Published in final edited form as:

Multivariate Behav Res. 2017 ; 52(5): 593–615. doi:10.1080/00273171.2017.1342203.

Bayesian Estimation for Item Factor Analysis Models with Sparse Categorical Indicators

Sierra A. Bainter^{1,2}

¹University of Miami

²University of North Carolina at Chapel Hill

Abstract

Psychometric models for item level data are broadly useful in psychology. A recurring issue for estimating IFA models is low item endorsement (*item sparseness*), due to limited sample sizes or extreme items such as rare symptoms or behaviors. In this paper, I demonstrate that under conditions characterized by sparseness, currently available estimation methods, including maximum likelihood (ML), are likely to fail to converge or lead to extreme estimates and low empirical power. Bayesian estimation incorporating prior information is a promising alternative to maximum likelihood estimation for IFA models with item sparseness. In this article I use a simulation study to demonstrate that Bayesian estimation incorporating general prior information improves parameter estimate stability, overall variability in estimates, and power for IFA models with sparse, categorical indicators. Importantly, the priors proposed here can be generally applied to many research contexts in psychology, and they do not impact results compared to ML when indicators are not sparse. I then apply this method to examine the relationship between suicide ideation and insomnia in a sample of first year college students. This provides an important alternative for researchers who may need to model items with sparse endorsement.

Keywords

item factor analysis; bayesian estimation; sparse categorical indicators

Factor analysis applied to item-level data, or item factor analysis (IFA; Wirth & Edwards, 2007), is integral for measuring many constructs in social science research. Applied to categorical items, IFA is useful for studying many symptoms, behaviors, or beliefs from rating scales. Unfortunately, sparse endorsement for categorical items can easily arise in psychological research due to a combination of rare behaviors or symptoms, limited sample sizes, and complex modeling needs. It is well known that studies in psychology in general

Correspondence should be addressed to Sierra Bainter, Department of Psychology, University of Miami, Coral Gables, Florida, 33146-0751. sbainter@miami.edu.

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

are chronically underpowered (Maxwell, 2004), and for rare symptoms or behaviors such as substance use it becomes especially challenging to collect a sample large enough to reliably test a research question (Chassin, Presson, Lee, & Macy, 2013). For research involving IFA, the challenge is to collect enough data to reasonably ensure convergence, stable parameter estimates, and power to detect meaningful effects in a properly specified model, and a sample being “large enough” is dependent on many factors, including the choice of model estimator.

Common estimators for IFA models are full information maximum likelihood (ML) and a variety of limited information methods (e.g. estimators based on weighted least squares). Unfortunately, traditional methods for estimating IFA models are not appropriate in conditions characterized by sparseness (Forero & Maydeu-Olivarez, 2009; Moshagen & Musch, 2013; Wirth & Edwards, 2007). Bayesian estimation as an alternative approach for IFA estimation has been demonstrated extensively (Albert & Chib, 1993; Béguin & Glas, 2001; Edwards, 2010; Patz & Junker, 1999; Song & Lee, 2002; Lee & Song, 2012; Lee & Tang, 2006). This previous research has motivated the use of Bayesian estimation for advantages such as scaling to high-dimensional latent variable models (Edwards, 2010) or advantages for testing hypotheses related to fit (Béguin & Glas, 2001). Bayesian estimation has not yet been demonstrated for the problem of estimating models with sparse item endorsement. In particular, these past applications for IFA models in psychology have focused on using relatively flat prior distributions in an effort to avoid including subjective prior information.

However, inclusion of prior information can be motivated simply to decrease variability in parameter estimates, which is the motivation for regularization procedures such as ridge and lasso regression (Tibshirani, 1996), which are used in a frequentist framework but also have a Bayesian interpretation (Park & Casella, 2008). Regularization procedures have recently been extended to latent variable models with the goals of creating simpler models and minimizing over fitting (Jacobucci, Grimm, & McArdle, 2016; Yuan, Wu, & Bentler, 2011). In this article, I demonstrate that Bayesian estimation with reasonable prior information improves parameter estimate stability, overall variability in estimates, and power for IFA models with sparse, categorical indicators. Importantly, the priors proposed here can be generally applied to many research contexts in psychology, and they do not impact results relative to maximum likelihood when indicators are not sparse. This provides an important alternative for researchers who may need to model items with sparse endorsement.

In the following sections I present the IFA model, discuss traditional methods of estimation, and present Bayesian estimation with general priors for items with sparse endorsement. I then demonstrate the advantages of Bayesian estimation with moderately informative prior distributions in conditions where ML fails using a focused simulation study and an application to suicide data. I conclude with a discussion including recommendations for applied researchers and opportunities for future research.

Item Factor Analysis

A univariate IFA model for responses to item i for respondent j (y_{ij}) consists of three components: (1) a linear predictor (μ_{ij}), (2) a conditional response distribution, and (3) a link function g to relate the linear predictor to the probability of response. For binary y_{ij} (e.g. yes/no or true/false item responses) a Bernoulli response distribution is specified for each item (i.e. $\sim \text{Ber}$), and the linear predictor is defined as an item intercept V_i plus a factor loading λ_i expressing the regression of item i on the continuous latent factor η_j

$$\begin{aligned} y_{ij} | \eta_j &\sim \text{Ber}(\mu_{ij}). \quad 1 \\ \mu_{ij} &= V_i + \lambda_i \eta_j \\ \eta_j &\sim N(\alpha, \psi). \end{aligned}$$

Note that uncertainty in the response is modeled only through the response distribution (rather than through residual variances. The indicators are assumed independent, conditioned on the latent factor. Because the factor scores η_j are unobserved they are modeled as normally distributed, i.e. $\sim N(\alpha, \psi)$. The model is usually scaled either by fixing one item loading per factor to 1 and its intercept to 0, or by setting the mean and variance of the latent factor to 0 and 1, respectively. The logit (inverse logistic) link function is commonly used to map the range of the linear predictor ($-\infty$ to ∞) onto the permissible range for the conditional response distribution, which can only take on the values $[0,1]$, defined as

$$g(\mu_{ij}) = \ln[\mu_{ij}(1 - \mu_{ij})^{-1}]. \quad 2$$

In this case, choosing the logit link for the IFA yields

$$v_i + \lambda_i \eta_j = \ln[\mu_{ij}(1 - \mu_{ij})^{-1}]. \quad 3$$

Equivalently this model is expressed as

$$\mu_{ij} = \frac{1}{1 + \exp[-(v_i + \lambda_i \eta_j)]} \quad 4$$

which is an alternative expression of the well-known 2PL IRT model (Takane & de Leeuw, 1987). Parameters in the IRT and IFA models can be directly transformed from one parameterization to the other; only their interpretations differ (see Wirth & Edwards, 2007 for conversion formulas). In the next section I discuss estimation approaches for IFA and the challenge of modeling items with sparse endorsement.

