Bayesian Regularization in Multiple-Indicators Multiple-Causes Models

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

Joint work with Dr. Xinya Liang
Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas, USA, xl014@uark.edu

20 July 2021

Overview

Introduction

Regularized Structural Equation Modeling (SEM)

Trade-off between model fitting and model complexity in SEM.

- Complex model - overfitting, less generalizability
- Simple model - omit important variables, poor model fitting

Regularization for achieving model parsimony and high generalizability (e.g., Jacobucci & Grimm, 2018a).

- Shrink nuisance parameters toward zero & identify essential parameters
- Retain accurate parameter estimates & improve the generalizability of estimates

Bayesian Regularization

Bayesian regularization assigns penalty priors to regularize the posterior distributions of parameters.

It is flexible in (Polson & Sokolov, 2019; van Erp et al., 2019):

- estimating the shrinkage parameters
- quantifying the uncertainty of parameter estimates
- handling small sample sizes

Bayesian Regularization

Different penalty priors:

- Ridge: global shrinkage
- Lasso (least absolute shrinkage and selection operator; Park & Casella, 2008; Tibshirani, 1996): global shrinkage
- Adaptive lasso (alasso; Zou, 2006): local shrinkage
- Spike-and-slab prior (SSP; Mitchell & Beauchamp, 1988): assign a discrete mixture of normal distributions on parameters
- Horseshoe (Carvalho et al., 2010): global-local shrinkage

Bayesian Regularized SEM

Methods	Measurement Models	Structural Models
Ridge	Muthén & Asparouhov (2012, 2013)	
Lasso	Chen et al. (2021), Pan et al. (2017), Zhang et al. (2021);	-
Alasso	Chen (2021), Pan et al. (2021);	Feng et al. (2017), Jacobucci & Grimm (2018b), Brandt et al. (2018);
SSP	Lu et al. (2016)	Brandt et al. (2018)
Horseshoe	-	-

Tabel 1: Integration of Different Penalty Priors with SEM

Method Comparison

	Comparison		Model
Chen et al. (2021)	Ridge	Lasso	Measurement
Lu et al. (2016)	Ridge	SSP	Measurement
Feng et al. (2017)	Alasso	Lasso	Structural
Brandt et al. (2018)	Alasso	ALasso+SSP	Structural

Tabel 2: Comparison between Different Penalty Priors

- Lasso and SSP have advantages in achieving parsimonious factor structures than ridge.
- Alasso has benefits in reducing appreciable bias caused by the global lasso shrinkage.

Purpose

Investigate the performance of different Bayesian regularization methods in parameter estimation and variable selection using MIMIC models:

- Penalty priors vs Non-informative prior
- Global vs Local vs Global-local shrinkage
- Under different modeling conditions (sample sizes, multicollinearity, effect sizes)

MIMIC Models

Suppose there are K latent factors $\boldsymbol{\omega}$ measured by J indicators $\mathbf{y}$ and regressed on P predictors $\mathbf{X}$, a MIMIC model (Jöreskog & Goldberger, 1975) can be expressed as follows:

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, i = 1, 2, ..., n, \tag{1}$$

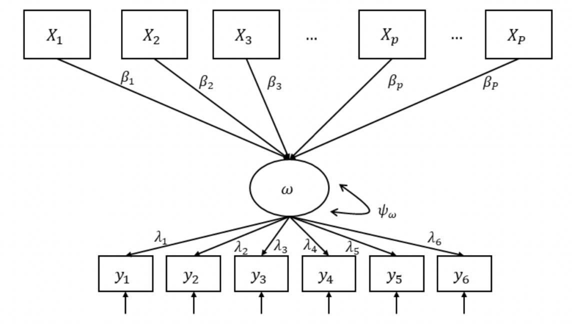
$$\boldsymbol{\omega}_i = \boldsymbol{\mu}_\omega + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\delta}_i \tag{2}$$

- $\mathbf{y}_i$: observed values of J indicators for the i-th participant.
- $\boldsymbol{\mu}, \boldsymbol{\mu}_\omega$: vector of intercepts.
- $\boldsymbol{\Lambda}$: factor loading matrix.
- $\boldsymbol{\omega}_i$: latent factors.
- $\boldsymbol{\epsilon}_i$: measurement errors.
- $\boldsymbol{\beta}$: path coefficients.
- $\boldsymbol{\delta}_i$: factor disturbances.

