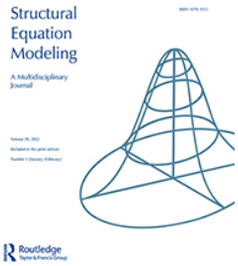
Routledge
Taylor & Francis Group



Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Advantages of Spike and Slab Priors for Detecting Differential Item Functioning Relative to Other Bayesian Regularizing Priors and Frequentist Lasso

Siyuan Marco Chen, Daniel J. Bauer, William M. Belzak & Holger Brandt

To cite this article: Siyuan Marco Chen, Daniel J. Bauer, William M. Belzak & Holger Brandt (2022) Advantages of Spike and Slab Priors for Detecting Differential Item Functioning Relative to Other Bayesian Regularizing Priors and Frequentist Lasso, Structural Equation Modeling: A Multidisciplinary Journal, 29:1, 122-139, DOI: 10.1080/10705511.2021.1948335

To link to this article: https://doi.org/10.1080/10705511.2021.1948335

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2022, VOL. 29, NO. 1, 122–139
https://doi.org/10.1080/10705511.2021.1948335

Advantages of Spike and Slab Priors for Detecting Differential Item Functioning Relative to Other Bayesian Regularizing Priors and Frequentist Lasso

Siyuan Marco Chen [a], Daniel J. Bauer[a], William M. Belzak [a], and Holger Brandt[b]

[a]University of North Carolina at Chapel Hill; [b]University of Zurich

ABSTRACT

An important step in scale development and assessment is to evaluate differential item functioning (DIF) across segments of the population. Recent approaches use lasso regularization to simultaneously detect DIF in all items and avoid incorrect anchor item assumptions that incur inflated error rates for classical DIF evaluation methods. Although promising, lasso methods cause underestimated standard errors and incorrect *p*-values. An alternative is Bayesian regularization that provides empirical standard errors. However, we point out that using empirical criteria such as credible intervals for selecting DIF parameters has limited validity. We argue that using a spike-and-slab prior with an inclusion probability criterion provides more theoretically coherent DIF selection and inference over Bayesian regularizing priors with empirical selection rules or frequentist lasso. We demonstrate this by simulation studies with Multi-group Item Response Theory and Moderated Nonlinear Factor Analysis models. Practical utility of the spike-and-slab prior selection criterion is discussed.

KEYWORDS

Differential item functioning; measurement invariance; Bayesian regularization; integrative data analysis

Psychological research relies on a consistent measurement scale for the latent construct across individuals who may differ in a variety of ways (e.g., gender, ethnicity, or age level). *Measurement Invariance* (MI), defined as how well a scale can measure a latent construct equivalently for different background populations (Mellenbergh, 1989) is a necessary property for a scale to produce the same trait scores for individuals who have the same latent trait level but differ on background characteristics. Scales without MI have items whose frequency of endorsement depends on both latent trait levels and background characteristics of the respondents. For example, Asian students are more likely to endorse an "Avoidance" item than US students for the same level of behavior activation (Chen et al., 2019). Such items display *Differential Item Functioning* (DIF) and will bias the trait score estimates (Steinberg & Thissen, 2006) if used without adjustment. It is thus essential for researchers to evaluate scales for DIF and avoid confounding item-level differences with latent trait differences between population groups.

Though DIF detection methods have advanced in recent years, known limitations still exist. Traditional iterative DIF detection methods such as IRT-LR-DIF (Kim et al., 1995; Thissen et al., 1993) need to constrain an a priori anchor item set to be equal across groups for model identification. These methods rely on repeated significance tests that may capitalize on chance (Belzak & Bauer, 2020; Draper, 1995; MacCallum et al., 1992) and are prone to inflated error rates when the anchor items are not selected correctly (Ankenmann et al., 1999; Finch, 2005; Stark et al., 2006). High error rates in DIF detection are challenging as psychological scales often have limited item pools. Yet correctly identifying anchor items before DIF evaluation can be difficult even for content experts (Scheuneman, 1987). More recently, regularization methods have been applied to detect DIF (Bauer et al., 2019; Magis et al., 2015; Tutz & Schauberger, 2015), with the advantage that they can avoid a priori designation of potentially incorrect anchor items by imposing model complexity constraints such as the lasso penalty (least absolute shrinkage and selection operator; Tibshirani, 1996) to shrink unimportant DIF parameters and select nonzero DIF effects. However, the lasso method produces biased and inconsistent parameter estimates and problematic standard errors because all parameters are shrunk with the same penalty parameter (Bauer et al., 2019; Kyung et al., 2010; Zou, 2006), which becomes an issue in subsequent analyses using these estimates or when researchers assess the statistical significance of the detected DIF estimates.

On the other hand, we believe Bayesian regularization could improve upon lasso-regularized DIF detection techniques. Bayesian penalty priors can be more flexible in retaining valid inferential properties and in extending to complex models. Past studies (e.g., Liang & Jacobucci, 2020; Shi et al., 2017 and preliminary results from; Brandt & Bauer, 2020) show the feasibility and promise of this approach. However, an important issue in Bayesian regularization is that common Bayesian shrinkage methods do not shrink any parameters to exactly zero and rely on post-hoc criteria to select important parameters (Carvalho et al., 2009; Hans, 2010; Leng et al., 2014). We point out in the next section that many selection criteria for Bayesian regularization (e.g., 95% posterior credible intervals or magnitude of posterior point estimates) are empirically chosen, do not conduct proper model selection, and have inconvenient interpretations. This makes it difficult to determine which criterion may be the most suited for DIF selection. This study aims to evaluate the spike-and-slab prior

CONTACT Siyuan Marco Chen mchen@unc.edu University of North Carolina at Chapel Hill.
Supplemental data for this article can be accessed on the publisher's website

(SSP) with inclusion probability parameters as an alternative selection criterion (Ishwaran & Rao, 2005; Lu et al., 2016; Mitchell & Beauchamp, 1988). The SSP method explicitly models the probability that a DIF parameter should be included and selects DIF by evaluating whether the posterior probability of this inclusion parameter exceeds a pre-determined threshold. This allows the SSP method to accurately account for model selection uncertainty and provides more reliable parameter estimates. Existing studies on SSP (Brandt & Bauer, 2020; Brandt et al., 2018; Lu et al., 2016) have demonstrated its utility, but none has evaluated the inferential quality of SSP relative to other regularization methods or the optimal inclusion parameter threshold, nor has this been done in the DIF detection context.

We argue in this study that the spike-and-slab prior with inclusion parameters is a theoretically coherent Bayesian shrinkage method that represents uncertainty in DIF detection better than Bayesian methods with empirical criteria. This study makes two unique contributions to the DIF evaluation literature. First, this study comprehensively compares SSP with other empirical Bayesian shrinkage methods and frequentist lasso on several performance measures, including DIF detection error rates, estimation quality, and uncertainty measures. The Bayesian shrinkage priors considered besides SSP include the small-variance normal prior (Shi et al., 2017), the Laplace prior (Pan et al., 2017), and the adaptive lasso prior (aLasso; Brandt et al., 2018). The comparison also involves measurement model conditions with simultaneous DIF effects from continuous and categorical covariates. Second, this study demonstrates the practical utility of the SSP approach for optimal DIF detection, yielding preferable inferential properties, and improving the generalizability of the DIF selection process. In the following sections, we lay out the measurement models used for DIF detection. We discuss the advantages and limitations of using Bayesian regularization for DIF evaluation and describe our research questions. We present a simulation study and discuss the results in the practical measurement invariance assessment context.

Measurement model setup

For generality, we base this study on the *moderated nonlinear factor analysis* model (Bauer, 2017; Bauer & Hussong, 2009; Curran et al., 2014) DIF effects are represented in the MNLFA model through the conditioning of item parameters on external covariates (e.g., group membership). A unidimensional MNLFA for binary items may be written as

$$\begin{aligned} P(y_{ij}=1|\eta_i) &= g^{-1}(\nu_{ij}+\lambda_{ij}\eta_i) \\ \eta_i &\sim \mathcal{N}(\alpha_i,\psi_i^2) \end{aligned} \tag{1}$$

$$\begin{aligned} \nu_{ij} &= \nu_{0j}+\kappa_j^{\mathrm{T}}\boldsymbol{x}_i \\ \lambda_{ij} &= \lambda_{0j}+\omega_j^{\mathrm{T}}\boldsymbol{x}_i \\ \alpha_i &= \alpha_0+\gamma^{\mathrm{T}}\boldsymbol{x}_i \\ \psi_i &= \psi_0 exp(\beta^{\mathrm{T}}\boldsymbol{x}_i) \end{aligned} \tag{2}$$

where y_{ij} is the observed response for person i on item j, $j=1,\ldots,p$. ν_{ij} and λ_{ij} are item intercept and slope parameters, respectively. η_i is the latent factor score for person i, assumed to follow a normal distribution with mean α_i and standard deviation ψ_i. $g(\cdot)$ is a link function (e.g., logit) between the response probability and the latent factor. Because in the Bayesian setting it is easier to specify models in terms of the standard deviation, here we differ from earlier presentations of the MNLFA by expressing the standard deviation of the factor as a log-linear function of the covariates as opposed to the variance, with the two alternative expressions being equivalent to one another.[1]

For each of the p items, κ_j and ω_j are, respectively, $q\times 1$ vectors of DIF parameters on the item intercepts and slopes. γ and β are, respectively, $q\times 1$ vectors of *impact* on the factor mean and standard deviation. Impact represents covariate influence on the latent score distribution. All of these effects on item and factor parameters are for $\boldsymbol{x}_i$, a $q\times 1$ vector of covariates for person i.[2] The null-subscripted parameters are the baseline item and factor parameters. To be identified, MNLFA requires that either the impact parameters be constrained to be null (typically undesirable) or alternatively that there exists at least one anchor item for each covariate that shows no DIF as a function of the covariate. Additionally, the scale of the latent factor must be set, for instance, by constraining α_i and ψ_i respectively to zero and one. By conditioning measurement model parameters on a vector of background variables, MNLFA allows for the simultaneous influence of multiple categorical and/or continuous background variables, improves scale score performance, and helps to generalize psychological theories over heterogeneous populations (Bauer & Hussong, 2009; Curran et al., 2018; Huang, 2018).

DIF detection with regularized MNLFA

Given the preponderance of potential DIF parameters in MNLFA, an accurate method is needed to guide model specification. Prior research has considered lasso regularization within the frequentist framework (Bauer et al., 2019; Belzak & Bauer, 2020; Magis et al., 2015; Tutz & Schauberger, 2015) to estimate all DIF parameters together without requiring pre-selected anchor items or multiple significance testing. The lasso penalty achieves both a shrinkage effect and variable selection by reducing maximum likelihood estimates by a constant amount, truncating them below at zero, so that unimportant effects are shrunken to zero and excluded (see Hastie et al., 2009, Ch. 3 for more details). When added to the DIF parameters (κ_j and ω_j in Equation 2), the lasso penalty constrains their parameter space and leaves only a subset of active DIF effects.