Estimation Approaches for IFA Models

There are two families of traditional estimation approaches for IFA models that include categorical indicators. These are limited-information estimators (e.g., modified weighted least squares methods, Jöreskog & Sörbom, 2001; Muthén, du Toit, & Spisic, 1997; the polychoric instrumental variable estimator, Bollen & Maydeu-Olivarez, 2007) and full-information maximum likelihood (ML) estimation. Limited-information estimators are computationally faster than ML for high-dimensional models, have well-established tests for model fit (Wirth & Edwards, 2007; Forero & Maydeu-Olivares, 2009), and are therefore in widespread use (e.g. the default estimator in Mplus, WLSMV, is limited-information, see Muthén & Muthén, 2015). While more computationally challenging, ML estimation is statistically preferable to limited-information approaches for the problem of sparse endorsement because limited-information estimators are sensitive to bivariate sparse frequencies¹ whereas ML estimation is sensitive to univariate (item-level) sparse frequencies (Wirth & Edwards, 2007). Previous research using simulation studies has shown that ML performs better than limited information approaches in conditions characterized by sparseness; though both estimators perform poorly in these conditions (Forero & Maydeu-Olivares, 2009). For this reason, I limit my focus in this article to compare ML as a traditional estimator to Bayesian estimation.

Maximum Likelihood Estimation

ML estimation of IFA models relies on numerical integration using quadrature techniques (e.g. Gauss-Hermite integration), and because quadrature-based integration is needed for each latent dimension, the computational burden increases exponentially with the number of factors (Wirth & Edwards, 2007). However, some promising developments reduce the computational complexity of ML for large models (see Wirth & Edwards, 2007; Cai, 2010a; 2010b). Regardless of approach to ML estimation, many factors including number of items, item categories, latent factors, and sample size, magnitudes of factor loadings, and additional sources of model complexity (e.g. cross-loadings) are important for expected convergence to a proper solution (i.e., solution propriety) and stability of parameter estimates (Forero & Maydeu-Olivares, 2009; Moshagen & Musch, 2014; Wirth & Edwards, 2007). Besides this, sparseness – as an issue distinct from sample size – is a concern for models with categorical indicators. High intercepts, low factor loadings, or a combination of both can lead to low probabilities of endorsement and therefore sparse observed frequencies in finite samples.²

Previous simulation studies suggest that ML estimation performance is poor in conditions characterized by sparseness – namely smaller sample sizes combined with smaller probabilities for item endorsement (Forero & Maydeu-Olivares, 2009; Moshagen & Musch, 2014). Forero and Maydeu-Olivares (2009) found that ML estimation failed in small samples (200 observations) for binary items with low endorsement (10%), especially with fewer items per factor and low factor loadings. Moshagen and Musch (2014) found that ML

¹Specifically, the polychoric correlation coefficients that are used in limited-information estimation approaches are sensitive to low frequencies in bivariate contingency tables (Olsson, 1979; Savalei, 2011).

²The reverse is also true, for example very low intercepts could lead to an item that is almost always endorsed and non-endorsement is sparse.

estimation of IFA models in smaller samples could yield highly distorted parameter estimates and standard errors in smaller samples, even when ML estimation converges. Wirth and Edwards (2007) caution against applying IFA models to items with sparse endorsement. As a result of their comprehensive study comparing ML estimation to limited-information estimators, Forero and Maydeu-Olivarez (2009) suggested that “Future research should investigate if new estimators are able to yield adequate results in the conditions identified in this study for which both FIML [ML] and CIFA-ULS [limited-information] fail” (p. 294), and recommended Bayesian estimation as an alternative to investigate.

Researchers currently facing these estimation challenges must simplify their models, combine items, collapse item categories (if more than two categories), or drop items. In many domains of psychology it is not always an option to avoid sparse items when the pool of items is limited, sample size is limited, or items are particularly important to measure a construct. For example, if the intended measure is a tendency towards self-harm, a rare behavior, it is important to include items about self-harm behaviors, even if they have low base rates. Next I introduce Bayesian estimation as an alternative when a research question requires modeling items with sparse endorsement and ML estimation is not appropriate.

Bayesian Estimation

A Bayesian estimation³ approach requires selection of an appropriate prior distribution for each parameter in the model. The prior distribution $\pi(\theta)$ is combined with the model likelihood function $L(y, \theta)$ — the same likelihood maximized by ML estimation— to arrive at the posterior distribution $\pi(\theta|y)$ via Bayes’ theorem:

$$\pi(\theta|y) = \frac{\pi(\theta)L(y; \theta)}{\int \pi(\theta)L(y; \theta)d\theta} \propto \pi(\theta)L(y; \theta). \quad 6$$

From the posterior distribution, detailed information is available about the distributions of individual parameters, and standard errors or credible intervals (the Bayesian analogue to confidence intervals) are based on percentiles of the posterior.

Important components of a Bayesian analysis are: prior specification, model specification, posterior computation, and evaluating the posterior solution. The model specification does not differ in a Bayesian analysis, so I focus on the other three components in the next three sections. I refer interested readers to Gelman et al. (2013) for further details on all aspects of Bayesian inference.

Prior Specification—Priors may vary in distributional form and shape, and the parameters (scale, location, etc.) governing the prior distributions of parameters are called

³The Bayesian framework I focus on is not the only possible approach. Maximum a posteriori (MAP or modal Bayes) estimation pairs prior distributions from Bayesian statistics with a method of estimation similar to ML estimation (Mislevy, 1986). I focus on “full” Bayesian inference and MCMC to describe the posterior distribution for its generality and potential to scale to higher dimensional problems.

hyperparameters. Priors can be diffuse or have relatively more mass near a range of plausible values, and the level of diffusion in the prior is usually expressed by the hyperparameter values.⁴ Prior distributions and their hyperparameters can be based on prior knowledge (e.g. previously published studies, pilot data, knowledge from experts), general recommendations, or from the data (data-dependent priors). Including prior information can improve an analysis by building on existing knowledge and is a way to be transparent about prior beliefs, incorporating hypotheses into the analysis. It is fairly common to at least restrict parameter values to their admissible range, for example constraining variances to be positive (Gelman et al., 2013).