MIMIC Models

The utility of MIMIC models is versatile (Finch & Miller, 2019)

- Control the influence of covariates on latent variables
- Test the measurement invariance between groups
- Identify differential item functioning



MIMIC Model with One-factor and Six-indicators

Bayesian Regularization

Bayesian Ridge

$$\beta_p \sim N(0, \sigma^2) \tag{3}$$

- β_p : p-th parameter to be regularized.
- σ^2 : variance hyperparameter, determines the penalty strength.
- The prior variance can be fixed at a preassigned value such as 0.01 (Muthén & Asparouhov, 2012) or 0.001 (Jacobucci & Grimm, 2018a), or be estimated through a hyperprior.

Applications

- Identify cross-loadings and residual correlations (e.g., Falkenström et al., 2015)
- Handle small sample sizes (e.g., Crenshaw et al., 2016)
- Assess measurement invariance (e.g., de Bondt & van Petegem, 2015)

Bayesian Lasso

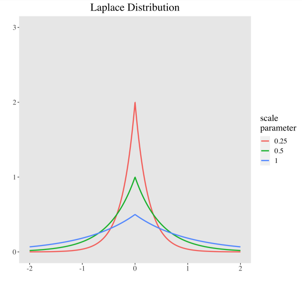
$$\beta_p \sim N(0, \psi_\omega \tau_p^2), \psi_\omega^{-1} \sim Gamma(\alpha_\omega, \beta_\omega) \tag{4}$$

$$\tau_p^2 \sim Gamma(1, \frac{\gamma^2}{2}), \gamma^2 \sim Gamma(a_l, b_l) \tag{5}$$

- τ_p is included to obtain the desired Laplace distribution of the conditional prior.
- γ is the global penalty parameter.

Applications

- Identify cross-loadings and residual correlations (Chen et al., 2021; Pan et al., 2017; Zhang et al., 2021)



Bayesian Adaptive Lasso

$$\beta_p \sim N(0, \psi_\omega \tau_p^2), \psi_\omega^{-1} \sim Gamma(\alpha_\omega, \beta_\omega) \tag{6}$$

$$\tau_p^2 \sim Gamma(1, \frac{\gamma_p^2}{2}), \gamma_p^2 \sim Gamma(a_l, b_l) \tag{7}$$

- γ_p : local penalty parameter.

Bayesian adaptive lasso has been extended to:

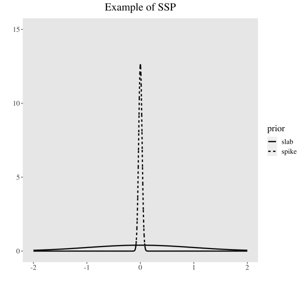
- SEMs with ordinal variables (Feng et al., 2017)
- Latent change score models (Jacobucci & Grimm, 2018b)
- Detect multiple linear and nonlinear effects in SEM with SSP (Brandt et al., 2018)

Spike-and-Slab

$$\beta_p \sim r_p N(0, c_p^2) + (1 - r_p) N(0, \sigma_p^2) \tag{8}$$

$$r_p \sim Bernoulli(.5) \tag{9}$$

- r_p : selection variable
- $N(0, \sigma_p^2)$: a point mass function (spike) commonly with a small prior variance to shrink the parameter to zero
- $N(0, c_p^2)$: the fuzzy prior (slab) that is typically assigned a large prior variance

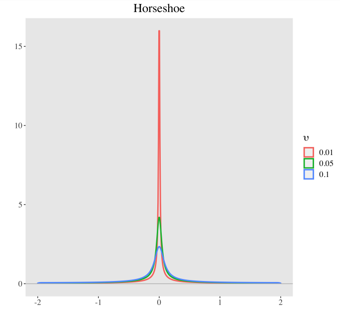


spike-$N(0, 0.001)$
slab-$N(0, 1)$

Horseshoe

$$\beta_p \sim N(0, \rho_p^2 \upsilon^2), \rho_p \sim C^+(0,1), \upsilon \sim C^+(0,1) \qquad (10)$$

- ρ_p, υ : local and global shrinkage parameters, respectively.
- Placing the half-Cauchy distributions $C^+(0,1)$ is similar to putting a beta(0.5, 0.5) prior on the shrinkage weight $\kappa_p = 1/(1+\rho_p^2\upsilon^2)$.