Previous studies show that lasso produces notably lower false positive rates and high power when compared to anchor item methods (e.g., Belzak & Bauer, 2020); however, inferential and computational limitations still exist. The first issue is bias in DIF parameter estimates. Lasso shrinks all DIF parameters

[1]The baseline latent factor SD is the square root of the baseline variance and β in the SD formulation here are 1/2 the β in the variance formulation.
[2]To express a traditional two-group 2-parameter logistic item response theory (2PL IRT) model as a special case of MNLFA, $g(\cdot)$ will be a logit link function, $q=1$, and x_i is a scalar that only takes the value 0 or 1 for two group membership.

uniformly and thus produces biased and inconsistent estimates for non-zero parameters (Hastie et al., 2009 pg. 91; Meinshausen & Bühlmann, 2006 see also; Fan & Li, 2001). To mitigate the bias, one can re-estimate the model without the lasso penalty after having used lasso for DIF item selection (e.g., Huang, 2018; Hastie et al., 2009 Ch. 3.8.5); however, this re-estimation approach ignores model selection uncertainty and causes underestimated standard errors. The resulting *p*-values are only correct *conditional on* the parameter estimate belonging to this active subset and can become an issue when researchers wish to evaluate the statistical significance of the remaining DIF effect estimates (Belzak & Bauer, 2020; Draper, 1995; Lu et al., 2016; Raftery et al., 1997). In addition, standard errors of lasso estimates are not available when the estimates are fixed at zero and not reliable when the regularized estimates are near zero, even with advanced bootstrap or robust methods (see; Kyung et al., 2010; Section 2 and references therein). As a consequence, when using frequentist lasso, we neither have a measure of uncertainty regarding selection of the active subset of parameters nor reliable standard errors for the included DIF parameters (Liang & Jacobucci, 2020). A third issue is that, in the maximum likelihood framework, an optimal shrinkage parameter needs to be selected through repeated model comparisons or cross-validation over a large set of candidate values. The additional computational burden from successive model fitting can limit extensions to more flexible shrinkage methods such as adaptive lasso or complex MNLFA models.

Advantages and caution in Bayesian regularization

We believe that certain Bayesian regularization methods have the potential to address many of the issues associated with regularized DIF detection by lasso in the frequentist framework (see, e.g., Gelman et al., 2013 for a Bayesian estimation overview). The first advantage of Bayesian regularization is the use of shrinkage priors and their penalty hyperpriors. Bayesian regularization methods (e.g., Brandt et al., 2018; Lu et al., 2016; Park & Casella, 2008) use informative priors on the target parameters to achieve a similar shrinkage effect to that produced by tuning the penalty in frequentist regularization. These shrinkage priors have direct connections in their forms and properties to frequentist procedures, such as between the double exponential (Laplace) prior and frequentist lasso or between the small-variance normal prior and frequentist ridge (see Jacobucci & Grimm, 2018; Tibshirani, 1996).

The use of shrinkage priors brings several strengths to Bayesian regularization. First, unlike frequentist regularization, Bayesian shrinkage methods can estimate the penalty (i.e., the scale parameter of the shrinkage prior) together with the rest of the model by assigning a hyperprior on the penalty parameter. Penalty parameter values sampled from this hyperprior, whose range can be set by the researchers, are used in regularizing the model. Whereas frequentist lasso chooses one model produced by a fixed penalty value according to model selection criteria, Bayesian regularized point estimates such as posterior means average over the distribution of potential penalty parameter values. This allows the model to account for uncertainty in the penalty value, resulting in better out-of-sample inferential performance (Hans, 2009; Leng et al., 2014; Park & Casella, 2008). Second, the penalty prior distribution can be easily modified to adapt to advanced shrinkage methods. If multiple penalty parameters are specified to accommodate differential shrinkage among many regularized coefficients, these penalties can be estimated simultaneously with the model, which reduces computational burden for Bayesian regularization in complex models with a large number of penalized coefficients. For example, it is possible to assign independent Laplace priors to each of the coefficients in an IRT or MNLFA model, rather than having all coefficients penalized by Laplace priors that share the same penalty parameter. Doing so results in Bayesian adaptive lasso, which maintains consistent parameter estimates (Brandt et al., 2018; Kyung et al., 2010; Leng et al., 2014). All the Laplace prior penalties can be estimated simultaneously, in contrast to the frequentist adaptive lasso (Zou, 2006) with which one must sequentially search through each dimension of the regularized parameters for the best penalty value.

A third strength of Bayesian regularization is that it gives all the regularized estimates empirical standard errors and credible intervals from posterior distributions (e.g., Kyung et al., 2010). These provide valid measures of uncertainty in parameter values than those from frequentist shrinkage methods: they do not rely on asymptotic assumptions or bootstrapped samples. Valid standard errors and CIs for DIF estimates under Bayesian regularization improve the validity of researcher's inference about these estimates. Further, they can be used to appropriately quantify and model the uncertainty of latent variable scores (e.g., Liu & Yang, 2018).

Despite the promise of Bayesian regularization, one issue that remains largely unaddressed is what criteria to use for parameter (e.g., DIF effect) selection. Prior studies of Bayesian regularized factor analysis have employed a range of selection criteria to empirically designate parameters as included or excluded from the model. For example, Muthén and Asparouhov (2012) used 95% equal-tailed credible intervals of the regularized estimates (based on percentile), while Pan et al. (2017) relied on 95% highest density intervals. Shi et al. (2017) used rankings of the DIF effect estimate magnitudes, and Feng et al. (2017) directly evaluated whether the posterior point estimates exceeded 0.1 in absolute value. The main reasons for using these criteria appeared to be that they showed desirable empirical performance in the respective simulation studies. It is unclear how to justify extending one of these empirical criteria across study contexts to DIF detection in multi-group IRT or MNLFA models.

Moreover, we caution that many existing selection criteria used in Bayesian regularization are *ad hoc* and not theoretically well supported. Whereas the frequentist lasso estimator conducts variable selection by setting unimportant parameters to zero, Bayesian regularizing priors lose this model selection property because they lack built-in optimal thresholding rules for setting shrunken estimates to zero (Brandt et al., 2018; Hans, 2010; Leng et al., 2014). As a result, the final Bayesian regularized model includes the same number of parameters as the unpenalized model. This issue can be attributed to the property of Bayesian estimation: A continuous Bayesian shrinkage prior such as the double-exponential prior assigns no probability to the event that the true parameter is *equal to*

zero, so there is also zero posterior probability of having a true parameter value at zero. That is, the posterior probability for a parameter θ_j to be zero given data $\boldsymbol{y}$ $h(\theta_j = 0|\boldsymbol{y})$ equals zero (Hans, 2010). Since no parameters are excluded to create alternative model specifications, Bayesian shrinkage priors with empirical selection criteria in fact consider only a single model and do not account for model selection uncertainty (Lykou & Ntzoufras, 2013). This incoherence would lead to DIF selection outcomes that are difficult to justify or interpret. Taken together, the debatable inferential validity and lack of generalizability in Bayesian regularization with empirical selection criteria may run counter to the main purpose of measurement invariance studies, that is to evaluate and improve the external validity of psychological scales. In the next section, we describe an alternative Bayesian regularization approach that addresses these problems.

Coherent DIF selection with spike-and-slab priors

This study proposes to focus on DIF selection using posterior inclusion probability parameters from the spike-and-slab prior as a coherent and generalizable DIF selection alternative. The main idea is to assign prior probability mass to the event $\{\theta_j = 0\}$, so that a space of models is spanned by the events $\{\theta_j = 0\}$ and $\{\theta_j \neq 0\}$ for each regularized parameter. Then we evaluate the posterior probability of each model with its own subset of parameters and decide the most likely one (Lykou & Ntzoufras, 2013; Hans, 2010; Yuan & Lin, 2005; Kruschke, 2010 Ch. 18). Prior distributions that include a parameter selection probability are called "spike-and-slab" priors (SSP), with several specifications available (e.g., Ishwaran & Rao, 2005; Lu et al., 2016; Mitchell & Beauchamp, 1988). Our SSP specification applies an inclusion parameter that quantifies the importance of θ_j parameters and whether they should be entered into the model (Brandt et al., 2018; Kuo & Mallick, 1998; Lykou & Ntzoufras, 2013). This SSP can achieve both lasso-style shrinkage and DIF effect selection when built onto a double exponential prior,

$$\begin{aligned} \theta_j &= \theta_j^* r_j \\ \theta_j^* &\sim dexp(0, (\sigma^2/\tau_j^2)) \\ r_j &\sim Beta(0.5, 0.5) \end{aligned} \tag{3}$$

where *dexp* stands for the Laplace distribution. r_j is the inclusion parameter that represents the probability of having a parameter of interest θ_j^* in the model and is treated with an diffuse Beta prior here. r_j can be seen as a rescaling factor on the penalizing lasso prior; it could alternatively be treated with a Bernoulli prior and represent a binary inclusion decision, but this setup may have convergence and computational disadvantages (Bhadra et al., 2019). In a DIF detection model, σ^2 is the latent factor variance, τ_j^2 is its Laplace penalty parameter (hyperprior not shown), and θ_j is the DIF effect parameter scaled by SSP. Importantly, SSP relies on the estimated marginal posterior inclusion probabilities $\hat{r}_j$ to accomplish DIF selection instead of ad-hoc criteria. Parameter selection decisions are made based on whether the posterior inclusion probabilities exceed a threshold. SSP parameter estimates $E(\theta_j|\boldsymbol{y})$ integrate out $\hat{r}_j$ to average over estimates under different inclusion probabilities, referred to in Lu et al. (2016) as the SSP marginal estimator.[3]

This SSP selection method noticeably differentiates from other empirical selection criteria. By explicitly estimating uncertainty regarding parameter inclusion, the SSP approach yields valid estimates and standard errors that account for model selection uncertainty. The resulting DIF selection is straightforwardly interpreted as having the highest inclusion probabilities in the model and is generalizable across study contexts. In addition, it is possible that an optimal parameter selection result exists for the SSP method. In linear regression setting, predictors selected with SSP inclusion probabilities above a 0.5 threshold provide a model with the optimal predictive error (Barbieri & Berger, 2004; Hans, 2010; Lu et al., 2016). However, in practice the 0.5 cutoff may not be optimal for finite samples, correlated predictors, or complex models (Dey et al., 2008 see Piironen et al., 2020 Section 2 for illustrations; see Narisetty & He, 2014 for alternatives). We evaluate in this study where an optimal inclusion threshold may exist in the DIF detection context. This is possible because the SSP model allows us to examine DIF detection performance over a range of inclusion probability cutoffs rather than a fixed one (Bainter et al., 2020). Doing so does not compromise inference, because changing the inclusion cutoffs does not require model re-estimation or affect other parameter inference in contrast to frequentist lasso. We note that evaluating empirical cutoff values based on data conditions does add some subjectivity to the parameter selection process, but this does not diminish the advantage of using inclusion probability parameters as a more coherent basis for model selection relative to empirical selection criteria like credible intervals. Current literature has not applied SSP in this fashion to measurement model DIF selection.

Summary and research questions

The motivation for this study is to evaluate the use of SSP with inclusion probability parameters (referred to henceforth as the SSP Criterion for DIF detection) in comparison to Bayesian shrinkage priors with empirical selection rules and the frequentist lasso method. Existing lasso-regularized DIF detection methods incur unreliable standard errors and underestimated *p*-values from model re-estimation. Bayesian regularizing priors such as Bayesian adaptive lasso can maintain consistent estimates without re-estimation or unreasonable computational burden. These methods provide theoretically valid empirical standard errors and incorporate uncertainty in penalty value selection. Yet Bayesian regularization priors do not consider the probability of any alternative model specifications, so parameter selection based on post-hoc empirical decision rules does not incorporate model selection uncertainty. This study argues that the SSP Criterion inherits the strength of Bayesian regularization while presenting a coherent and generalizable DIF selection rule for scale assessment. In evaluating

[3]The above SSP parameter selection and estimation procedures have been referred to in a broader context as stochastic search variable selection (e.g., Bainter et al., 2020) and Bayesian model averaging (Raftery et al., 1997).

this argument, we investigate the following research questions concerning DIF detection performance, consistency of DIF estimates, and inferential quality.