Flat priors can also be used to obtain results consistent with maximum likelihood estimation, using Bayesian estimation methods simply as a computational tool (Gelman et al., 2013).⁵ With adequate sample size, Bayesian and ML estimation converge on the same solution; this means that Bayesian estimation can be expected to perform as well as ML estimation when ML is converging to a stable solution (See Gelman et al., 2013, Ch. 4; Wasserman, 2005). For properly specified models⁶, even informative, inaccurate priors will be overwhelmed with adequate data, as long as there is non-zero probability at the true values (Depaoli, 2014). With limited sample sizes, parameter estimates are more sensitive to prior values (Berger & Bernardo, 1992; Kass & Wasserman, 1996). There are also hazards to relying on default priors of any kind, including default flat priors (Kass & Wasserman, 1996).

Whereas in ML estimation, asymptotic properties may not hold in finite samples, in a Bayesian analysis, inferences may be influenced by choices made about the prior distribution. Even when reasonable information is available to guide prior specification, a sensitivity analysis should be conducted to see whether the results are robust to prior specification (e.g. Song & Lee, 2012, Ch. 3) and will often be requested by reviewers if a Bayesian analysis is used. This can be done for example by perturbing the prior hyperparameter values or by considering other prior choices. Findings that are robust to sensitivity analyses strengthen confidence in your conclusions. If sensitivity analyses reveal that important inferences are dependent on a specific prior choice, this must be transparently reported to qualify your results. As with sensitivity analyses more generally (i.e. assumptions about missing data), choosing among competing models is a matter of adopting the most reasonable set of assumptions (reasonable to the analyst and also to potential reviewers). If a sensitivity analysis reveals problems, the analyst may need to conclude that the available data cannot effectively answer some questions. I discuss this important issue of prior sensitivity in more detail later in the context of an applied example (see also Gelman et al, 2013, Chs. 6 & 7).

⁴The labeling of informative versus uninformative for peaked and diffuse priors, respectively, is in wide use but can be misleading as a flat prior may be highly informative for some purposes, and the level of information in a particular prior varies case-by-case (see Zhu & Lu, 2004). I avoid referring to flat prior distributions as “uninformative” for this reason.

⁵Many flat priors do not have “proper” probability distributions, meaning they do not integrate to 1. For example a uniform distribution on the real line ($U(-\infty, \infty)$) is improper. The use of improper priors can lead to an improper posterior distribution, invalidating inference, therefore using improper priors requires care to ensure that the posterior distribution is proper.

⁶The influence of informative prior distributions, accurate or inaccurate, for misspecified IFA models is an important topic for future research.

An example prior specification for a univariate IFA model with binary indicators is as follows, selecting unbounded uniform priors for λ_i and v_i :

$$\begin{aligned}\pi(\lambda_i) &\sim U(-\infty, \infty) \\ \pi(v_i) &\sim U(-\infty, \infty)\end{aligned}$$

where the model is scaled by setting the mean and variance of the latent factor to (0,1). However there is a reasonable basis to restrict these priors. General ranges and typical values of these parameters are known in social science research. If intercepts v_i are expected to range from about negative 5 to 5, a reasonable prior could be normal, centered at zero, with standard deviation hyperparameter $\sigma = 2.5$ [i.e. $\pi(\lambda_i) \sim N(0, 2.5)$] to concentrate the prior plausible range. If theory would strongly dictate that all items should be positively related to the latent variable (e.g., from previous measurement studies), the prior distribution could favor or be restricted to positive values.

Posterior Simulation—Bayesian estimation of most interesting models relies on Markov chain Monte Carlo (MCMC) simulation methods to describe the posterior distribution (Tanner & Wong, 1987; Gelfand & Smith, 1990). MCMC algorithms draw a chain of correlated samples that asymptotically converge to the posterior distribution (see Edwards, 2010, for an overview). Most existing work for Bayesian IFA has focused on two types of MCMC algorithms: Gibbs and Metropolis-Hastings (Albert & Chib, 1993; Béguin & Glas, 2001; Edwards, 2010; Patz & Junker, 1999; Song & Lee, 2002, 2012; Lee & Tang, 2006). However the rules controlling implementation require careful oversight and fine-tuning in order to effectively sample from the whole posterior distribution, and convergence for some high-dimensional or correlated posterior distributions can be effectively impossible (Gelman et al., 2013). An alternative to Gibbs and MH algorithms designed to explore the posterior distribution more systematically and efficiently is Hamiltonian Monte Carlo (HMC, sometimes called Hybrid Monte Carlo; Duane, Kennedy, Pendleton, & Roweth, 1987)⁷. HMC is also flexible with respect to prior choice, and is implemented in the general and flexible Bayesian program and language Stan (Stan Development Team, 2015) with the No-U-Turn sampler (Hoffman & Gelman, 2014) to automate tuning of the algorithm.

Posterior Evaluation—There are many techniques to help assess MCMC convergence (see Gelman et al., 2013, for a review), however it is generally impossible to know for sure that any single chain has converged because convergence diagnostics assess conditions that are necessary but not sufficient for convergence. A useful visual diagnostic tool is a traceplot, which shows the iteration number plotted against the sampled values for a parameter. The potential scale reduction statistic ($\hat{R}$; Gelman and Rubin, 1992) compares variability within a chain to variability between other randomly initiated chains; a value of $\hat{R} = 1$ is evidence of convergence. Importantly, all parameters in a model must show evidence of convergence before it is suitable to make inferences from the posterior

⁷because many concepts of Hamiltonian dynamics and HMC are unfamiliar to non-physicists, a detailed description of HMC is beyond the scope of this paper. I refer interested readers to Neal (2011) and Gelman et al. (2013, pp. 300-308) for more details, however note that this material is necessarily technical.

distribution. Because the samples from the posterior are not independent, an estimate of effective sample size (n_{eff}) is useful to measure efficiency of the chain and determine whether sufficient uncorrelated samples have been drawn for posterior inference (see Gelman, et al., 2013, pp. 284-87).