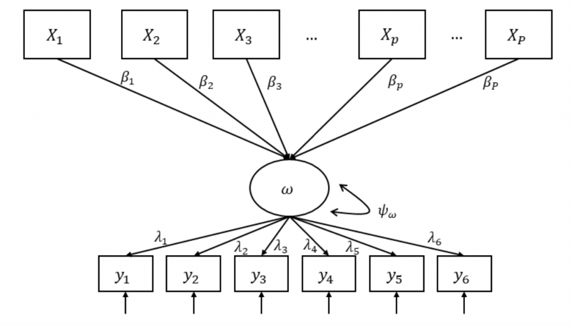


Simulation Study

Purpose and Design

We conducted a simulation study with a similar design to Jacobucci et al. (2019):

- Collinearity among covariates: 0, .2, .5, .8, .95
- Sample size: 100, 200, 300, 500, 1000



Purpose and Design

Other Settings

- Effect Sizes:
  $\beta_1 - \beta_{70} = 0, \beta_{71} - \beta_{80} = .2, \beta_{81} - \beta_{90} = .5, \beta_{91} - \beta_{100} = .8$
- Factor loadings: c(1, .8, .8, .8, .5, .5)
- Residual variances of indicators and factor disturbance: 1
- Number of replications: 200 datasets per condition

Model Estimation

Software: R, JAGS (Plummer, 2003)

Hyperparameters of priors for path coefficients:

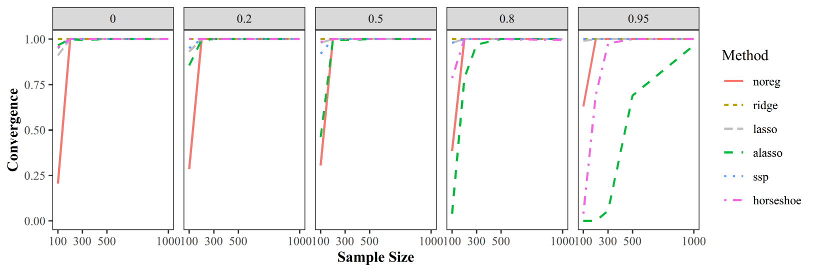
- Diffuse prior: $N(0, 1000)$
- Ridge: $N(0, 0.01)$ (Muthén & Asparouhov, 2012)
- Lasso and Alasso: $\alpha_l = 1, \beta_l = 0.01$ (Chen et al., 2021)
- SSP: $\sigma_p^2 = 0.001, c_p^2 \sim IG(0.5, 0.5)$ (van Erp et al., 2019)
- Horseshoe: $\rho_p \sim C^+(0, 1), \upsilon \sim C^+(0, 1)$

- Priors for other parameters (e.g., loadings): diffuse priors.
- Model convergence criteria: The estimated potential scale reduction (EPSR; Gelman et al., 1996) value should be less than 1.05 within 5,000 - 20,000 burn-in iterations.

Evaluation Criteria

- Convergence Rate
- Rejection Rate of 95% Highest Posterior Density (HPD) Interval
- Rejection Rate of Threshold: the proportion of converged replications where $\mid \beta_{est} \mid > .1$ (Feng et al., 2017)
- 95% Coverage Rate
- Relative Bias
- Root Mean Square Error (RMSE) $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\beta_{est} - \beta_{true}\right)^2}$ where N is the number of converged replications