(1) We evaluate whether the SSP Criterion and other Bayesian shrinkage methods achieve comparable, if not superior, DIF detection performance relative to the frequentist lasso method, in each case without the use of a priori anchor items. In particular, treating all other model parameters as random variables in Bayesian estimation could decrease DIF detection accuracy for Bayesian shrinkage methods in general relative to frequentist lasso. Further, we explore how to parameterize the inclusion probability in MNLFA models and seek to identify desirable inclusion probability thresholds to improve the outcomes.
(2) We investigate the consistency (decreasing bias as the sample size increases) of the DIF effect estimates from the SSP Criterion without the re-estimation step that sometimes follows frequentist lasso (e.g., re-estimation without penalty after parameter selection). Since the spike-and-slab prior has the same adaptive shrinkage components as Bayesian adaptive lasso (i.e. independent Laplace prior for each regularized parameter), we evaluate whether these estimates are (a) comparably unbiased at large sample sizes as the Bayesian adaptive lasso method; (b) more unbiased at large sample sizes than the non-adaptive Bayesian lasso; and (c) close to the re-estimated DIF effect results from frequentist lasso, where minimum shrinkage bias exists because the model penalty has been relaxed.
(3) We evaluate whether the SSP model has preferable inferential quality in its DIF effect estimates. Since the SSP approach incorporates uncertainty in model selection and penalty values better than other Bayesian shrinkage priors and frequentist lasso, it can be hypothesized that DIF estimates from the SSP approach have better inferential quality as measured by standard error accuracy and interval coverage.

The DIF detection methods being examined include the SSP Criterion, the frequentist lasso method, and a few other Bayesian shrinkage priors that have been studied in past literature with 95% equal-tailed credible intervals as their empirical selection criterion, including the small-variance normal prior, the Laplace prior, the adaptive lasso prior, and the spike-and-slab prior using only the credible interval selection criterion. We study two SSP model specifications to better understand the effect of model complexity on the SSP Criterion performance. They include (a) having more than one inclusion parameter per item adaptive over DIF effects (labeled as SSP_VI, VI as "variable" inclusion according to Hans (2010) and (b) having only one inclusion parameter by each item (SSP with variable inclusion parameters by item, labeled as SSP_VI_bi, VI_bi as "variable" inclusion by item). In the next sections we present two simulation studies using 2-group IRT and MNLFA models to examine our research questions. Prior specifications used in the studies are shown in Table 1.

Study I: DIF detection in a 2-Group IRT model

We present here a DIF detection simulation study in a 2-group IRT model to evaluate our research questions

Table 1. Regularizing prior specifications for simulation.

Prior	Specification	Parameter
Normal	$\pi(\theta\|\sigma^2,\tau) = \prod_{j=1}^{p} \frac{\tau}{\sqrt{2\pi\sigma^2}} e^{-(\tau\theta_j)^2/2\sigma^2}$	$\theta_j \in \{\kappa_j, \omega_j\}$
Laplace	$\pi(\theta\|\sigma^2,\tau) = \prod_{j=1}^{p} \frac{\tau}{2\sqrt{\sigma^2}} e^{-\tau\|\theta_j\|/\sqrt{\sigma^2}}$	$\theta_j \in \{\kappa_j, \omega_j\}$
aLasso	$\pi(\theta\|\sigma^2,\tau) = \prod_{j=1}^{p} \frac{\tau_j}{2\sqrt{\sigma^2}} e^{-\tau_j\|\theta_j\|/\sqrt{\sigma^2}}$	$\theta_j \in \{\kappa_j, \omega_j\}$, $\tau_j \in \{\tau_{\kappa_j}, \tau_{\omega_j}\}$
SSP_VI	$\theta_j = \theta_j^* r_j$ $\theta_j^* \sim dexp(0, (\sigma^2/\tau_j^2))$ $r_j \sim Beta(0.5, 0.5)$	$\theta_j \in \{\kappa_j, \omega_j\}$, $\tau_j \in \{\tau_{\kappa_j}, \tau_{\omega_j}\}$, In IRT model $r_j \in \{r_{\kappa_j}, r_{\omega_j}\}$, In MNLFA model $r_j \in \{r_{age_j}, r_{gender_j}, r_{study_j}\}$
SSP_VI_bi	$\theta_j = \theta_j^* r_j$ $\theta_j^* \sim dexp(0, (\sigma^2/\tau_j^2))$ $r_j \sim Beta(0.5, 0.5)$	$\theta_j \in \{\kappa_j, \omega_j\}$, $\tau_j \in \{\tau_{\kappa_j}, \tau_{\omega_j}\}$, In IRT model $r_j \in \{r_j\}$, In MNLFA model $r_j \in \{r_j\}$

$\{\kappa_j, \omega_j\}$ are the DIF parameters; τ is the penalty parameter that is shared across p items in non-adaptive priors (Normal, Laplace) and varies over DIF parameters in the adaptive priors; note that he adaptive penalties always vary by items and by intercepts and slopes (not by MNLFA covariates) to reduce excessive model complexity; aLasso = adaptive lasso prior; *dexp* is the Laplace distribution; two SSP specifications are studied, with more than one inclusion parameter per item (SSP_VI) and only one inclusion parameter per item (SSP_VI_bi), and they also differ in IRT and MNLFA; r_j is the inclusion probability parameter that varies over DIF effects or only over items.

above. This case comprises of the same data generation model, simulation design factors, and replication data sets as the study on frequentist lasso DIF detection that was conducted by Belzak and Bauer (2020) By re-using this data, we can directly reference the frequentist lasso results from this previous study, in particular, those obtained using the tuning parameter with the best Bayesian Information Criterion (BIC; Belzak and Bauer (2020) found this to be superior to the model tuned under the best Akaike Information Criterion (AIC) in terms of DIF selection performance). We briefly describe the simulation conditions setting below.

Model specification and estimation for Bayesian shrinkage methods

A Bayesian hierarchical 2-group IRT model, expressed by Equation (4), is used to estimate and detect DIF effects.

$$\begin{aligned} Logit\left[P\left(y_{ij}=1|\eta_i,\lambda_{0j},\nu_{0j},\kappa_j,\omega_j;x_i\right)\right] &= \left(\nu_{0j}+\kappa_j x_i\right)+\left(\lambda_{0j}+\omega_j x_i\right)\eta_i \\ \eta_i|\alpha_0,\psi_0,\gamma,\beta;x_i &\sim N\left(\alpha_0+\gamma x_i,\left(\psi_0 exp(\beta x_i)\right)^2\right) \\ \lambda_{0j},\nu_{0j},\gamma,\beta &\sim \pi_{diffuse}(\cdot) \\ \kappa_j,\omega_j|\tau_j &\sim \pi_{reg}(\tau_j) \\ \tau_j &\sim p(\cdot) \end{aligned} \tag{4}$$

where the DIF effect parameters $\{\kappa_j, \omega_j\}$ receive their shared or different regularizing prior distributions $\pi_{reg}(\tau_j)$ according to the prior chosen in Table 1. The baseline factor mean and standard deviation parameters, α_0 and ψ_0, are respectively fixed to zero and one to set the latent factor scale. $p(\cdot)$ is the hyperprior on the penalty parameters. The penalty parameters τ are specified based on previous literature (e.g., Park & Casella, 2008; Shi et al., 2017) to ensure acceptable model convergence rates. For the Normal prior, $\tau^2 = 10$; for the Laplace distributions in all other priors the penalty hyperprior is set to be $\tau^2 \sim gamma(10, 1)$. In this way, the τ^2 prior has a mean of 10 and a large variance to keep the prior flat. Since the reference group latent variance is fixed at 1, these gamma priors at their means gave the same shrunken scale value τ as the Normal prior on the DIF priors.

The Bayesian parameter estimates use the posterior mean. The analysis model is saturated with DIF effects. All DIF effect parameters are freely estimated under shrinkage priors (achieving "approximate model identification"; Muthén & Asparouhov, 2012 while the data generation models designate a specific pattern of DIF parameters. Impact parameters are freely estimated to capture covariate influence on the latent score distributions. The diffuse prior distribution $\pi_{diffuse}(\cdot)$ on the impact and non-DIF item parameters is chosen as a normal prior $N(0, 5^2)$, reflecting

trast group.

Table 2. Population parameter values for the 2-Group IRT case.

				Intercept			Slope		
	One-Item DIF	1/3-Item DIF	1/2-Item DIF	Baseline (ν_{0j})	DIF-Small (κ_j)	DIF-Large (κ_j)	Baseline (λ_{0j})	DIF-Small (ω_j)	DIF-Large (ω_j)
1				0.80			0.70		
2	*	*	*	0.20	−0.4	−1	1.00	−0.2	−0.4
3			*	−0.40	0.4	1.1	1.30	0.2	0.5
4		*	*	−1.00	0.4	1.2	1.60	0.2	0.6
5				−1.60			1.90		
6				−2.20			2.20		
7				0.80			0.70		
8		*	*	0.20	−0.4	−1	1.00	−0.2	−0.4
9			*	−0.40	0.4	1.1	1.30	0.2	0.5
10		*	*	−1.00	0.4	1.2	1.60	0.2	0.6
11				−1.60			1.90		
12				−2.20			2.20		

The 6-item condition includes the first 6 items, and the 12-item condition includes all 12 items. DIF-Small and DIF-Large indicate differential item functioning effect for the small and large DIF magnitude conditions, respectively. Asterisks indicate items whose DIF effect is included in the data generation of a specific cell. The baseline item parameter values represent the mean of a uniform distribution that generates the specific parameter values for each replication. The simulated intercepts ranged $\pm$ .30, and the slopes ranged $\pm$ .15 from these means. This current study re-uses the identical simulation data sets in the original Belzak and Bauer Belzak and Bauer (2020) study.

a default belief that the item parameter value could be zero and helps to control for any potential influence on DIF effect estimation, our interest of study. One exception is that we constrain the item slope parameter to be positive, so that $\lambda_{0j} \sim N^{+}(0, 5^2)$ to avoid sign indeterminacy of the estimates (i.e., factor reflection) that can occur because a Bayesian multiple-group factor model with freely estimated impact and loadings is only locally identified (e.g., Bainter, 2017). In one-factor models, doing so imposes the assumption that all the item responses are scored to be ordered in the same direction as the latent factor.

We fit our Bayesian models with 4 chains of Hamiltonian Markov Chain Monte Carlo (MCMC) samples in the *rstan* package (Carpenter et al., 2017; Stan Development Team, 2019) on 4-core Intel Xeon CPUs on a high-performance computing cluster. Each MCMC chain has 2,000 warmup iterations and 2,000 sampling iterations. Model convergence is defined as having strictly zero divergent transitions in the sampling process and having no "r-hat" indices above 1.01 (Betancourt & Girolami, 2013; Vehtari et al., 2020). These thresholds indicated that the chains successfully explored the defined posterior parameter space and that the four independent MCMC sampling chains agreed with each other without incurring multiple solutions. In practice, convergence rates vary depending on the data and types of prior penalties, but they can be improved by tuning the step sizes of the MCMC chains and the sampler maximum iterations (Betancourt, 2018; Hoffman & Gelman, 2014).

Study I results

The 2-group IRT models achieved reasonable convergence rates ranging from 83% to 100% with an average above 90% for all but the Normal prior. Convergence below 90% occurred mostly in the cells with large DIF, high DIF proportion, and large sample sizes. The Normal prior has about 75% convergence in these cells. Full results on convergence rates are shown in Supplemental Materials. Time taken by individual model replications varied from a few minutes to over an hour, depending on the sample size and number of items. Non-convergence is likely due to sample characteristics and might be avoidable with further tuning. However, we did not modify the tuning, prior specification, or penalty strength post-hoc for individual non-converging datasets so as not to introduce additional sources of variability to the results. Only the converged replications are considered for the results.

