	-	-	1.86	1
200	0.8	0.3	0.3	3.73%	0.82	0.92	3.42	1
200	0.8	0.3	0.7	4.02%	0.85	0.96	3.62	1
200	0.8	0.7	0.3	3.64%	1	1	3.6	0.98
200	0.8	0.7	0.7	4.23%	1	1	3.86	0.99
500	0.5	0	0.3	4.31%	-	-	1.94	1
500	0.5	0	0.7	4.93%	-	-	2.17	1
500	0.5	0.3	0.3	3.84%	0.98	1	3.67	1
500	0.5	0.3	0.7	4.11%	0.96	1	3.81	1
500	0.5	0.7	0.3	3.58%	1	1	3.54	1
500	0.5	0.7	0.7	3.73%	1	1	3.64	1
500	0.8	0	0.3	4.29%	-	-	1.93	1
500	0.8	0	0.7	4.07%	-	-	1.79	1
500	0.8	0.3	0.3	3.59%	1	1	3.58	1
500	0.8	0.3	0.7	4.13%	1	1	3.86	1
500	0.8	0.7	0.3	3.66%	1	1	3.61	1
500	0.8	0.7	0.7	3.52%	1	1	3.55	1

true:0

true:2

Results: BLCFA

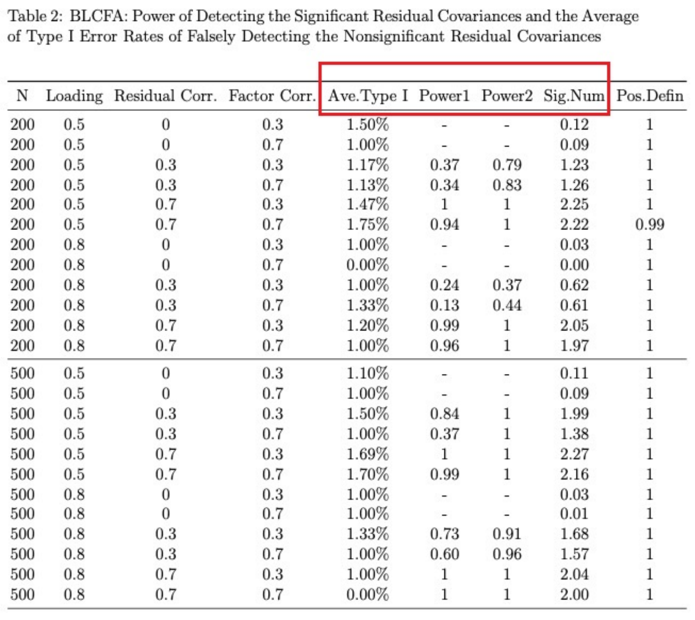
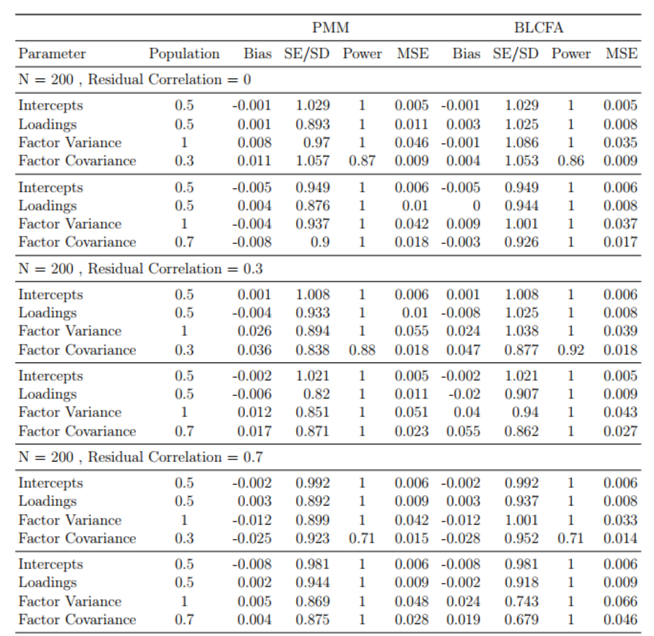


Table 2: BLCFA: Power of Detecting the Significant Residual Covariances and the Average of Type I Error Rates of Falsely Detecting the Nonsignificant Residual Covariances