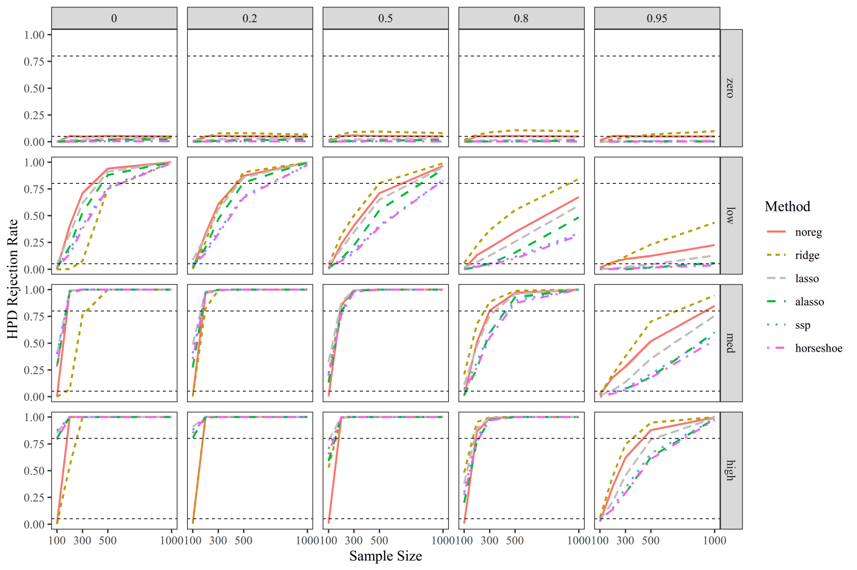
Results: Convergence Rates



noreg: diffuse prior; collinearity: 0, .2, .5, .8, and .95

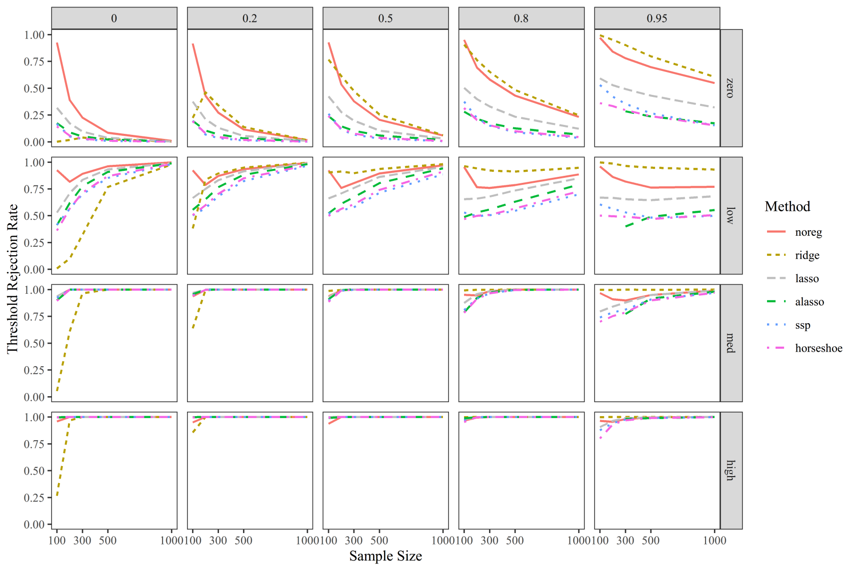
- Diffuse prior: low convergence rates (.21 - .63) with small sample size
- Ridge, lasso, and SSP: excellent convergence rates
- Alasso and horseshoe: convergence rates decreased as the collinearity increased