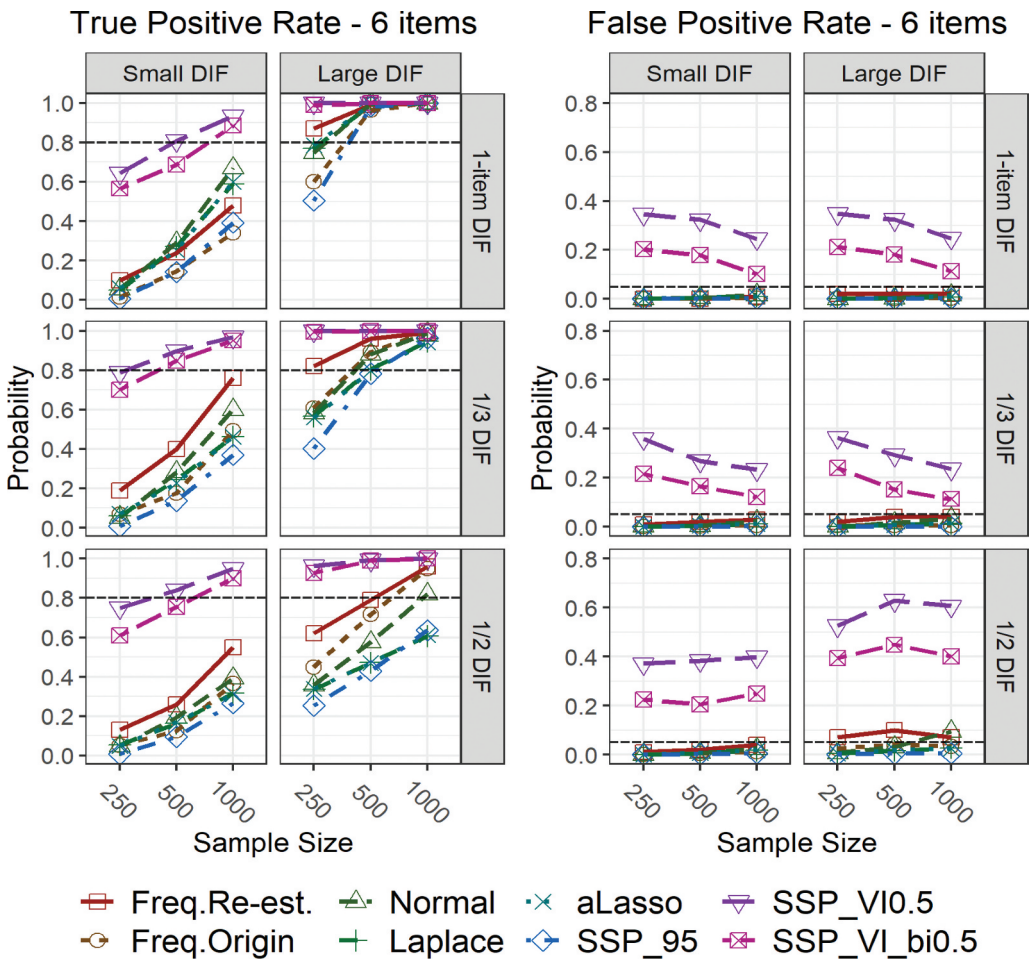


Figure 1. False and true positive rates for frequentist and Bayesian differential item functioning detection in 2-group Item Response Theory (IRT) models with 6 items. Freq.Re-est. = Frequentist lasso IRT model re-estimated without the penalty effect; Freq.Origin = Frequentist lasso best fitted IRT model selected by Bayesian Information Criterion without re-estimation; Normal = Bayesian IRT model with small-variance normal prior; Laplace = Bayesian IRT model with Laplace (lasso) prior; aLasso = Bayesian IRT model with adaptive lasso prior; SSP_95 = Bayesian IRT model with spike-and-slab prior, inclusion parameter, and a 95% credible interval selection criterion; SSP_VI0.5 = Bayesian IRT model with spike-and-slab prior, one inclusion parameter per DIF effect, and an inclusion probability (>0.5) selection criterion; SSP_VI_bi0.5 = Bayesian IRT model with spike-and-slab prior, one inclusion parameter per item, and an inclusion probability (>0.5) selection criterion.

DIF detection false and true positive rates

False and true positive rates are used to examine performance differences among DIF detection methods for Research Question 1. These rates are calculated by first taking the proportions of replications in which each item is detected as having DIF (on either intercept or slope), and then taking the averages of these proportions across items that do not have DIF effects or do have DIF effects for the empirical false and true positive rates, respectively. Figure 1 presents the DIF detection false and true positive rates for the 6-item data conditions, using an inclusion probability of 0.5 or greater for SSP. The 12-item conditions displayed similar patterns and are shown in the supplemental materials. When possible, we included the original frequentist model output estimated with the lasso penalty (labeled as Freq.Origin), whose active DIF subset was selected to have the minimum BIC. We also included the re-estimated frequentist lasso output (labeled as Freq.Re-est) based on this best lasso-selected DIF subset but not the lasso penalty, in order to better distinguish between the performance of the frequentist lasso model and the effect of the penalty on the DIF detection outcome.

The result showed that, when applying a 0.5 inclusion probability threshold, the SSP Criterion incurred notably inflated false positive rates relative to all other frequentist or Bayesian selection methods. Bayesian shrinkage priors using empirical credible intervals for DIF selection performed similarly to the original frequentist lasso model with the penalty effect when the proportion and magnitude of DIF were small and worse when the DIF proportion and magnitude were large, and they almost always had lower true positive rates than the re-estimated frequentist model. The overall result suggested very limited practical applicability for Bayesian shrinkage priors with the empirical interval selection rule.

We next considered alternative inclusion probability thresholds for SSP models. To reduce false positives, we opted to examine a 0.6 inclusion probability cutoff. Figure 2 shows that a 0.6 inclusion probability cutoff effectively reduced false positive errors in the SSP Criterion to a level comparable to the frequentist methods, while achieving better or similar DIF detection power than the frequentist methods. Thus, the SSP Criterion with a 0.6 inclusion cutoff achieved preferable empirical DIF detection performance overall in the 2-group IRT context, outperforming the frequentist lasso model with re-estimation and the Bayesian shrinkage priors that relied on credible intervals for the posterior DIF estimates.

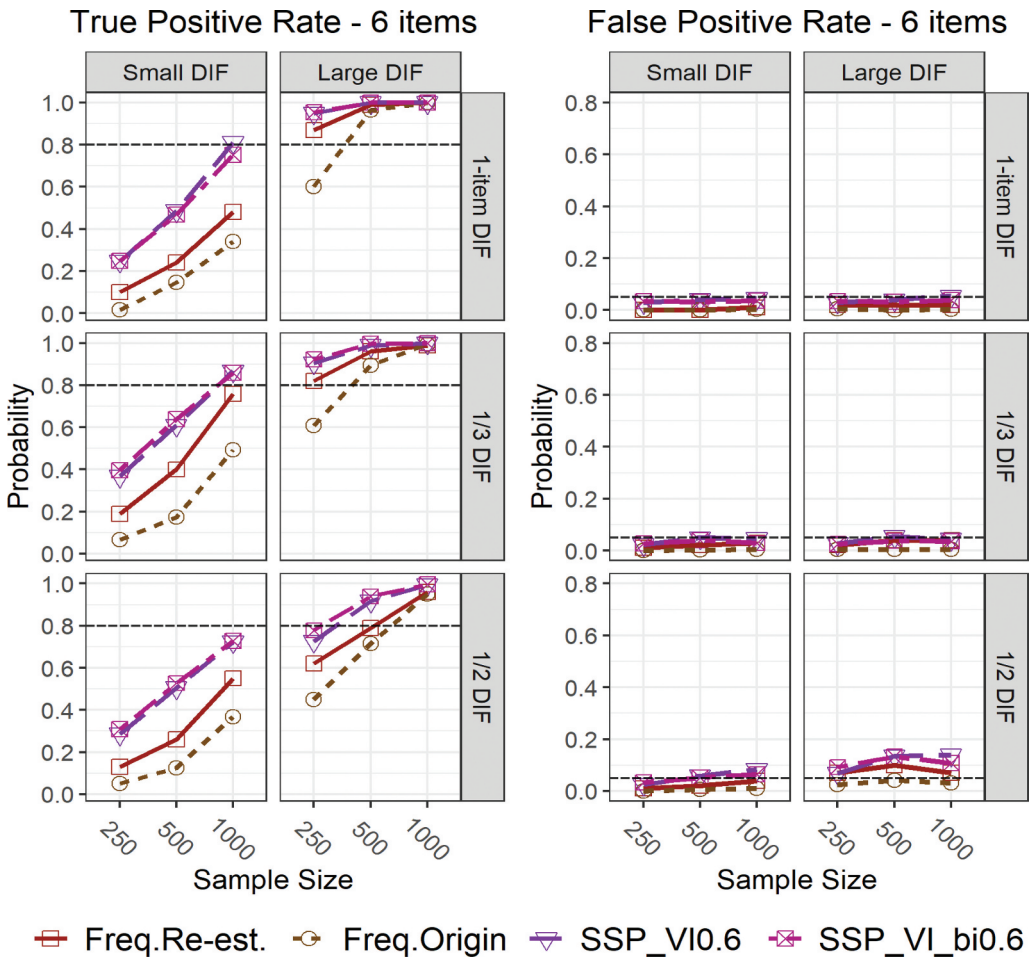


Figure 2. False and true positive rates for the SSP criterion in 2-group Item Response Theory (IRT) models with a 0.6 inclusion threshold and 6 items.
Freq.Re-est. = Frequentist lasso IRT model re-estimated without the penalty effect; Freq.Origin = Frequentist lasso best fitted IRT model selected by Bayesian Information Criterion without re-estimation; SSP_VI0.6 = Bayesian IRT model with spike-and-slab prior, one inclusion parameter per DIF effect, and an inclusion probability (>0.6) selection criterion; SSP_VI_bi0.6 = Bayesian IRT model with spike-and-slab prior, one inclusion parameter per item, and an inclusion probability (>0.6) selection criterion.