Advantages of Bayesian Estimation for Sparse Indicators

Though Bayesian estimation has been profitably used to estimate complex IFA models (e.g. Edwards, 2010; Song & Lee, 2012), it has not been studied for the problem of estimating IFA models with sparse categorical indicators, and past demonstrations have focused on flat prior distributions. However, incorporating prior information has been shown to be especially useful in sparse data settings (Dunson & Dinse, 2001; Peddada, Dinse, & Kissling, 2007). Dunson and Dinse (2001) suggest a Bayesian method for studying tumor incidence rates, which are rare events and often difficult to predict because of small sample sizes. By incorporating historical data as prior information, their method leads to more interpretable results and can improve detection of small but biologically important changes in incidence rates (Dunson & Dinse, 2001; Peddada, Dinse, & Kissling, 2007).

Priors can have a stabilizing, shrinkage effect on parameters with limited data available for their estimation. In the case of estimates nearing extreme values due to sparse data, shrinking extreme values should be computationally advantageous, more reliable, and lead to higher power. Often applied researchers prefer the unbiasedness property of maximum likelihood estimation, but in cases of sparseness, it is arguably advantageous to prefer estimation with some bias in exchange for lower variance to avoid overfitting. This rationale (i.e., increased stability at the cost of some bias) is the same used for regularized regression methods such as ridge regression or lasso regression (Tibshirani, 1996), which are used in a frequentist framework but also have a Bayesian interpretation (Park & Casella, 2008). Regularization procedures have been extended to structural equation models with the goal of creating simpler models and minimizing over fitting (Jacobucci, Grimm, & McArdle, 2016; Yuan, Wu, & Bentler, 2011). The stabilizing effect of reasonable priors should also be beneficial for computational problems arising from sparse categorical data because the priors can be used to avoid improper solutions and extreme estimates. This prior influence may be undesirable in some circumstances depending on the purposes for specific model inferences, however in general if reasonable priors are chosen prior-driven stabilization should be advantageous.

In summary, Bayesian inference can be adapted to provide good performance even in less than ideal circumstances with large models, small samples, or sparseness; however it is important to evaluate posterior simulation carefully. As such, Bayesian estimation is a promising alternative to ML estimation for IFA with sparse indicators. In the next section I demonstrate Bayesian estimation with moderately informative priors for IFA models with sparse indicators in a targeted simulation study and compare performance to ML estimation.

Simulation Study

In this simulation, I compare ML and Bayesian estimation for IFA models across a set of targeted conditions, varying the pattern of sparse items, degree of sparseness, and sample

size. There are several aims for this study. First, the equivalence of ML and Bayesian estimation is demonstrated in baseline conditions with no sparse items, using flat priors and a range of moderately informative prior specifications. The stability of results across prior specifications demonstrates that the prior information is not influential when adequate information is available in the data. In the sparseness conditions, I demonstrate the effects of sparseness for ML estimation compared to Bayesian estimation including priors in terms of convergence, efficiency, and power. Finally, I use this simulation to illustrate concrete details of Bayesian estimation.

Data for this simulation were generated in R (R Core Team, 2015) from a distribution with fixed population values consistent with a two-factor IFA with five binary indicators per factor. The correlation between factors was moderate, $\psi_{12} = .3$. All item loadings were $\lambda_i = 1.5$, (logit-scale). This item loading is consistent with estimates encountered in practice (e.g. Hussong, Flora, Curran, Chassin, & Zucker, 2008) and simulation studies for similar models (e.g. Cai, 2010; Edwards, 2010; Forero & Maydeu-Olivares, 2009).

The sample size, pattern of sparse items, and probability of endorsement for sparse items were varied. Sample size was varied at $N=250$ and $N=500$ for each of 500 replications⁸ per condition. These sample sizes were chosen to be representative of modestly large samples with which substantive researchers would typically feel confident estimating and interpreting structural equation models. The pattern of sparse items was varied across three conditions: one baseline condition with no sparse items (all $v_i = 0$), a condition with four of five items sparse on only one factor, and a condition with four of five items sparse on both factors. To induce sparseness, threshold parameters were set to $v_i = 4.11$ and $v_i = 4.9$ (logit-scale); this corresponds to marginal endorsement probabilities $p = .04$ and $p = .02$ with $\lambda_i = 1.5$. Expected marginal frequencies for sparse items were 10 ($v_i = 4.11$, $p = .04$) and 5 ($v_i = 4.9$, $p = .02$) for the sample size of 250, and 20 ($v_i = 4.11$, $p = .04$) and 10 ($v_i = 4.9$, $p = .02$) for the sample size of 500. Endorsement was non-zero for all items in all replications.

ML Estimation

Maximum likelihood estimation was performed in Mplus version 7.31 (Muthén & Muthén, 2015) using the default integration method (adaptive numerical integration) with 15 integration points. This integration method and number of integration points is well-suited for a IFA with two latent factors, though alternative methods of integration are useful for more complex models with more latent factors (Wirth & Edwards, 2007). Estimation for each replication was automated using the MplusAutomation R package (Hallquist & Wiley, 2014). The latent factors were identified by setting the variance to unity for each factor and estimating all factor loadings⁹. The program syntax is provided in Appendix A.

⁸In this study, 500 replications per condition were sufficient for Monte Carlo convergence. This was checked by examining running average plots for key estimates across replications.

⁹This model specification is only locally identified (Bollen & Bauldry, 2010; Loken, 2005), as there is a sign indeterminacy for the factor loadings on one or both factors. For the estimation routines used in Mplus for these models and data, the sign indeterminacy is not an issue and leads to solutions with a majority of positive factor loadings.

Bayesian Estimation

Bayesian estimation was performed with Stan version 2.10.0 implemented in R, using Hamiltonian Monte Carlo and the No-U-turns sampler. Stan software can be used with many interfaces, including R, but is coded in C++ for efficiency. To write a Stan program, users define the statistical model and priors for each parameter, and the program adapts the sampling algorithm. Using HMC in Stan, there is no restriction for prior distributions. Stan allows users to specify improper priors (i.e. integral of prior is infinity) and diagnoses improper posteriors automatically when parameters overflow to infinity during simulation (Carpenter et al., 2017). In contrast to other statistical programs that offer Bayesian estimation, using the Stan programming language allows users a high degree of flexibility in model and prior choice, oversight of MCMC convergence, and fast computation. The Stan program syntax used to specify the IFA model is provided in Appendix B.