Results: 95% HPD Rejection Rates

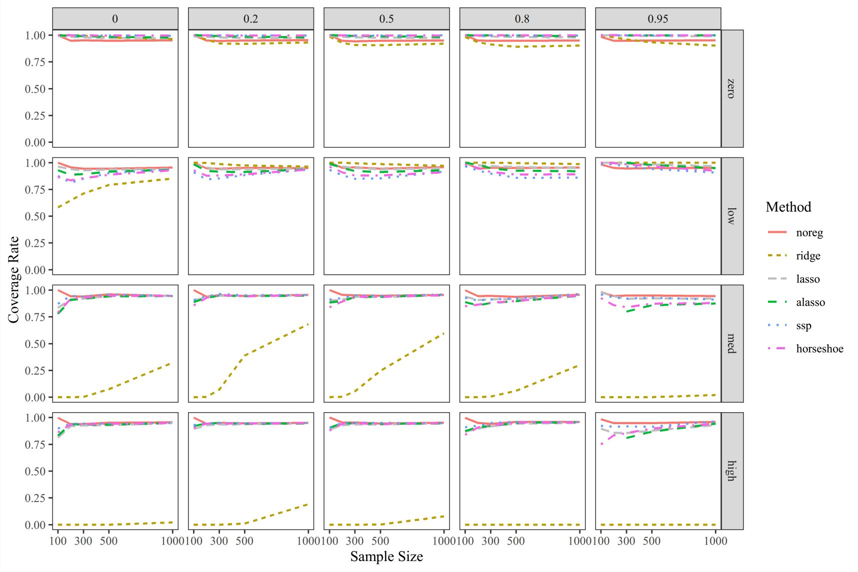


effect size: zero, low (.2), medium (.5), and high (.8)

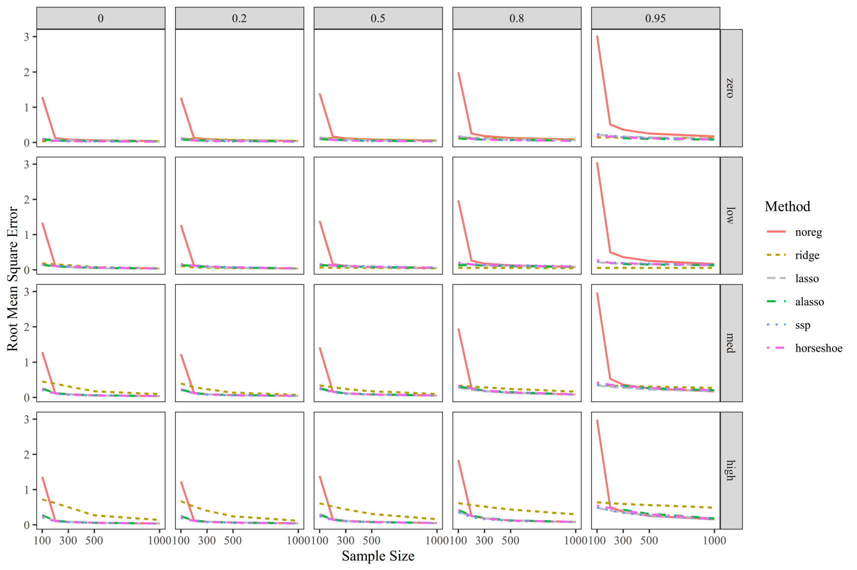
Results: Threshold Rejection Rates



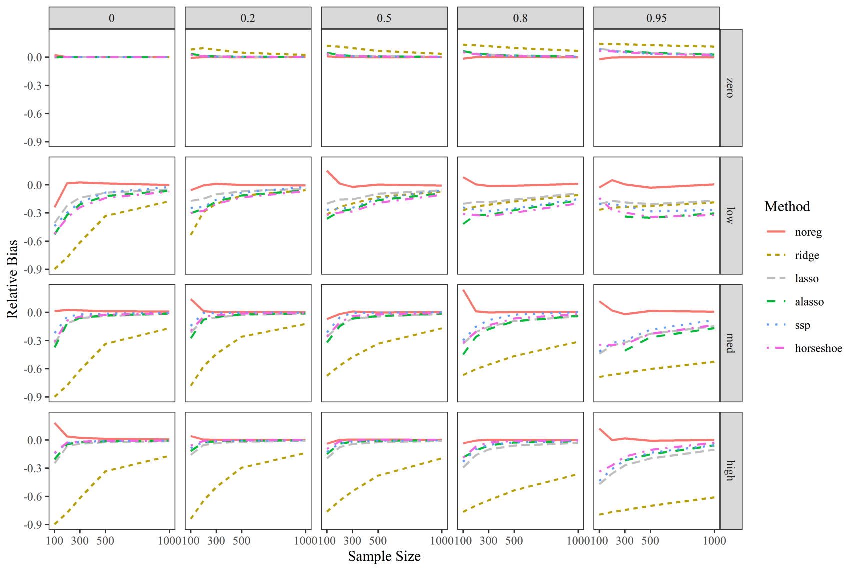
Results: 95% Coverage Rate



Results: RMSE



Results: Relative Bias



Results: Other Parameters

Diffuse prior with small sample sizes

- Unacceptable relative bias and RMSE for factor disturbance
- Low coverage rates for factor loadings