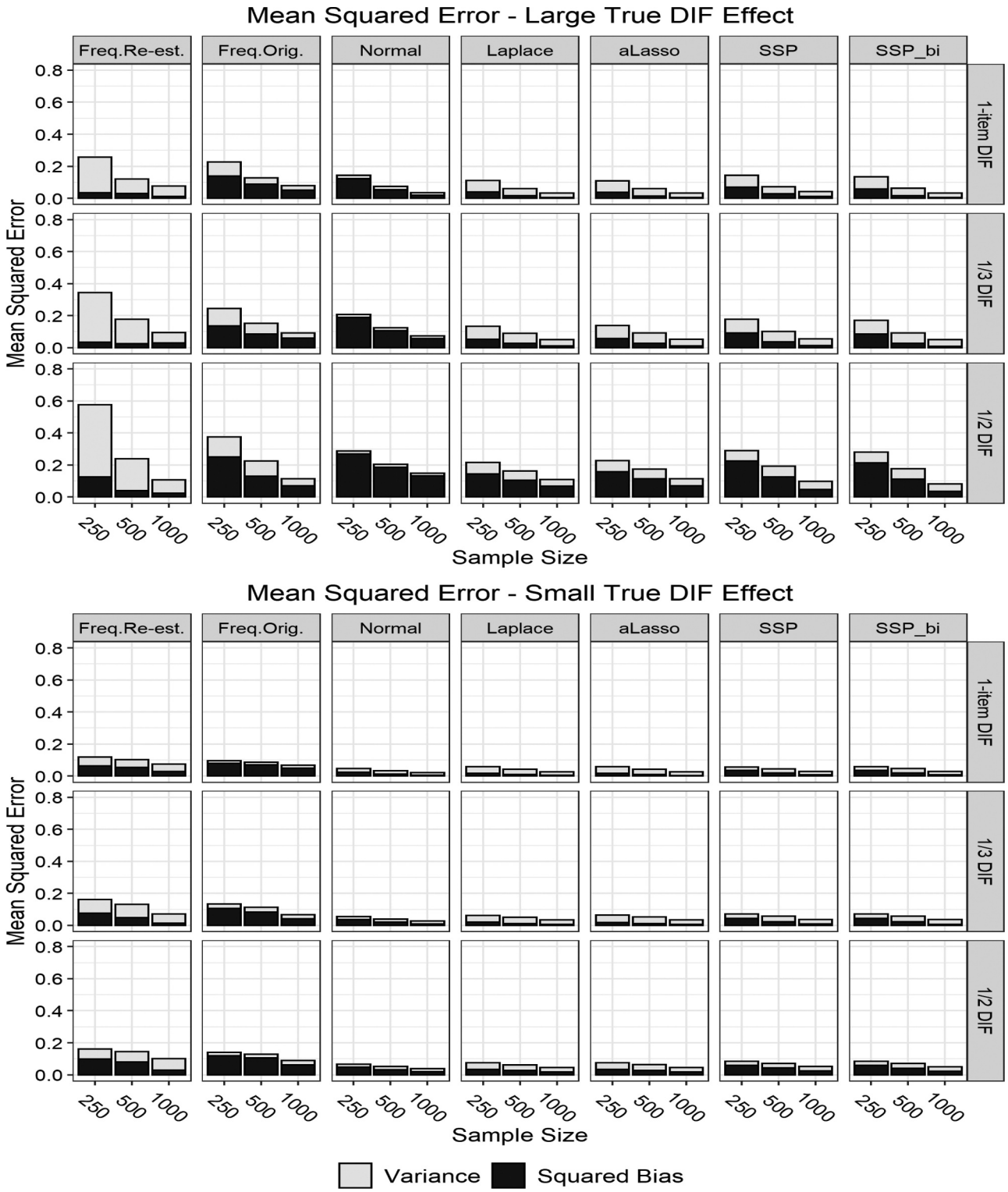


Figure 3. Mean squared error for frequentist and Bayesian differential item functioning effect (DIF) estimates in 2-group item response theory (IRT) models. Each bar represents the MSE per DIF parameter averaged over the 6- and 12-item conditions. Freq.Re-est. = Frequentist lasso IRT model re-estimated without the penalty effect; Freq.Origin = Frequentist lasso best fitted IRT model selected by Bayesian Information Criterion without re-estimation; Normal = Bayesian IRT model with small-variance normal (ridge) prior; Laplace = Bayesian IRT model with Laplace (lasso) prior; aLasso = Bayesian IRT model with adaptive lasso prior; SSP = Bayesian IRT model with spike-and-slab prior and one inclusion parameter per DIF effect; SSP_bi = Bayesian IRT model with spike-and-slab prior and one inclusion parameter per item.

Mean squared error

Mean squared error (MSE) of DIF effect estimates, defined as the squared difference between the DIF estimates and the true population values, averaged over the number of replications and number of DIF parameters in each condition, is used to evaluate parameter estimate consistency among DIF detection

methods for Research Question 2. MSE as decomposed into squared bias and variance is shown in Figure 3. Panels show results from the large and small DIF magnitude conditions respectively, with each bar averaged over the simulation cells with 6 and 12 items. Here we can evaluate the DIF estimation accuracy by comparing the absolute MSE levels of DIF detection methods across all conditions. We can also investigate the DIF estimation consistency by focusing on the MSE levels when the sample size is large.

Overall MSE levels are comparable between frequentist lasso and the Bayesian models, with a slight advantage for the latter. The regularized estimates achieved lower MSE, particularly in small samples, relative to the re-estimated frequentist lasso estimates. However, the re-estimated frequentist lasso provided the lowest bias across most conditions. Among the Bayesian models, the SSP models had slightly higher bias in small samples, but improved their MSE and bias levels more with increasing sample sizes than the models with non-adaptive penalties, suggesting that they benefited from the consistency property of adaptive shrinkage priors.

Standard error accuracy and interval coverage

Research Question 3 investigates the quality of inferences about DIF parameters by evaluating standard error (SE) accuracy and confidence/credible interval coverage rates. The former is calculated as the ratio between the average estimated SE across replications over the empirical standard deviation of the estimates across replications. The latter is calculated as the number of replications where the equal-tailed credible or confidence intervals covered the true population DIF effect values divided by the total number of replications and DIF parameters in a condition. For both measures a ratio closer to one indicates more valid uncertainty estimates for the population DIF effects. However, we first caution that the absolute SE accuracy deviation level in the two frequentist models may not be reliable, because frequentist lasso fixed some DIF estimates at zero, for which no valid standard errors exist. In calculating the SE accuracy, we treated the SE of those DIF estimates left out by lasso as zero. Doing so better accounts for the feature of frequentist lasso as a point estimator that does not have built-in measures for the certainty in selecting active parameter subsets (Kyung et al., 2010). These artificial zero standard errors would bring down both the numerator and the denominator of individual DIF effect SE accuracy, so they may artificially influence the absolute magnitude of frequentist averaged SE accuracy, but they should not easily reverse the direction of the measure .[4]

Here we provide a brief summary of results on SEs, reserving more detailed results to Supplemental Materials to conserve space. Averaging over 6 and 12 item conditions, we found that both the re-estimated model and the original best-fitted frequentist lasso model underestimated standard errors by about 50%. The re-estimated model has further underestimation than the original best-fit model even though the former is not affected by the penalty effects and shares the same active DIF parameter subsets with the latter in each condition. In contrast, the Bayesian DIF selection models overestimated the DIF effect variability. Among the Bayesian models, the small-variance normal prior tends to give SE estimates two times larger than their empirical standard deviation, while the Laplace and adaptive lasso prior overestimated by about 50%. With larger sample size the SSP approach generates more accurate SE estimates than the other priors, overestimating by only 10%-25% at $N = 1000$. This is potentially due to SSP having inclusion probability parameters, but in small samples this advantage diminished.

In calculating coverage rates, frequentist lasso replications that fixed true DIF effects to zero (i.e., false negative) are treated as not covering the population parameter value. Given space limits we briefly summarize the results here and provide more details in the Supplemental Materials. All Bayesian models except for the normal prior provided close to 80% coverage across all conditions, while the frequentist lasso coverage averaged below 50%. Frequentist lasso coverage is confounded with its DIF detection power, but this is not the case for the Bayesian methods. The overestimated Bayesian SE compensated for the estimation bias by providing wide credible intervals and thus increased coverage.

Summary

For the 2-group IRT models, we observed that the SSP Criterion with one inclusion parameter per item and an inclusion probability threshold of 0.6 is adequate in preserving superior DIF detection power relative to the frequentist lasso model while mitigating the excessive false positive rate observed with a 0.5 threshold. On the other hand, DIF selection using shrinkage priors and empirical credible intervals has limited empirical power, either comparable to or worse than the lasso-penalized frequentist model. Bayesian models performed similarly to the frequentist lasso model with respect to the DIF estimation accuracy. With smaller sample sizes, Bayesian model tended to produce DIF effect estimates with higher bias but similar or lower overall MSE levels than the re-estimated frequentist lasso model. With large sample sizes the bias levels of the SSP models resemble the re-estimated lasso. With respect to inference, the SSP Criterion model provided more accurate standard errors for DIF estimates relative to other Bayesian shrinkage priors in most conditions, whereas the frequentist lasso model underestimated the standard errors. The Bayesian methods also produced superior coverage relative to the frequentist lasso. To determine how well these results generalize, we next compare these regularization methods with more complex MNLFA models.

[4]The alternative, to omit zero DIF estimates from accuracy calculations, would leave too few DIF estimates available across replications, which would cause instability in calculating the empirical standard deviation and SE accuracy (often giving accuracy values above 3). Additionally, in frequentist lasso conditions, only estimates considered non-zero at least 5 times across all replications are included in the accuracy calculation to avoid having minuscule empirical parameter SD and outlier SE accuracy incidents.

Table 3. Population parameter values for the MNLFA case.

	6 Items		12 Items		Intercept (Small DIF | Large DIF)				Slope (Small DIF | Large DIF)			
Item	DIF 33%	DIF 66%	DIF 33%	DIF 66%	Baseline (v_{0j})	Age	Gender	Study	Baseline (λ_{0j})	Age	Gender	Study
1					−.5				1			
2		*	*	*	−.9	.125|.25	−.5|-1	.5|1	1.3	.05|.075	−.2|-.3	.2|.3
3		*	*	*	−1.3	−.125|-.25	.5|1	.5|1	1.6	−.05|-.075	.2|.3	.2|.3
4	*	*	*	*	−1.7	.125|.25			1.9	.05|.075		
5	*	*	*	*	−2.1		−.5|-1	.5|1	2.2		−.2|-.3	.2|.3
6					−2.5				2.5			
7					−.5				1			
8				*	−.9	.125|.25	−.5|-1	.5|1	1.3	.05|.075	−.2|-.3	.2|.3
9				*	−1.3	−.125|-.25	.5|1	.5|1	1.6	−.05|-.075	.2|.3	.2|.3
10				*	−1.7	.125|.25			1.9	.05|.075		
11				*	−2.1		−.5|-1	.5|1	2.2		−.2|-.3	.2|.3
12					−2.5				2.5			

The 6-item condition includes the first 6 items, and the 12-item condition includes 12 items. The 24-item condition repeated the parameter values of these first 12 items. The 24-item 33% DIF condition has the identical DIF pattern with the 12-item 66% DIF condition, and the 24-item 66% DIF case repeats the DIF pattern of the 12 items in the 12-item 66% DIF condition. DIF-Small and DIF-Large indicate differential item functioning effect for the small and large DIF magnitude conditions, respectively. Asterisks indicate items whose DIF effect is included in the data generation of a specific cell. The item parameter values are fixed when generating the data in each replication. This current study re-uses the identical simulation data sets in the original Bauer et al. Bauer et al. (2019) study.

Study II: DIF detection in a MNLFA model

MNLFA models extend beyond multi-group IRT models to allow for DIF from multiple covariates which may be either continuous or categorical (Bauer, 2017). These models have wide applicability but have been used frequently for scale equating when pooling data across multiple studies for the purpose of integrative data analysis (Hussong et al., 2013). To examine the feasibility of the current Bayesian SSP method, we analyze the previously simulated data described by Curran et al. (2016) for which a subset of conditions were analyzed using the frequentist lasso in Bauer et al. (2019). We again compare to the best-BIC lasso results.

Data generation

The background covariates for this simulation are Study (50% in Study 1, 50% in Study 2), Gender (50% male), and Age (ranging from −2.67 to 2.67 after standardization). The binary covariates are effect coded, and all three covariates have variances of 1. The data-generation model applies correlated covariate values to Equations 1 and 2 to produce item parameters and binary item responses. Simulation conditions include Sample Size (500, 2000), Number of Items (6, 12, 24), Proportion of Items with DIF (1/3 items, 1/2 items), and Magnitude of DIF (small, large), for a total of 24 distinct cells. Each cell has 100 replication data sets. The specific data-generating population parameters for the item responses from the original studies are detailed in Table 3. The population parameters for the impact equations for the mean and standard deviation are shown below, following the notations of Equation 1 and 2.

$$\begin{aligned}\alpha_i &= .51 \times Age_i + .56 \times Study_i - .21 \times Study_i \times Age_i \\ \psi_i &= \sqrt{.65} \times exp(.1875 \times Age_i - .025 \times Gender_i \\ &\quad +.025 \times Study_i)\end{aligned} \tag{5}$$

Model specification and estimation

The MNLFA model fitted to the data is expressed in Equation 4. The difference from the 2-group IRT model is that instead of a single group membership covariate or scalar DIF parameters, the MNLFA model has a $q \times 1$ vector of covariates $\boldsymbol{x}_i$ and $q \times 1$ vectors of DIF and impact parameters $\boldsymbol{\kappa}_j, \boldsymbol{\omega}_j, \boldsymbol{\gamma}, \boldsymbol{\beta}$. The prior specifications followed Table 1. The small-variance normal prior and spike-and-slab prior using empirical 95% intervals for DIF selection are omitted here to conserve space, as they performed highly similarly to Bayesian lasso and adaptive lasso in Study I. Since the MNLFA model has about three times more DIF parameters and a continuous covariate with wider range than the IRT models, stronger shrinkage was needed to identify the MNLFA model and achieve adequate convergence rates. After exploring different levels of shrinkage in individual data-sets, we decided to rescale all MNLFA shrinkage prior penalties to 1/30 of their counterparts from Study I to maintain reasonable model convergence. That is, $\tau^2 = 300$ for the MNLFA model Normal prior and $\tau^2 \sim gamma(300, 1)$ for the Laplace distribution penalty hyperprior. Other model specification decisions mimic Study I.

