Bayesian Factor Analysis as a Variable-Selection Problem

Lu, Z. H. , Chow, S. M. , & Loken, E. . (2016). Bayesian factor analysis as a variable-selection problem: alternative priors and consequences. *Multivariate Behav Res*, *51*(4), 519-539. DOI: 10.1080/00273171.2016.1168279

Backgroud

Disadvantages of strict restrictions in CFA: However, prior knowl- edge may not be available, and specifying oversimplified factor-loading structures may lead to biased estimates of the parameters in □ (Asparouhov & Muthén, 2009).

Hybrid approach between EFA and CFA: modification indices

Limitations of MI method:

- Low generalization ability
- The best subset selection requires the com- parison of 2^p models, which is not feasible even for moderate p.
- without a proper correction procedure, the Type I error rate may be far from the specified significance level.
- the parameter and standard error estimates may be severely inflated by collinearity problems and may lead to falsely identified significant variables (Harrell, 2001).

-- > an alternative way to combine the strengths of EFA and CFA is to structure factor analysis as a variable selection problem.

variable selection:ridge, ssp

the current study: test the performance of ssp and compare it with ridge in loadings identification and estimation in one-step. (distinct from the two-step procedure adopted by Muthén and Asparouhov (2012) in their BSEM-RP.). a real data analysis

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*, 85–117. doi:10.1214/09-BA403

"To ease presentation, we assume that no intercept term exists in the model. "

Frequentist ridge and lasso

risge: The ridge estimator shrinks the magnitude of all elements of β toward zero. The L2 penalty function reflects the subjective belief that estimates of β with large magnitudes are less preferable.

lasso: As illustrated in Hastie et al.(2009), the lasso estimator sets the elements of the MLE estimator that are smaller than $1/\tau^2$ to exactly zero and shrinks the other elements to zero by $1/\tau^2$.

Bayesian ridge and lasso

two popular prior choices for β are the Gaussian distribution and the Laplace distributions, expressed respectively as

$$p(\boldsymbol{\beta}|\sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\beta_j^2\right\}, \qquad (5)$$

$$p(\boldsymbol{\beta}|\sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{p}|\beta_j|\right\}, \qquad (6)$$

Bayesian ridge

-

$$p(\lambda_{jk}|\psi_j) = N(\lambda_{0jk}, \psi_j\sigma^2_{\lambda 0jk}),$$
$$p(\mu_j|\psi_j) = N(\mu_{0j}, \psi_j\sigma^2_{\mu 0j}),$$
$$p(\psi_j) = IG(\alpha_{1j}, \alpha_{2j}),\ p(\boldsymbol{\Phi}) = IW(\rho_0, \boldsymbol{\Phi}_0), \qquad (8)$$

where IG and IW stand for inverse gamma and inverse Wishart distributions, respectively. The terms $\rho_0 > q-1$, $\alpha_{1j} > 0$, $\alpha_{2j} > 0$, λ_{0jk}, $\sigma^2_{\lambda 0jk} > 0$, μ_{0j}, $\sigma^2_{\mu 0j} > 0$, and posi-

spike and slab

we propose using the spike-and-slab prior (SSP; Ishwaran & Rao, 2005; Mitchell & Beauchamp, 1988),

$$p(\lambda_{jk}|r_{jk}) = (1 - r_{jk})\delta_0 + r_{jk}N(0, c_{jk}^2),$$
$$\text{with } p(r_{jk}) = Bernoulli(p_{jk}), \quad (10)$$

for all cross loadings in $\mathbf{\Lambda}$ except for the cross loadings that are fixed to zero to satisfy the minimum constraints of the model.[8] Here, r_{jk} is a binary latent variable commonly used in the formulation of mixture models, which indicates whether the loading in the jth row and kth column of $\mathbf{\Lambda}$ should be set to zero or freed to deviate from zero; δ_0 is a point mass function at 0, and c_{jk}^2 (where c_{jk}^2 is usually substantially greater than 0) and p_{jk} are hyperparameters reflecting the user's prior knowledge or subjective belief.

hierarchical normal mixture prior (George & McCulloch, 1993),

$$p(\lambda_{jk}|r_{jk}) = (1 - r_{jk})N(0, \sigma_{jk}^2) + r_{jk}N(0, c_{jk}^2),$$
$$\text{with } p(r_{jk}) = Bernoulli(p_{jk}), \quad (11)$$

That is, when σ2 is small (e.g., 0.01), the first component would center narrowly around zero, as opposed to being specified as a point mass (i.e., a spike) at zero.

This prior is a mixture of two components: (a) the "spike"—representing the prior belief that λjk should be fixed at zero and (b) the "slab" (e.g., $c_{jk}^2 = 10000$)—the wider normal distribution representing the belief that λjk should be estimated from data because it may deviate substantially from 0.

Posterior

- SSP confirmatory estimator: $E(\lambda_{jk}|Y, R = \widetilde{R})$ This expectationc an be estimated by the mean of the subset of MCMC samples for which $R^{(t)} = \widetilde{R}$ As demonstrated in the simulation study, the SSP conditional estimator provides similar results to CFA models where the loading structures characterized by R˜ are assumed to be known.
- SSP conditional estimator: $E(\lambda_{jk}|Y, r_{jk} = 1)$
- SSP marginal estimator:

  $E(\lambda_{jk}|Y)$, mixture distribution as $E(\lambda_{jk}|Y, r_{jk} = 1)P(r_{jk} = 1|Y) + E(\lambda_{jk}|Y, r_{jk} = 0)P(r_{jk} = 0|Y)$, This estimator may thus be preferred in situations where the researcher does not regard the final model as "definitive" but rather wants to weigh the cross-loading estimates by their variable-selection uncertainties.