Empirical Illustration

Data and Model

Data

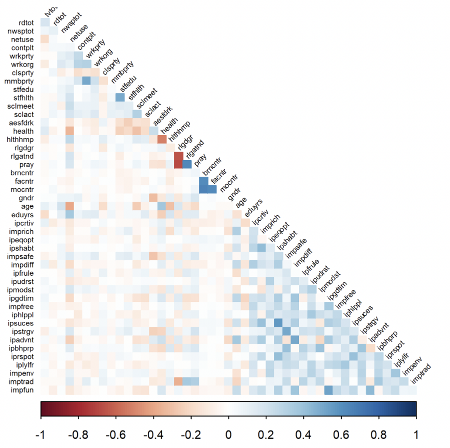
- The third round of the European Social Survey
- Randomly selected 1,000 samples (45.5% male and 54.5% female, Age: mean = 46.69, sd = 18.04)

Factor

- Center for Epidemiologic Studies - Depression Scale (CES-D, Radloff, 1977)
- e.g., "Felt depressed, how often past week"
- Eight items, 4-point Likert-type scale, treated as continuous following Van de Velde et al. (2009).

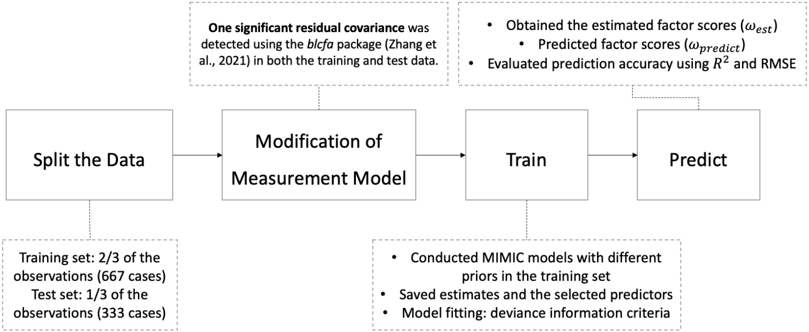
Multicollinearity

Forty-six covariates include demographic variables, health status, family status, and portrait values