Study II results

Convergence rate discrepancies exist more notably among the Bayesian penalized MNLFA models than the 2-group IRT models. Among the 24 data conditions for the Bayesian methods, the SSP Criterion model with one inclusion parameter per item (SSP_VI_bi) again achieved an average and a median convergence rate above 95%, but the SSP Criterion model with one inclusion parameter per DIF effect (SSP_VI) has around 60% convergence in three 2000-sample cells. The adaptive lasso prior had an average 77% and a median 79% convergence rate, while Bayesian lasso model only had an average 62% and a median 67%. These divergent replications can often be salvaged by model tuning, stronger penalties, or even different algorithm starting values, and many low convergence cases appeared in the 2000-sample large DIF conditions. Overall, the divergence can be attributed to the increased difficulty maintaining an approximately identified model with the shrinkage priors as sample sizes and model complexity increase. While it is possible to improve the overall convergence with stronger prior penalties, the resulting estimation bias will also increase. In practice, it is

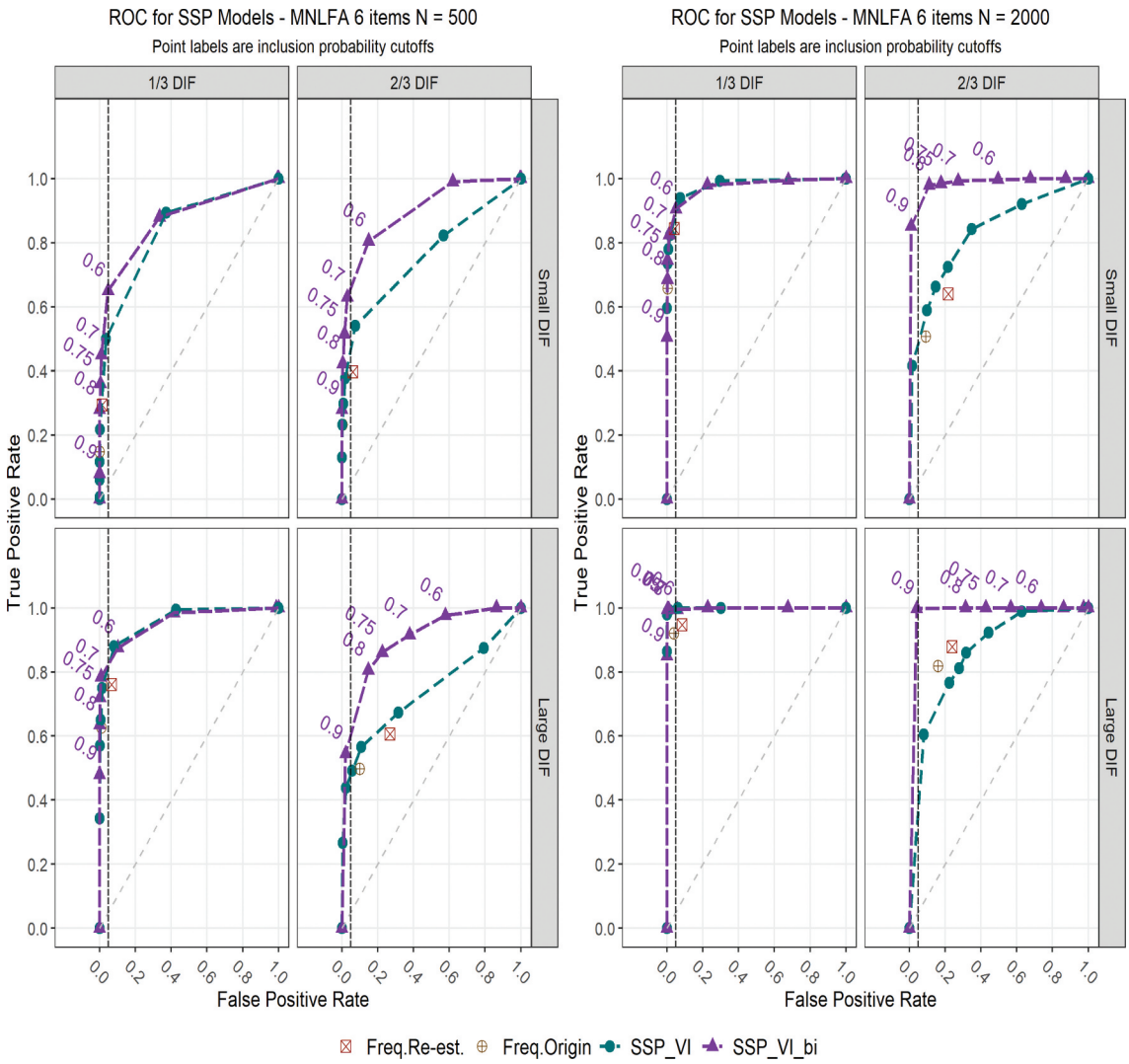


Figure 4. Receiver operating characteristic curves for Bayesian differential item functioning effect detection methods in moderated nonlinear factor analysis (MNLFA) models. Frequentist lasso performance is pointed out for reference. Vertical line is 0.05 false positive rate. Each data label indicated the respective inclusion probability threshold. Freq.Re-est. = Frequentist lasso MNLFA model re-estimated without the penalty effect; Freq.Origin = Frequentist lasso best fitted MNLFA model selected by Bayesian Information Criterion without re-estimation; SSP = Bayesian MNLFA model with spike-and-slab prior and one inclusion parameter per DIF covariate for each item; SSP_VI_bi = Bayesian MNLFA model with spike-and-slab prior and one inclusion parameter per item.

worth exploring the appropriate magnitude of shrinkage that balances model identification and estimation accuracy. The current study only considered converged replications. Since we controlled for the strength of penalties and model tuning in these Bayesian models, these differential convergence rates (shown in Supplemental Materials) could indicate that the SSP Criterion models worked better than common shrinkage priors in complex MNLFA models.

DIF detection false and true positive rates at optimal performance

For the MNLFA models an item can have DIF effects on one covariate but not on another covariate, so false and positive rates were calculated by averaging across all item-covariate combinations that do not or do have DIF, except for the SSP_VI_bi model, where DIF detection could be done only by item. Similar to the 2-group IRT case, the shrinkage priors using credible intervals (lasso and adaptive lasso) had about the same false and true positive rate as the original frequentist lasso model. The SSP Criterion incurred increased false positive rates even at a 0.6 inclusion probability cutoff relative to the frequentist model. To determine whether the SSP Criterion can achieve a more desirable trade-off between the false and true positive rates, we evaluate a range of inclusion probability thresholds. Figure 4 shows receiver operating characteristic (ROC) curves that depict this trade-off at varying SSP inclusion probability thresholds for the 6-item conditions. The frequentist model performance is also indicated for reference.

Overall, the range of desirable SSP inclusion thresholds that gave the optimal DIF detection performance (closest to perfect detection performance at the top-left corner) varied depending on the data conditions. The magnitude and proportion of DIF effects in the data, as well as the sample size, all influence the value of the optimal inclusion thresholds. When the proportion of items with DIF was small, an inclusion probability cutoff of about 0.7 consistently kept the false positive rate below 0.05 and preserved an effective true positive rate.

When the proportion of items with DIF was large, an inclusion threshold of 0.9 appeared to be optimal given a large sample size but less so at a smaller sample size. In the large-sample and high DIF-proportion conditions and using the SSP model with one inclusion parameter per item, we can improve upon the frequentist lasso method and achieve the most desirable DIF detection performance at a conservative 0.9 inclusion probability cutoff. In most conditions, the true positive rates are comparable or more advantageous than the frequentist result when the false positives are controlled at the respective 0.9 or 0.7 inclusion thresholds. This pattern persisted in the 12- and 24-item conditions, which are shown in Supplemental Materials.

MSE

Figure 5 shows the MSE levels from the MNLFA models with large and small DIF magnitude conditions respectively, averaged over 6

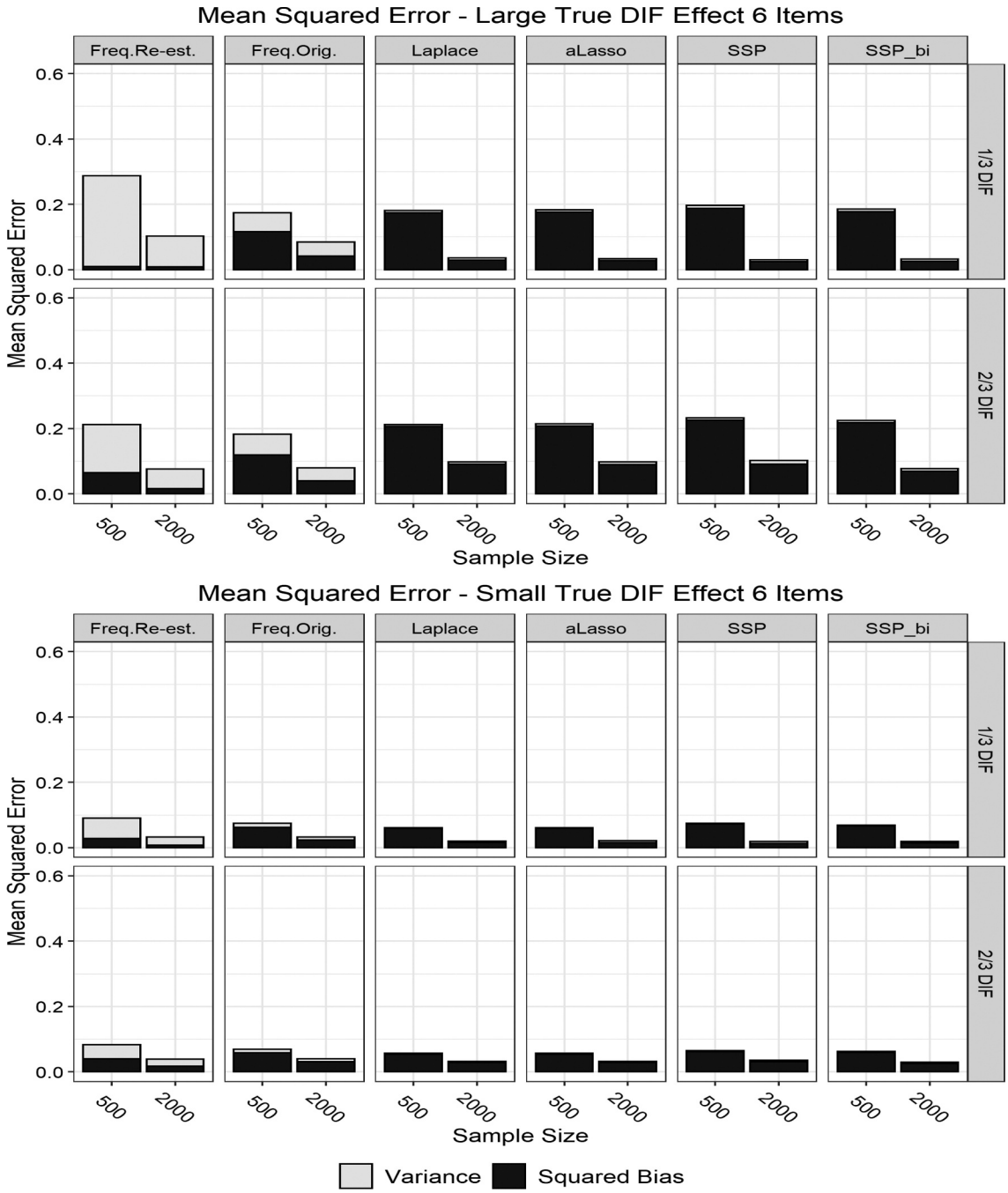


Figure 5. Mean squared error for frequentist and Bayesian differential item functioning (DIF) effect estimates in moderated nonlinear factor analysis (MNLFA) models with 6 items. Freq.Re-est. = Frequentist lasso MNLFA model re-estimated without the penalty effect; Freq.Origin = Frequentist lasso best fitted MNLFA model selected by Bayesian Information Criterion without re-estimation; Laplace = Bayesian MNLFA model with Laplace (lasso) prior; aLasso = Bayesian MNLFA model with adaptive lasso prior; SSP = Bayesian MNLFA model with spike-and-slab prior and one inclusion parameter per DIF covariate for each item; SSP_bi = Bayesian MNLFA model with spike-and-slab prior and one inclusion parameter per item.

items. The 12- and 24-item conditions display similar patterns and are shown in Supplemental Materials. The overall true DIF MSE levels of the Bayesian models were either comparable to or improved upon the frequentist re-estimated lasso model, depending on whether the proprotion of DIF was high or low. Similar to the IRT case, the frequentst lasso model still had the lowest bias level, whereas the Bayesian MSE levels are almost entirely due to bias rather than variance, suggesting strong shrinkage due to the priors. Increased bias was observed when the magnitude of DIF was large, with a more pronounced increase for Bayesian methods versus frequentist lasso. No noticeable MSE level difference existed among the Bayesian models. Bayesian models with adaptive penalties did not achieve a notable decrease in bias relative to non-adaptive ones, likely because our adaptive penalties vary only by item intercept and slope but not by covariate effects within each item. This specification reduces model complexity but may not be flexible enough when there are multiple covariates.