one-step --> two-step method

To select a final model from a collection of candidate models, one approach is to select a final model in which only cross loadings with $p(r_{jk} = 1|Y) > 0.5$ This specific model is the optimal predictive model in regressions with independent and identically distributed Gaussian errors.

advantages:

- facilitates the classification of zero cross loadings by shrinking them closer to zero, leading to more parsimonious models and interpretations.
- For nonzero loadings, the SSP provides higher power for detection by learning the posterior distribution of rij = 1 and exerting less effect of shrinkage on the moderate or large cross loadings.
- one-step approach for model selection, parameter estimation, and inference, which makes the data analytic process more convenient.
- the less informative prior distributions of the cross loadings in the BSEM-SSP exert less influence on the posterior distribution. (less sensitive,By comparison, Muthén and Asparouhov (2012) and Asparouhov, Muthén, and Morin (2015) suggested running BSEM-RP multiple times by using several prior variances in the range of [0.001, 0.1] for sensitivity checking)
- enables fast calculation of Bayes factors (Pang & Gill, 2009) not available for Bayesian lasso and BSEM-RP

Simulation

Comparison

spike and slab: two prior variance: 0.01, 10000; identification criteria: $p(r_{jk} = 1|Y) > 0.5$

ridge: identification criteria: est/se >1.96 / 2.99(without / with Bonferroni correction)

frequentist:

- forward-backward stepwise (FBS) model modification with wald test, (significant criteria: 21 levels of alpha in the range of exp(-1.5) - - exp(-6.5), 0.05 as defult)
- a forward selection (FS) procedure with a Scheffé-type test (FS; Hancock, 1999), which was developed to better control the Type I error across the set of all possible post hoc model modifications. chi-square test (significant criteria: 25 levels of alpha in the range of exp(-0.25) - - exp(-6.25))
- EFA: Confdence intervals were used to decide the significance of the cross loadings with and without the Bonferroni correction.

sensitivity analysis:

The informative prior distributions were specified to have means that were equal to the true values used in the simulation. In addition, we considered the prior settings with means equal to two times and one half of the true values

ridge: prior variance: 0.001, 0.01, 0.1

Results

Power and Type I error rate

receiver operating characteristic (ROC) curves : x axis: false positive rate; y axis: true positive rate

Reduced power for the BSEM-RP due to the discrepancy between the prior and the data.

The performance of the BSEM-SSP was comparable to that of the FBS-MI in the conditions considered. In FBS method, without appropriate multiple comparison correction, the Type I errors of the sequential procedure appeared inflated

These ROC curves demonstrate that the FS approach may be too conservative, with low false detection rates of zero cross loadings and the low detection power of nonzero cross loadings

EFA: less power. By comparison, parsimonious models (BSEM, CFA) with few spurious nonzero parameters are generally associated with better detection power and higher true discovery rates.

N increases, the difference between methods decrease

Estimates

For cross-loadings that were truly zero, BSEM-SSP provided more efficient estimates that are closer to zero compared to BSEM-RP and FBS-MI due to the use of the spike-and-slab prior.

the marginal and conditional estimators of BSEM-SSP are more informative concerning the nonzero cross loadings than the posterior mean or median of BSEM-RP.

BSEM-RP had better performance due to the correct informative prior distributions when the nonzero cross loadings are small.

Empirical study

Compared to the BSEM-RP, the BSEM-SSP tended to identify cross loadings that were greater than 0.1(based on EFA) with less shrinkage and shrink the others to be closer to zero

The strong priors imposed in the BSEM-RP might have overshrunk some truly nonzero cross loadings and forced them to exert their associations with the latent factors through other main or cross loadings, as was found in our simulation study.

Discussion

Overall, the BSEM-SSP outperformed the other approaches considered when the sample size and effect size were relatively large and yielded more parsimonious loading structures (by shrinking weaker cross loadings to zero) in situations involving smaller sample and/or effect sizes.

FBS-MI with the default p-value threshold is characterized by heightened false positive rates—a problem that actually aggravates as the sample size grows.

many other frequentist model ft indices exist for model-selection purposes (see Hu & Bentler, 1998, 1999, for reviews)

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecifcation. Psychological Methods, 3, 424–453. doi:10.1037/1082-989X.3.4.424 Hu, L.-T., & Bentler, P. M.(1999). Cutoff criteria for ft indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6, 1–55. doi:10.1080/10705519909540118

Shrinkage priors

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis, 1, 515–534. doi:10.1214/06-BA117A

Leng, C., Tran, M.-N., & Nott, D. (2014). Bayesian adaptive lasso. Annals of the Institute of Statistical Mathematics, 66, 221–244. doi:10.1007/s10463-013-0429-6 Li, Q., & Lin, N. (2010). The Bayesian elastic net. Bayesian Analysis, 5, 151–170. doi:10.1214/10-BA506

Limitations of the two-step procedure of BSEM-RP suggested by Muthén and Asparouhov (2012)

- reuse of the same data -- overfitting in the second step: possible solution: cross validation
- the variation of the cross-loading selection in the first step of the BSEM-RP is not accounted for in the second step (estimation step), which leads to underestimation of the standard errors.

My Thoughts

Good title and good story: extend the simulation study to a big question. very comprehensive simulation study (especially for the sensitivity analysis)

The ridge prior in Mplus is not the common ridge prior, the prior variance and the penalty strength is fixed, perhaps even cannot outperformed the frequentist ridge method.

ssp: combine non-infor with ridge can also be adapted using laplace prior (aBlasso-ssp). maybe the perfomance of ssp is similar to horseshoe which consider two shrinkage parameters (local and global)

- Return