Correlation Heatmap for Covariates

Hold-out Method



Results: Variable Selection and Parameter Estimation

covariate	definition	Non-informative	Ridge	Lasso	Alasso	Horseshoe	SSP
wrkprty	Worked in political party or action group last 12 months.	**.46 (.083, .822)**[1]	.121 (-.049, .286)	.081 (-.06, .284)	.22 (-.041, .56)	.195 (-.05, .572)	.298 (-.035, .592)
sclmeet	How often socially meet with friends, relatives or colleagues.	**-.076 (-.14, -.011)**	**-.071 (-.13, -.01)**	-.057 (-.115, .002)	**-.061 (-.123, -.004)**	-.045 (-.108, .011)	-.051 (-.116, .007)
sclact	Take part in social activities compared to others of same age.	**-.133 (-.195, -.07)**	**-.122 (-.18, -.064)**	**-.118 (-.178, -.056)**	**-.124 (-.182, -.064)**	**-.127 (-.189, -.061)**	**-.135 (-.202, -.066)**
aesfdrk	Feeling of safety of walking alone in local area after dark.	**.07 (.004, .135)**	**.073 (.012, .132)**	**.071 (.011, .133)**	**.069 (.006, .13)**	.074 (-.001, .137)	**.067 (.003, .14)**
health	Subjective general unhealthy.	**.22 (.142, .298)**	**.2 (.132, .271)**	**.21 (.138, .283)**	**.216 (.141, .29)**	**.235 (.157, .314)**	**.234 (.158, .312)**
hlthhmp	Hampered in daily activities by illness/disability/infirmity/mental problem.	**-.112 (-.182, -.043)**	**-.108 (-.171, -.044)**	**-.095 (-.16, -.027)**	**-.099 (-.167, -.032)**	-.088 (-.155, .001)	**-.091 (-.167, -.015)**
gndr	gender	**-.139 (-.268, -.011)**	**-.111 (-.213, -.004)**	-.086 (-.193, .015)	-.101 (-.22, .011)	-.077 (-.21, .021)	-.106 (-.246, .02)
age	age	**-.099 (-.177, -.021)**	**-.081 (-.149, -.011)**	-.062 (-.127, .005)	**-.076 (-.143, -.003)**	-.061 (-.136, .008)	-.067 (-.151, .001)
ipeqopt	Important that people are treated equally and have equal opportunities.	**.068 (.006, .136)**	**.064 (.004, .123)**	.053 (-.005, .11)	.057 (-.001, .117)	.05 (-.008, .111)	.051 (-.007, .115)
impsafe	Important to be humble and modest, not draw attention.	**-.08 (-.151, -.005)**	**-.071 (-.138, -.004)**	-.057 (-.122, .005)	-.063 (-.133, .002)	-.048 (-.119, .012)	-.054 (-.132, .01)
ipsuces	Important to be successful and that people recognize achievements.	**.114 (.039, .193)**	**.099 (.029, .168)**	**.08 (.011, .149)**	**.092 (.023, .164)**	.076 (-.004, .144)	**.084 (.012, .168)**
ipstrgv	Important that government is strong and ensures safety.	**-.103 (-.176, -.03)**	**-.091 (-.158, -.026)**	**-.077 (-.146, -.01)**	**-.083 (-.151, -.014)**	-.067 (-.14, .007)	**-.076 (-.158, -.004)**
Number of Significant Covariates		12	11	6	8	2	6
DIC		14331.906	14287.801	14289.324	14331.017	14195.412	14235.981

Note. [1] estimate (95% highest posterior density). Bold: significant estimates detected by the 95% HPD interval. DIC = deviance information criteria.

Results: Prediction

	Correlation Among the Predicted Factor Scores and Estimated Values						R^2	RMSE
	Non-informative	Ridge	Lasso	Alasso	Horseshoe	SSP		
Non-informative	-						.171	0.642
Ridge	.977**						.181	0.628
Lasso	.921**	.941**					.191	0.614
Alasso	.948**	.969**	.973**				.181	0.621
Horseshoe	.836**	.853**	.910**	.904**			.225	0.598
SSP	.920**	.940**	.999**	.974**	.925**		.195	0.616
Estimated Values	.414**	.426**	.437**	.425**	.474**	.441**	-	-

Note. ** $p < 0.01$.

Discussion

Penalty Priors vs Non-informative Priors

Variable selection

- Ridge has advantages in handling high collinearity.
- For low collinearity conditions, penalty priors except for ridge performed better than diffuse prior in small sample sizes.

Parameter Estimation

- Penalty prior except for ridge outperformed the diffuse prior in maintaining low RMSEs with small sample sizes.

Benefits of regularization in making predictions (empirical study) and achieving model convergence (ridge, lasso, and SSP).

Global Penalty vs Local Penalty

Convergence

- Global shrinkage has advantages in model convergence.
- Alasso and horseshoe yielded low convergence rates with the co-existence of small sample size and high multicollinearity.

Variable Selection and Parameter Estimation

- Global shrinkage methods (ridge, lasso): variable selection.
- Methods with local shrinkage parameter: parameter estimation.
- SSP and Horseshoe had a similar performance in most conditions.

Takeaway

Penalty priors compared to diffuse priors

- Robust results in small sample size conditions (simulation).
- High generalizability even with a relatively large sample size (empirical study).

Choice of different penalty priors

- For variable selection: global shrinkage (e.g., ridge in high collinearity conditions).
- For parameter estimation: penalty priors which include local shrinkage.
- Model fit indexes (e.g., DIC).

Thanks for listening!