Prior specification—The first prior specification included flat priors for the intercepts and item loadings. Flat priors were not expected to perform well in conditions with sparseness. These priors were normal with extremely high variance, essentially uniform on the admissible range for all parameters:

$$\begin{aligned}\pi(\lambda_i) &\sim N(0, \sigma = 1000) \quad 8 \\ \pi(v_i) &\sim N(0, \sigma = 1000) \\ \pi(\psi_{21}) &\sim U[-1, 1].\end{aligned}$$

The second prior specification used priors with increased probability for plausible values and constrained factor loadings as positive:

$$\begin{aligned}\pi(\lambda_i) &\sim N(0, \sigma = 3.57), \quad \lambda_i \geq 0 \quad 9 \\ \pi(v_i) &\sim N(0, \sigma = 3.57) \\ \pi(\psi_{21}) &\sim U[-1, 1].\end{aligned}$$

The variance in the moderately informative priors specifies 95% prior probability that item intercepts lie within $[-7, 7]$, and 97.5% prior probability that factor loadings lie within $[0, 7]$. Note that these ranges are general for many applications in psychology. Specifying factor loadings as positive as I do here may or may not be appropriate depending on the research context; however there is an additional computational advantage to this restriction which I explain in the next section. Note also that these priors, though concentrated, place the highest probability near zero. The information contained in these priors therefore encourages “shrinkage” of parameter estimates rather than, for example, positing that parameter estimates be large.

Finally, to examine sensitivity to the variance of the prior distribution, I included prior conditions with somewhat more diffuse priors (prior $\sigma = 5$, expressing prior probability that 95% of intercepts lie within $[-10, 10]$, factor loadings lie within $[0, 10]$), and a prior condition with more concentrated priors (prior $\sigma = 2$, specifying prior probability that 95% of intercepts lie within $[-4, 4]$, factor loadings lie within $[0, 4]$)).

Posterior simulation—The simulations were run on computing clusters. For each condition and prior, replications were submitted in parallel in sets of 20. The method of identification used for ML (setting each factor mean and variance to 0 and 1, respectively), although only locally identifying the model (Bollen & Bauldry, 2010) lead to all solutions with a majority of positive factor loadings (i.e. sign indeterminacy was not an issue using ML estimation for this model and data in Mplus). However, it is important to note that sign indeterminacy does become an issue in posterior simulation. Specifically, solutions with either all positive factor loadings or “flipped” solutions with all negative factor loadings are log-likelihood equivalent. Similarly, a solution with all positive loadings for one factor and all negative loadings for the other factor, and a negative covariance between factors, is equivalent. This sign indeterminacy can be resolved either by restricting the sign of factor loadings in the prior distribution (as in the moderately informative prior specification, Equation 9) or using the alternate scaling: by setting a single indicator to 1 for each factor and estimating the variance of each factor. However choice of scaling can also impact the efficiency of posterior simulation. For multidimensional latent variable models, estimation may be more challenging when the scale of the latent variable is freely estimated than when it is set to 1.¹⁰

In order to maximize efficiency in posterior simulation, the more efficient scaling was used (setting latent factor variances to 1), and flipped solutions in the baseline condition with flat priors were post-processed after estimation to the preferred scaling for inference (constraining factor loadings to be positive in the moderately informative priors also removed sign indeterminacy). Post-processing to an inferential parameterization has been used in a similar modeling context with continuous indicators (Ghosh & Dunson, 2009). To avoid opposite solutions within replication, a single chain with 20,000 iterations (half burn-in) was run for each replication, and $\hat{R}$ was monitored for each chain to determine if any chains switched between solutions (i.e., $\hat{R}$ above 1 should signal switching between solutions within a chain).

Performance of each estimator was evaluated in terms of convergence, bias, efficiency, confidence interval coverage, and empirical power.

Results

Results are presented first for ML followed by Bayesian estimation, organized by outcome.

Maximum Likelihood

Tables 1 and 2 summarize results for ML estimation. To simplify the presentation, results are grouped for item loadings and intercepts on items with 50/50 endorsement (λ , ν) and loadings and intercepts for sparse items (λ_{SP} , ν_{SP}).

¹⁰Specifically, for the baseline condition with no sparse items and moderately informative priors, scaling to a latent factor resulted in small estimated effective sample size (e.g., less than 100) for multiple parameters in approximately 10% of replications after 10,000 replications (half burn-in). Scaling by setting the factor variances to 1, however, resulted in higher estimated effective sample size (e.g., minimum 569) and sampling was twice as fast.

Convergence and extreme values—All Mplus warnings and errors were monitored to screen for serious errors indicating nonconvergence versus ignorable warnings (e.g., warning that an estimate has been fixed). Mplus may fix estimates if they reach boundaries (e.g., logit thresholds outside $[-15, 15]$) at certain points in the estimation routine, but estimates outside of this range may also be reported (Muthén & Muthén, 2015) and in this simulation estimates were not automatically fixed to boundary values. In addition to convergence to proper maximum likelihood solutions, solutions were monitored for extreme estimates that would seem suspicious in practice.

In the baseline conditions (i.e. no sparse items) there were no convergence failures. With 4/5 items sparse on one factor, all replications converged with $N=500$ and 4% endorsement, but otherwise convergence rates ranged between 97.8% (2% endorsement, $N=500$) and 94.2% (2% endorsement, $N=250$). In the most extreme conditions, with 2% endorsement on 8/10 items 91.2% of replications technically converged with $N=500$, and 81.8% converged for $N=250$. Parameters were occasionally fixed to their estimated value (i.e. no standard error is reported); this occurred in less than 2% of replications.

Although convergence criteria were technically satisfied for most ML solutions, extreme values were frequently reported in conditions with sparseness. While any single set of limits for extreme values is necessarily arbitrary, it is illustrative to consider how frequently extreme estimates arose. As one gauge of the number of extreme solutions, I considered the number of replications that converged with intercept estimates larger than ± 15 or item loadings larger than ± 8 . In practice, estimates this large would be considered extremely suspicious. In most conditions with sparse endorsement, the rate of converged replications with extreme values was high, especially with the smaller sample size or higher threshold. For example, 9% (44/489) of solutions included at least one estimate outside this range for the model with sparse items on one factor with $N=500$, 2% endorsement, and 55% (224/409) of solutions with sparse items on both factors with $N=250$, 2% endorsement. Extreme intercept and loading estimates often (but not necessarily) corresponded to observed endorsement of about 5 cases (1% for sample size of 500). For these ranges of extreme values, extreme intercepts and item loadings were almost equally common (49% extreme loadings, 51% extreme intercepts for $N=500$ conditions). High loadings frequently corresponded to high intercepts: about 75% of items with extreme intercept estimates also had extreme loading estimates. This means that items with low endorsement were generally estimated as strongly related to the latent variable.