Standard error accuracy and interval coverage

Comparing to the 2-group IRT case, the advantage of the SSP models in uncertainty measures decreased to some extent. The Bayesian models again produced overestimated standard errors, while the frequentist models produced underestimated ones. At a sample size of 500, the Bayesian model standard errors were twice as large the empirical standard deviations of the estimates, whereas the frequentist model standard errors were about half the size of the empirical standard deviations. SSP models did not outperform other Bayesian priors in SE accuracy. When the sample size reached 2000, the Bayesian SEs improved more than the frequentist ones, to about 1.5 times the SDs of the estimates, and the SSP models produced slightly more accurate SEs than other shrinkage priors.[5] These trends are consistent with what we observed in the 2-group IRT case.

Coverage rates for MNLFA true DIF effect estimates similarly favored Bayesian methods over the frequentist one, as a result of the overestimated Bayesian SE. At a 2000 sample size, the Bayesian models consistently produced about 75% coverage, while the frequentist model coverage ranged from 40% to 60%. Coverage rates were most similar when the DIF magnitude and proportion were both large. See Supplemental Materials for the complete results of these two sections above.

Non-DIF parameter estimates

It is often the case that we are interested in the baseline item parameter estimates and impact estimates together with the DIF effects in MNLFA models (Bauer, 2017). We evaluated the MSEs of these non-DIF estimates to better understand regularized MNLFA model performance. The Bayesian item parameter estimates showed slightly higher bias (to the first decimal place) and variance than the frequentist re-estimated model when the DIF effect magnitude is large. This increase in bias disappeared when the magnitude of DIF effects was small. The Bayesian impact estimates showed higher mean impact bias (to the second decimal place) but comparable overall mean and SD impact MSE levels relative to the frequentist estimates in the large DIF magnitude conditions. In the small DIF magnitude conditions, the Bayesian mean and SD impact estimates showed comparable bias and overall MSE levels as the frequentist estimates. (The related figure is presented in Supplemental Materials.) We posit that the bias of the non-DIF estimates can be attributed to shrinkage bias on DIF estimates, which would explain the difference between the large and small DIF magnitude conditions. Overall, the shrinkage effect on the DIF estimates seems to produce reasonable impact estimates while adding to a small bias in the item parameter estimates. The Bayesian item estimate bias is perhaps consequential to item calibration in a regularized Bayesian MNLFA model only when there is a small sample size and a large magnitude of DIF effects.

Identifying one anchor item

Rather than estimating all DIF effects at once, an alternative scale evaluation strategy is to identify at least one invariant item as the anchor, so that researchers can estimate a measurement model without penalization (as in Curran et al., 2016) and conduct conventional DIF testing methods such as likelihood ratio tests.[6] Since the Bayesian models did not fix any DIF estimate to exactly zero, it is possible to determine one item as the "most likely invariant" in the Bayesian context. For the non-SSP Bayesian models, we determined an item as the most likely invariant one in each replication if all of its DIF effect credible intervals included zero and its sum of absolute DIF effect point estimates was the smallest among all items, which can be viewed as a empirical proxy of how closely centered the DIF parameter posterior distributions were around zero. For the SSP models, the most likely invariant item was determined as having the smallest DIF effect inclusion probability (which was a single estimate for the SSP_VI_bi model and a product of DIF probability estimates for the SSP_VI model). Our evaluation results showed that, in the small DIF small sample size conditions, the SSP models achieved about 90% success (true positive) rates in identifying a correct anchor item, while the other Bayesian models had 80–85% success rates. Across all other data conditions both the SSP models and other Bayesian models achieved over 95% success rates. The full result is shown in the Supplemental Materials. In addition to its more reliable performance, the SSP invariant item selection method has a more straightforward probabilistic interpretation than the empirical method in other Bayesian models. Choosing the most likely invariant item and then applying conventional DIF testing methods could provide a useful cross-reference to our simultaneous DIF evaluation strategy, though we note that this alternative involves model re-estimation in the same data without accounting for model

[5]The low convergence rates in Bayesian lasso and adaptive lasso may have contributed to downward bias (better accuracy) on their SE accuracy and upward bias on their coverage results, because the replications with less regular data (e.g., correlated DIF covariate effects, weaker effect sizes) are likely to cause divergence in the simulation without more detailed model tuning. These replications may have been more demanding to estimate and, if not left out of the results, added to SE inaccuracies or coverage error.

[6]We thank an anonymous reviewer for this suggestion.

selection uncertainty and this may affect inference in the same way as the frequentist lasso method.

Study summary

Our first research question evaluated the DIF detection performance of the SSP Criterion and other Bayesian shrinkage methods using empirical intervals in comparison to the frequentist lasso. Our results indicated that, when appropriate DIF inclusion probability cutoffs were chosen, the SSP Criterion with one inclusion parameter per item provided superior DIF detection power relative to the frequentist lasso and the empirical Bayesian shrinkage methods, while keeping the false positive errors below 0.05.

We found the SSP Criterion to be susceptible to false positives at the 0.5 probability cutoff used in past parameter selection literature, especially when the DIF magnitude and proportion were high, or when simultaneous covariate DIF effects were present in MNLFA models. However, we observed for 2-group IRT models consistently that SSP with one inclusion parameter per item and an inclusion probability threshold of 0.6 mitigated the excessive false positives and maintained desirable true positive rates. In the MNLFA models, depending on if the proportion of DIF items was small or large, a 0.7 or 0.9 inclusion threshold achieved near optimal DIF selection, outperforming the frequentist method. These performance patterns were notably consistent across DIF magnitudes and numbers of item conditions when the sample size was large. Although most commonly used in previous studies of Bayesian regularization, DIF detection using shrinkage priors and empirical credible intervals showed limited empirical power. The performance of this approach resembled that of the frequentist lasso model without re-estimation, which can be attributed to the penalty effect in the model and the overestimated SE interfering with the interval coverage.

Our second research question investigated the consistency property of the SSP Criterion and other adaptive shrinkage priors, i.e. how the estimation bias of the SSP model changes as the sample size increases. Our results showed that the SSP model, when applied to all covariates as in the IRT case, benefits from the consistency property of its adaptive penalty and incurs slightly less bias on the DIF estimates relative to the non-adaptive shrinkage priors in large samples. We also observed that the Bayesian regularized 2-group IRT and MNLFA models achieved overall comparable MSE levels but higher bias levels compared to the re-estimated frequentist lasso.

Specifically, in the 2-group IRT case, Bayesian models achieved a smaller (to one decimal place) MSE for DIF estimates at a small sample size and comparable (to two decimal places) MSE at a large sample size relative to the re-estimated frequentist lasso. In the more complex MNLFA case, the MSE for the Bayesian DIF estimate was only smaller than the frequentist estimates when the DIF magnitude was large, the proportion of DIF was small, and the sample size was small. In other conditions, the MSEs for the Bayesian estimates were comparable with the MSEs of the frequentist lasso. The MSEs of the Bayesian estimates are mostly composed of bias and no variance, displaying the "bias-variance trade-off" of regularized estimates (Hastie et al., 2009 Ch. 2.9) that might result from a strong penalty and might benefit subsequent model inferences. The SSP Criterion model performed comparably to the other shrinkage priors in terms of estimation quality, but the SSP model converged more easily in complex MNLFA models thanks to its use of inclusion parameters.

Our third research question evaluated the inferential quality of the SSP models as measured by standard error accuracy and coverage rates. The results demonstrated a preferable quality of inferences for the SSP model relative to other DIF detection models. We observed that the SSP model provided the most accurate DIF SEs in most 2-group IRT model conditions and in MNLFA models with larger sample sizes; the SSP and other Bayesian shrinkage models provided adequate coverage rates across all data conditions.

Specifically, for the SSP model, the use of inclusion probability parameters properly incorporated model selection uncertainty and improved SE accuracy, even though MNLFA model complexity may have impeded this improvement in smaller sample sizes. The frequentist re-estimated lasso model underestimated DIF effect SEs, because the re-estimation relied on the active lasso-selected DIF subset and failed to account for model selection uncertainty, corresponding to previous literature (e.g., Draper, 1995). All the Bayesian DIF selection models overestimated the DIF effect SEs, which could stem from the additional variability introduced into the model by the Bayesian priors on non-DIF parameters. Empirical credible intervals using these inflated SEs are too wide to have good DIF selection power. In terms of coverage, the SSP model and other Bayesian models provided similar and preferable DIF effect interval coverage over the frequentist method across most data conditions, particularly when the DIF magnitude was small (about 25% higher). As the magnitude of DIF increased, the coverage difference narrowed. Although these Bayesian DIF credible intervals could be too wide to inform of the true DIF values, they were stable and not confounded with DIF selection power. In contrast, because true DIF effects fixed at zero by frequentist lasso would not be counted as having coverage, low power translated into low coverage for this model in the small DIF conditions.

Discussion and future directions

We proposed applying the spike-and-slab prior with inclusion parameter (the SSP Criterion) as a theoretically coherent Bayesian shrinkage method with valid uncertainty measures for conducting DIF effect selection in multi-group IRT and more complex MNLFA models. The SSP Criterion uses the inclusion parameters to characterize the probability of DIF effects being included in the model. The SSP Criterion overcomes the issue of problematic standard errors and model re-estimation following frequentist lasso, as well as the failure to consider model selection uncertainty in using Bayesian shrinkage priors with empirical parameter selection rules (e.g., credible intervals). These theoretical advantages motivated us to consider the SSP approach as an improvement upon existing DIF detection methods. Our studies demonstrated that, compared with the frequentist lasso and other Bayesian shrinkage models, the SSP Criterion using one inclusion parameter per item can achieve more effective DIF selection, reasonable DIF and item parameter estimates,

and more reliable SEs, particularly as sample size increases. These more accurate estimates could facilitate DIF assessment in large scale studies (such as PISA; Kelava et al., 2014) or factor score computation in large-sample integrative data analyses (Curran et al., 2016). In fitting a SSP-regularized MNLFA model for scale assessment, researchers should examine a range of inclusion probability thresholds during parameter selection as a sensitivity measure; overlapping sets of selected DIF effects under various inclusion cutoffs would enhance the detection outcome reliability (Bainter et al., 2020). In some contexts, such as when new items can be generated at low cost, false positives may be less concerning than false negative. In such cases, the final inclusion cutoff value can be chosen lower to minimize the possibility of not excluding an item with true DIF effect. Researchers can also formally study where an optimal range of inclusion thresholds may exist in their measurement model contexts and data conditions. We believe these are valid practices because the SSP model has past theoretical support regarding the existence of optimal performance (under settings such as linear regressions; Barbieri & Berger, 2004; Dey et al., 2008), and because examining the probability thresholds does not require model re-estimation or causes inferential issues.