N	Loading	Residual Corr.	Factor Corr.	Ave.Type I	Power1	Power2	Sig.Num	Pos.Defin
200	0.5	0	0.3	1.50%	-	-	0.12	1
200	0.5	0	0.7	1.00%	-	-	0.09	1
200	0.5	0.3	0.3	1.17%	0.37	0.79	1.23	1
200	0.5	0.3	0.7	1.13%	0.34	0.83	1.26	1
200	0.5	0.7	0.3	1.47%	1	1	2.25	1
200	0.5	0.7	0.7	1.75%	0.94	1	2.22	0.99
200	0.8	0	0.3	1.00%	-	-	0.03	1
200	0.8	0	0.7	0.00%	-	-	0.00	1
200	0.8	0.3	0.3	1.00%	0.24	0.37	0.62	1
200	0.8	0.3	0.7	1.33%	0.13	0.44	0.61	1
200	0.8	0.7	0.3	1.20%	0.99	1	2.05	1
200	0.8	0.7	0.7	1.00%	0.96	1	1.97	1
500	0.5	0	0.3	1.10%	-	-	0.11	1
500	0.5	0	0.7	1.00%	-	-	0.09	1
500	0.5	0.3	0.3	1.50%	0.84	1	1.99	1
500	0.5	0.3	0.7	1.00%	0.37	1	1.38	1
500	0.5	0.7	0.3	1.69%	1	1	2.27	1
500	0.5	0.7	0.7	1.70%	0.99	1	2.16	1
500	0.8	0	0.3	1.00%	-	-	0.03	1
500	0.8	0	0.7	1.00%	-	-	0.01	1
500	0.8	0.3	0.3	1.33%	0.73	0.91	1.68	1
500	0.8	0.3	0.7	1.00%	0.60	0.96	1.57	1
500	0.8	0.7	0.3	1.00%	1	1	2.04	1
500	0.8	0.7	0.7	0.00%	1	1	2.00	1

Results: Estimates



		PMM				BLCFA			
Parameter	Population	Bias	SE/SD	Power	MSE	Bias	SE/SD	Power	MSE
N = 200 , Residual Correlation = 0									
Intercepts	0.5	-0.001	1.029	1	0.005	-0.001	1.029	1	0.005
Loadings	0.5	0.001	0.893	1	0.011	0.003	1.025	1	0.008
Factor Variance	1	0.008	0.97	1	0.046	-0.001	1.086	1	0.035
Factor Covariance	0.3	0.011	1.057	0.87	0.009	0.004	1.053	0.86	0.009
Intercepts	0.5	-0.005	0.949	1	0.006	-0.005	0.949	1	0.006
Loadings	0.5	0.004	0.876	1	0.01	0	0.944	1	0.008
Factor Variance	1	-0.004	0.937	1	0.042	0.009	1.001	1	0.037
Factor Covariance	0.7	-0.008	0.9	1	0.018	-0.003	0.926	1	0.017
N = 200 , Residual Correlation = 0.3									
Intercepts	0.5	0.001	1.008	1	0.006	0.001	1.008	1	0.006
Loadings	0.5	-0.004	0.933	1	0.01	-0.008	1.025	1	0.008
Factor Variance	1	0.026	0.894	1	0.055	0.024	1.038	1	0.039
Factor Covariance	0.3	0.036	0.838	0.88	0.018	0.047	0.877	0.92	0.018
Intercepts	0.5	-0.002	1.021	1	0.005	-0.002	1.021	1	0.005
Loadings	0.5	-0.006	0.82	1	0.011	-0.02	0.907	1	0.009
Factor Variance	1	0.012	0.851	1	0.051	0.04	0.94	1	0.043
Factor Covariance	0.7	0.017	0.871	1	0.023	0.055	0.862	1	0.027
N = 200 , Residual Correlation = 0.7									
Intercepts	0.5	-0.002	0.992	1	0.006	-0.002	0.992	1	0.006
Loadings	0.5	0.003	0.892	1	0.009	0.003	0.937	1	0.008
Factor Variance	1	-0.012	0.899	1	0.042	-0.012	1.001	1	0.033
Factor Covariance	0.3	-0.025	0.923	0.71	0.015	-0.028	0.952	0.71	0.014
Intercepts	0.5	-0.008	0.981	1	0.006	-0.008	0.981	1	0.006
Loadings	0.5	0.002	0.944	1	0.009	-0.002	0.918	1	0.009
Factor Variance	1	0.005	0.869	1	0.048	0.024	0.743	1	0.066
Factor Covariance	0.7	0.004	0.875	1	0.028	0.019	0.679	1	0.046

Discussion

Discussion

- We proposed a two-steps method for detecting significant residual covariances and cross-loadings in one estimation.
- Acceptable power except for the conditions when the sample size and residual correlations / cross-loadings are small.
- By re-analyzing the model with significant residual correlations in Mplus, the advantages of Bayesian Lasso CFA can be extended to complex models.

Discussion

- The Bayesian Lasso CFA method also circumvents the limitations of post-hoc model modifications.
- Simulation results showed that the proposed method performed better in terms of Type I error rates and the percentage of positive definite covariance matrix.
- This method can detect all the significant residual correlations and cross-loadings in one estimation without sequential model modifications.

Discussion

- Recently, we extented the lasso method to adaptive lasso method in detecting cross-loadings.
- This package can only handle with continuous variables and need to be extended to ordered categorical data.
- This package can also be extended to conduct bi-factor models with its advantages in maintaining the positive definiteness of covariance matrix.

Thanks for listening!