All technically converged solutions are included with the results in Tables 1 and 2 and in sections describing bias, efficiency, coverage, and empirical power. Robust statistics (e.g. medians) are useful for considering performance without undue influence from extreme observations; however note that “extreme” estimates occurred frequently in these conditions.

Bias and efficiency—Raw bias for parameters using ML estimation is summarized in Tables 1 and 2, calculated generally for parameter θ by subtracting the true value from the r th estimate ($\hat{\theta}_r$) and averaging across the total number of replications in the cell (R):

$$\frac{\hat{\theta}_r - \theta}{R} \quad 10$$

Because the mean is sensitive to extreme values, median bias was also recorded with the minimum, 5th quantile, 95th quantile, and maximum values for parameters in each condition. Root mean square error (RMSE) is included as a measure of parameter estimate efficiency, computed generally for parameter θ as

$$\sqrt{\frac{\sum_{r=1}^R (\hat{\theta}_r - \theta)^2}{R}} \quad 11$$

RMSE reflects both sampling variability and squared bias, with larger values reflecting greater variability in estimates relative to the true value. When estimates are unbiased, the RMSE can be thought of as the empirical standard error. When bias is present, efficiency measured by RMSE reflects overall accuracy. RMSE is a scale-dependent criterion, so there are no cut-offs to determine acceptable or unacceptable values, however RMSE can be used to compare efficiency across estimators. Because RMSE is sensitive to extreme values, the median absolute deviation about the median (MAD) is also presented in Tables 1 and 2 as a robust measure of efficiency (Huber & Ronchetti, 2009), calculated for each parameter in replication r as

$$MAD_K = Median\{|\hat{\theta}_r - M_K|\} \quad 12$$

where

$$M_K = Median\{\hat{\theta}_r\} \quad 13$$

There was no evidence of meaningful, systematic bias using ML estimation in any of the conditions studied, as seen in Tables 1 and 2. Some mean estimates are biased due to extreme values; however median estimates were very close to the true values. However, as expected, sparseness lead to decreased efficiency in parameter estimates compared to the baseline condition. For example, with N=500, RMSE for the estimated correlation between factors increased from .06 in the baseline condition to .11 with sparse items on one factor and .16 with sparse items on both factors. Average RMSE for loadings and intercepts on sparse items were exceedingly large due to extreme values; changes in efficiency for these parameters can be examined in terms of MAD.

Confidence interval coverage—As an indicator of bias in standard errors, the proportion of estimated 95% confidence intervals that contained the true population parameter are shown in Tables 1 and 2. If parameter estimates and standard errors are

unbiased, the 95% confidence interval should contain the true population value in 95% of replications. Collins, Schafer, and Kam (2001) consider coverage values that fall below 90% to be problematic. In the baseline conditions, 95% confidence interval coverage was between 92-96%. The range widened slightly with 4/5 items sparse on a single factor (88-96%). In the high sparseness conditions the lowest coverage was 88%, however coverage was slightly inflated to 98-99% for estimates of the threshold for non-sparse items. These results suggest that confidence intervals were not substantially biased in these conditions with sparse items; in other words, standard errors were generally appropriately large to account for variability in parameter estimates.

Empirical power—Empirical power was computed by recording the proportion of significant estimates for each parameter according to a standard alpha level of .05. Empirical power to detect significant effects (ψ_{12} , λ , λ_{SP} , ν_{SP}) was lower in conditions with sparseness. For example, with 4/5 items sparse with 4% endorsement, 90% of the correlation estimates between factors were significant at N=500 (62% at N=250). 35% of correlation estimates were significant with sparse items (2% endorsement) on both factors and a sample size of 250. The proportions of significant factor loadings also decreased in the conditions with sparseness, for example with N=250 and sparse items on both factors, 49% and 15% of the loadings for sparse items were significant, at 4% and 2% endorsement, respectively.

Given these results, it is clear that although ML estimation is generally unbiased, it is not appropriate in these conditions with very sparse item endorsement. This finding has previously been reported, especially in a comprehensive literature review and simulation study by Forero & Maydeu-Olivares (2009). If researchers wish to make inferences from data with sparse endorsement, they are likely to obtain suspicious parameter estimates and to lack power to detect significant effects. Next, I detail results for these models using Bayesian estimation.

Bayesian Estimation

Convergence assessment—Whereas ML has clear criteria for convergence, using Bayesian estimation there are only degrees of confidence in convergence. Convergence was assessed by monitoring the estimated potential scale reduction factor and effective sample size. Stan computes the potential scale reduction factor by splitting the chain in half (Stan Development Team, 2015), so it is possible to monitor $\hat{R}$ for a single chain. I also monitored diagnostic trace plots and autocorrelation plots for a small sample of replications. In general, most replications appeared to converge, except for the combination of flat priors and sparse items. With sparse items and flat priors, the replications failed to converge within a time limit of 7 days for sets of 20 replications.¹¹ Because there is no rationale to use Bayesian estimation with flat priors for sparse data, the flat priors are only reported here as a comparison for the baseline conditions.