Downsides also exist for the SSP Criterion. First, the SSP model produce similar overall MSE but more in-sample shrinkage bias than the re-estimated frequentist lasso. We see this as a reasonable trade-off from using the shrinkage prior and accounting for the uncertainty in penalty and DIF effect selection; the consistency property of the SSP model adaptive shrinkage may also partly compensate for the bias in large sample sizes. Second, the need to choose an inclusion probability cutoff introduces a certain level of arbitrariness in DIF modeling, such as when the DIF proportion is a priori unknown to researchers in fitting an MNLFA model. However, we believe that (a) varying the inclusion probability within a reasonable range (e.g., $0.7 - 0.9$) as mentioned above provides a measure of sensitivity and can add to the robustness of the conclusions; (b) the inclusion thresholds can be reasonably generalized across data and known model conditions examined in our simulation studies, and additional studies may extend our findings to alternative analysis scenarios such as having DIF covariates with a different scale or models with other link functions; and (c) the SSP approach possesses theoretical advantages and thus better validity over the existing frequentist lasso DIF detection or the Bayesian shrinkage method with credible intervals. It retains better external validity across study contexts, in contrast to other empirical Bayesian selection criteria such as credible intervals (which only suggests the range of likely parameter values) or the absolute magnitude of the point estimates (which may vary depending on the parameter set examined). We think the above advantages of the SSP Criterion outweigh its downsides and follow from the fact that it considers an appropriate decision space spanned by whether to include each of the concerned parameters, while maintaining valid inference on the regularized estimates. These advantages and downsides encourage further examination and application of the SSP Criterion. To facilitate use of this approach, a data analysis example with the SSP model using R (R Core Team, 2019) is included in the supplemental materials.

We see several ways in which the current study could be expanded in future research. First, the number of simulation conditions we considered is by no means exhaustive for determining optimal DIF inclusion thresholds. Overlooked conditions such as different DIF effect magnitudes and directions could better inform researchers regarding what range of optimal inclusion thresholds to expect in their assessment. Second, this study did not consider the influence of model misspecification other than the inclusion of extraneous DIF parameters. We expect model misfit to be an issue in an exploratory use of measurement models. Third, our study applied frequentist definitions of false positives rates and power but did not investigate other typical Bayesian hypothesis testing tools such as Bayes factor or posterior model probability (Kruschke, 2010 Ch. 18). Though the frequentist measures applied to Bayesian models suited the DIF evaluation purpose and have been used in this way in the past (e.g., Rubin, 1984), the alternative Bayesian testing framework is worth considering. Finally, more studies are needed to better understand the potential benefits the SSP model may bring to out-of-sample inference, such as item calibration or scoring latent traits.

Acknowledgments

We thank David Thissen, Patrick Curran, and two anonymous reviewers for helpful suggestions on this work. Correspondence concerning this article should be addressed to Siyuan Marco Chen, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, 235 E. Cameron Avenue, Chapel Hill, NC 27599-3270. E-mail: mchen@unc.edu

ORCID

Siyuan Marco Chen http://orcid.org/0000-0002-3346-5424
William M. Belzak http://orcid.org/0000-0001-6594-1651

References

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*, 277–300. Retrieved December 4, 2019. from www.jstor.org/stable/1435429

Bainter, S. A. (2017). Bayesian estimation for item factor analysis models with sparse categorical indicators. *Multivariate Behavioral Research*, *52*, 593–615. https://doi.org/10.1080/00273171.2017.1342203

Bainter, S. A., McCauley, T. G., Wager, T., & Losin, E. A. R. (2020). Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain [Publisher: SAGE Publications Inc]. *Advances in Methods and Practices in Psychological Science*, *3*, 66–80. https://doi.org/10.1177/2515245919885617

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*, 870–897. https://doi.org/10.1214/009053604000000238

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*, 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2019). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–13. https://doi.org/10.1080/10705511.2019.1642754

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*, 101–125. https://doi.org/10.1037/a0015583

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning [Publisher: American Psychological Association]. *Psychological Methods*, *25*, 673–690. https://doi.org/10.1037/met0000253

Betancourt, M. (2018). *A conceptual introduction to hamiltonian monte carlo*. *arXiv:1701.02434 [stat]*. Retrieved January 7, 2019, from http://arxiv.org/abs/1701.02434

Betancourt, M., & Girolami, M. (2013). *Hamiltonian monte carlo for hierarchical models*. *arXiv:1312.0906 [stat]*. Retrieved January 7, 2019, from http://arxiv.org/abs/1312.0906

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2019). Lasso meets horseshoe: A survey [Publisher: Institute of Mathematical Statistics]. *Statistical Science*, *34*, 405–427. https://doi.org/10.1214/19-STS700

Brandt, H., & Bauer, D. J. (2020). *Bayesian penalty methods for testing measurement invariance in moderated nonlinear factor analysis*. Manuscript in preparation. Department of Psychology, University of Zurich.

Brandt, H., Cambria, J., & Kelava, A. (2018). An adaptive bayesian lasso approach with spike-and-slab priors to identify multiple linear and nonlinear effects in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 946–960. https://doi.org/10.1080/10705511.2018.1474114

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1. https://doi.org/10.18637/jss.v076.i01

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the horseshoe. *Artificial Intelligence and Statistics*, *5*, 73–80. http://proceedings.mlr.press/v5/carvalho09a/carvalho09a.pdf

Chen, Y., Thissen, D., Anand, D., Chen, L. H., Liang, H., & Daughters, S. B. (2019). Evaluating differential item functioning (DIF) of the Chinese version of the behavioral activation for depression scale (c-BADS). *European Journal of Psychological Assessment*, *36*, 1–21. https://doi.org/10.1027/1015-5759/a000525

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 827–844. https://doi.org/10.1080/10705511.2016.1220839

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., & Hussong, A. M. (2018). Recovering predictor–criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 860–875. https://doi.org/10.1080/10705511.2018.1473773

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., Sher, K., & Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behavioral Research*, *49*, 214–231. https://doi.org/10.1080/00273171.2014.889594

Dey, T., Ishwaran, H., & Rao, J. S. (2008). An in-depth look at highest posterior model selection. *Econometric Theory*, *24*, 2. https://doi.org/10.1017/S026646660808016X

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 45–97. Retrieved December 4, 2019, from www.jstor.org/stable/2346087

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360. https://doi.org/10.1198/016214501753382273

Feng, X.-N., Wu, H.-T., & Song, X.-Y. (2017). Bayesian regularized multivariate generalized latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 341–358. https://doi.org/10.1080/10705511.2016.1257353

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with mantel-haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*, 278–295. https://doi.org/10.1177/0146621605275728

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013, November 1). *Bayesian data analysis, third edition [Google-Books-ID: ZXL6AQAAQBAJ]*. CRC Press.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*, 835–845. https://doi.org/10.1093/biomet/asp047

Hans, C. (2010). Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing*, *20*, 221–229. https://doi.org/10.1007/s11222-009-9160-9

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction, second edition* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/978-0-387-84858-7

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research*, *15*, 1593–1623. https://www.jmlr.org/papers/v15/hoffman14a.html

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *71*, 499–522. https://doi.org/10.1111/bmsp.12130

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology*, *9*, 61–89. https://doi.org/10.1146/annurev-clinpsy-050212-185522

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, *33*, 730–773. Retrieved November 10, 2019 from https://www.jstor.org/stable/3448605

Jacobucci, R., & Grimm, K. J. (2018). Comparison of frequentist and bayesian regularization in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 639–649. https://doi.org/10.1080/10705511.2017.1410822

Kelava, A., Nagengast, B., & Brandt, H. (2014). A nonlinear structural equation mixture modeling approach for nonnormally distributed latent predictor variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 468–481. https://doi.org/10.1080/10705511.2014.915379

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*, 261–276. Retrieved November 28, 2018 https://www.jstor.org/stable/1435297

Kruschke, J. (2010, November 25). *Doing bayesian data analysis: A tutorial introduction with r*. Academic Press.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, *60*, 65–81. Retrieved February 14, 2020 from https://www.jstor.org/stable/25053023

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, *5*, 369–411. https://doi.org/10.1214/10-BA607

Leng, C., Tran, M.-N., & Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, *66*, 221–244. https://doi.org/10.1007/s10463-013-0429-6

Liang, X., & Jacobucci, R. (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 722–734. https://doi.org/10.1080/10705511.2019.1693273

Liu, Y., & Yang, J. S. (2018). Bootstrap-calibrated interval estimates for latent variable scores in item response theory. *Psychometrika*, *83*, 333–354. https://doi.org/10.1007/s11336-017-9582-9

Lu, Z.-H., Chow, S.-M., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research*, *51*, 519–539. https://doi.org/10.1080/00273171.2016.1168279

Lykou, A., & Ntzoufras, I. (2013). On bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing*, *23*, 361–390. https://doi.org/10.1007/s11222-012-9316-x

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*, 111–135. https://doi.org/10.3102/1076998614559747

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*, 1436–1462. Retrieved December 31, 2019 from https://www.jstor.org/stable/25463463

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*, 1023–1032. https://doi.org/10.1080/01621459.1988.10478694

Muthén, B., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335. https://doi.org/10.1037/a0026802

Narisetty, N. N., & He, X. (2014). Bayesian variable selection with shrinking and diffusing priors [Publisher: Institute of Mathematical Statistics]. *Annals of Statistics*, *42*, 789–817. https://doi.org/10.1214/14-AOS1207

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The bayesian lasso. *Psychological Methods*, *22*, 687–704. https://doi.org/10.1037/met0000112

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686. https://doi.org/10.1198/016214508000000337

Piironen, J., Paasiniemi, M., & Vehtari, A. (2020). Projective inference in high-dimensional problems: Prediction and feature selection [Publisher: The Institute of Mathematical Statistics and the Bernoulli Society]. *Electronic Journal of Statistics*, *14*, 2155–2197. https://doi.org/10.1214/20-EJS1711

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*, 179–191. https://doi.org/10.1080/01621459.1997.10473615 [Publisher: Taylor & Francis _eprint:.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*, 1151–1172. https://doi.org/10.1214/aos/1176346785

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement*, *24*, 97–118. https://doi.org/10.1111/j.1745-3984.1987.tb00267.x

Shi, D., Song, H., Liao, X., Terry, R., & Snyder, L. A. (2017). Bayesian SEM for specification search problems in testing factorial invariance. *Multivariate Behavioral Research*, *52*, 430–444. https://doi.org/10.1080/00273171.2017.1306432

Stan Development Team. (2019). *RStan: The R interface to Stan [R package version 2.18.2]*. http://mc-stan.org/

Stark, S., Chernyshenko, O. S., & Chernyshenko, O. S. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified approach. *Journal of Applied Psychology*, *91*, 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*, 402–415. https://doi.org/10.1037/1082-989X.11.4.402

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning (pp. 67–113)*. Lawrence Erlbaum Associates, Inc.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288. Retrieved November 11, 2019 from https://www.jstor.org/stable/2346178

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in rasch models. *Psychometrika*, *80*, 21–43. https://doi.org/10.1007/s11336-013-9377-6

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). *Rank-normalization, folding, and localization: An improved $ \widehat{r}$ for assessing convergence of MCMC. arXiv:1903.08008 [stat]*. International Society for Bayesian Analysis. Retrieved June 2, 2020, from http://arxiv.org/abs/1903.08008

Yuan, M., & Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, *100*, 1215–1225. https://doi.org/10.1198/016214505000000367

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429. https://doi.org/10.1198/016214506000000735