For all conditions reported here, $\hat{R}$ was 1 for all parameters. The number of replications with effective sample size above 100 for all parameters (denoted R) are reported for each

¹¹As noted by a reviewer, the prior variance used here may simply be too large for models with categorical data. Convergence may be achievable in a condition with “flat” priors over a smaller region.

condition in Tables 3–5. For the baseline conditions with no sparse items, effective sample size was above 100 for all parameters in 99.6–100% of replications using flat priors, and in 100% of replications using moderately informative priors. For conditions with sparseness, 10,000 post burn-in iterations was sufficient to achieve 100 effective samples per parameter for most replications using moderately informative priors. Because convergence is very different in the Bayesian and ML frameworks, it is problematic to directly compare “convergence rates” from the two frameworks. Even though effective sample size was lower than the specified cutoff for 34 and 44 replications in the most extreme sparseness conditions at $N=500$ and 250, respectively, sampling for more iterations could be done to achieve the desired effective sample size. In these conditions with a high number of sparse items, using a moderately informative prior specification, it is possible to examine solutions in cases where an estimate was not available using ML estimation, either by sampling for more iterations or by inspecting solutions with lower effective sample size.¹²

For this simulation, effective sample size of at least 100 for all parameters was considered sufficient to interpret results for each replication. Replications with effective sample size below 100 for any parameter are not included in results tables. In practice, higher effective sample size may be preferable for any single replication for interpreting some quantities (e.g. 1000 for increased precision for interpreting posterior intervals; see Gelman et al., 2013, p. 267), and lower effective sample size may be adequate for other quantities such as posterior medians. For a single replication, it may be useful to sample for more iterations. However it is not currently possible to automate sampling until a desired effective sample size is reached using the HMC algorithm in Stan.

Posterior simulation (20,000 total iterations) for sets of 20 replications completed in approximately 10 hours or less running on a single Intel Xeon Processor (2.93 GHz). This means that estimation for single replications could be expected to run in about 30 minutes on a personal computer, for this model and sample size.

Bias and efficiency—Bias and efficiency for Bayesian estimation was evaluated based on posterior medians. Table 3 summarizes results for the baseline models with no sparse items, organized by prior specification. Results for conditions with sparse items are in Tables 4 and 5. With no sparse items, there was no evidence of bias in any parameter using any of the three priors studied.

Correlation estimates were downwardly biased when the models included sparse items, and this bias was more pronounced with more sparse items. For example, the mean correlation estimate was .27 and .26 (raw bias $-.03$ and $-.04$) with $N=250$ and 2% endorsement for 4/10 and 8/10, respectively. Bias in item loadings depended on the number of sparse items in the model and whether the item loading was for a sparse item. In general, loadings for sparse items were downwardly biased and loadings for non-sparse items were biased upwards. This pattern provides relatively more weight to items that were not sparse. For example, with $N=500$ and 2% endorsement on 4/10 items, bias was $-.11$ and $+.26$ for loadings on sparse

¹²For the conditions with sparseness, I separately examined results for all replications, including replications with effective sample size below the cutoff. The results did not differ meaningfully for any outcome.

and non-sparse items respectively. The upward bias was more pronounced for non-sparse item loadings with 8/10 items sparse. However, extreme estimates were eliminated. Considering the ranges for what were considered extreme estimates using ML, there were no intercept estimates outside of ± 15 or item loadings outside of ± 8 .

With no sparse items, parameter estimate efficiency as measured by RMSE and MAD was essentially the same using each prior specification; the efficiency of parameter estimates also closely matched efficiency using ML estimation for these baseline conditions. As expected, in conditions with sparse items, RMSE and MAD were larger with more sparse items. In the high sparseness condition with $N=250$, RMSE for item loadings was 0.60 (sparse item) and 1.87 (non-sparse item) for moderately informative priors. Figure 1 and Figure 2 compare MAD and RMSE for each parameter across estimators, using as examples both sparseness conditions with $N=500$ and 2% endoresement. For item loadings on non-sparse items in the high sparseness conditions, MAD was markedly higher using Bayesian estimation. For all other parameters, RMSE and MAD were higher using ML estimation or about equal. Note that different subsets of replications are included in this comparison, because the ML results are restricted to models that converged and Bayesian results are restricted to replications that met the minimum effective sample size for all parameters. Replications that did not converge using ML estimation were not the same replications with below intercept effective sample size using Bayesian estimation in any case.

Credible interval coverage—For the baseline conditions, credible interval coverage was between 94-96% for all parameters and all priors; this aligns with the coverage observed using ML. With moderately informative priors, coverage rates were comparable to those observed for ML estimation for the same conditions.

Empirical power—Empirical power for different estimates is summarized in the last columns of Tables 3-5. In the baseline condition, empirical power was 1.00 for true effects (correlation estimates and factor loadings) using both prior specifications. In all sparseness conditions with moderately informative priors, power to detect true effects was 1.00 for all parameters, matching empirical power in the baseline conditions.

Prior sensitivity analysis—To examine possible sensitivity to the chosen moderately concentrated prior distribution ($\sigma = 3.57$), all conditions were re-estimated using a less concentrated ($\sigma = 5$) and more concentrated ($\sigma = 2$) prior specification. All important quantities including average parameter estimates, RMSE and MAD did not appreciably differ under each prior specification. Due to space constraints, these tables are not included here, but are available by request.

Summary of Bayesian Estimation Results—Taken together, the results showed that the use of priors in Bayesian estimation can stabilize estimates in IFA models with sparse, categorical data. The use of a moderately informative prior specification eliminated extreme parameter estimates, improved estimate efficiency, and increased empirical power to detect true effects. Results also suggest that Bayesian estimation can be a useful alternative when models do not converge using ML estimation, although more iterations of posterior sampling may be needed to ensure an adequate number of effective samples. However, increased

overall efficiency and empirical power were tied to a trade-off with overall unbiasedness. Bayesian estimation performs similarly to ML estimation with flat or moderately informative priors in a baseline condition with a moderate sample size and even endorsement on all items.

Application to Suicidality and Insomnia

To illustrate the advantages of Bayesian Estimation for sparse categorical indicators, I consider an application to study the relationship between insomnia and suicidality. Data for this example is taken from a study of 457 first year college students (64% female, $M_{\text{age}}=19$ years, range: 17-33 years) who received research credit for participation (see Timpano et al., 2011). Participants completed questionnaires assessing several psychological risk factors, including the Beck Scale for Suicide Ideation (BSS; Beck et al., 1979) and the Insomnia Severity Index (ISI; Morin, 1993). Insomnia has been linked to risk for suicide in several studies (e.g., McCall & Black, 2013; Winsper & Tang, 2014). To assess the relationship between insomnia symptoms and suicidality, I estimated an item factor analysis model with two factors and their correlation for the ISI (7 items, dichotomized from a 0-4 scale) and BSS items (5 screening items, dichotomized from a 0-2 scale). Before constructing models and choosing an estimator, I recommend examining item-level frequencies for sparseness. In this example, endorsement for the 7 ISI items ranged from 15% -57%, however endorsement for the BSS items ranged from just 2% - 6% (or 9-27 cases).

The model was estimated for the full sample using both ML estimation in Mplus and Bayesian estimation in Stan with the moderate prior specification used in the simulation study (prior variance $N(0,3.57)$ to specify high probability that item intercepts and loadings were less than 7 in magnitude, and truncated at zero to specify positive item loadings). As in the simulation study, no cross-loadings were estimated, and the latent factor was scaled by setting the mean and variance of the latent factor to zero and unity and estimating all item loadings. To further demonstrate the comparison between ML and Bayesian estimation in limited samples with sparse endorsement, I also randomly split the full sample, and compare results using both estimators for each subsample ($N=234$ & $N=223$).

Using ML estimation in the full sample, the model converged to a solution without errors, but some of the parameter estimates appeared unstable. Estimates and confidence intervals using ML estimation in the full sample and in the split halves are plotted in Figure 3. This figure shows that some item parameters were estimated with narrow confidence intervals and little fluctuation across subsamples. However for a substantial number of parameters there was a high level of uncertainty in estimates, with very wide confidence intervals making it difficult to claim statistical significance.

Bayesian estimation was carried out as in the simulation study with 20,000 iterations for a single chain (half discarded as burn-in). After sampling, $\hat{R}$ was 1 and the effective sample size was above 1000 for all parameters (median ESS = 5149 out of a possible maximum of 10,000). As shown in Figure 4, results from Bayesian estimation were less variable compared to ML. Where ML estimates were measured with narrow confidence intervals, the Bayesian estimates look nearly identical. However for item parameters estimated with high

uncertainty using ML, the Bayes median estimates are somewhat shrunken and the intervals are much narrower (and therefore exclude zero more often). For example, using ML estimation in the full sample ($N=457$) the factor loading for being worried/distressed about a current sleep problem on the ISI was 9.49 ($SE=5.45$), and the factor loading for endorsing a weak or moderate wish to die on the BSS was 9.32 ($SE=13.03$). The thresholds for these items were also extreme, estimated as 8.52 ($SE=4.83$) and 16.51 ($SE=24.03$) for the ISI and BSS items, respectively. Neither the factor loading nor threshold for either item was significant. Using Bayesian estimation the factor loadings were 6.16 ($SD=1.34$) and 4.33 ($SD=0.99$) for the BSS and ISI items respectively, with corresponding thresholds of 5.51 ($SD=1.13$) and 7.81 ($SD=1.51$), and the smaller posterior standard deviations is practically meaningful (all estimates are now significant).

Finally, I assessed prior sensitivity by comparing results across the range of priors used in the simulation study, considering a relatively more concentrated prior $[N(0,2)]$ and one relatively less concentrated $[N(0,5)]$. As displayed in Figure 5 for the full sample, differences in posterior estimates were negligible across prior specifications. Posterior intervals are indistinguishable for the majority of item parameters. For parameters estimated with more uncertainty, the prior variance had more influence on shrinkage, but in no case did this result in meaningfully different estimates (e.g., change in significance). To see if the results were more sensitive to this prior specification in a smaller sample, Figure 6 shows results in the smaller subsample ($N=223$). Besides slight differences due to reduced sample size, it is difficult to detect any increased or unreasonable influence of the prior at this smaller sample size.

In this example the sensitivity analysis did not reveal that results were highly sensitive to prior inputs, but this may not always be the case. An important question is what an analyst should do (and what to conclude) if the results do meaningfully change. Statistical inference is ultimately about quantifying uncertainty, and inconclusive results from a sensitivity analysis provide insight into the degree of uncertainty we have in our results. There are two positions that can be argued when a sensitivity analysis reveals that results meaningfully differ: 1) both models are reasonable, therefore the results are inconclusive, the conclusion cannot be reliably drawn with the existing data or 2) the assumptions of one model are more reasonable, therefore we have more confidence in the results from the model with the assumptions we are most comfortable with. In one sense the results from this example *were* sensitive, sensitive to ML versus Bayesian estimation with moderately informative priors, but in this case the restricted prior range of values makes more sense (and led to more interpretable and stable estimates) than the unbounded range assumed using ML. A guiding principle is that a reader should be informed of any analysis decisions or assumptions that could compromise the conclusions made from the analyses and make his or her own judgment about the relative merit of each modeling choice.

Discussion

I have presented a method for improving IFA estimation with sparse, categorical indicators. Prior information about typical parameter values in psychological research can be utilized in a Bayesian framework to decrease variability in parameter estimates, eliminate extreme

estimates, and improve empirical power to detect true effects. I evaluated Bayesian estimation in conditions where ML performs poorly and in baseline conditions where ML performs well.

Performance of ML Estimation for Sparse Items

Previous research has suggested that categorical estimation methods break down under conditions of sparseness and fail to reliably produce converged, reasonable solutions (Forero & Maydeu-Olivares, 2009; Moshagen & Musch, 2014; Rhemtulla et al., 2012; Wirth & Edwards, 2007). Parameter estimate efficiency and power also decrease in conditions with sparseness.

This study extends these results specifically for conditions with a large proportion of items with low endorsement and moderate sample sizes. Sparseness has very little effect on bias in parameter estimates using ML estimation. In the conditions studied, ML convergence did not fall below 80%. However, as expected, sparseness led to suspiciously large parameter estimates in a substantial proportion of replications. Moshagen and Musch (2014) also reported suspicious ML estimates despite high convergence rates, and the present results support their finding that achieving convergence to proper ML solutions does not necessarily indicate that results are trustworthy. Besides decreased efficiency and the presence of extreme parameter estimates, empirical power to detect true effects unsurprisingly decreases in conditions with substantial sparseness.

Considering the broader literature on IFA models, the issue of very low endorsement for categorical items is analogous to continuous items with very low variance. Continuous items with low variance can cause estimation problems related to empirical under-identification (Bentler & Chou, 1987; Rindskopf, 1984). With item variances near zero, there is too little information available to perform estimation. While this research is not intended to identify exact frequencies or marginal probabilities where sparseness becomes an issue, the general principle is that sparse endorsement can lead to items with insufficient information to perform ML estimation. It is clear that frequencies play a more important role than endorsement rates; a 2% probability of endorsement with $N=250$ is more problematic than 2% probability of endorsement with $N=500$. The conditions included here are extreme in terms of marginal endorsement rates and numbers of sparse items. However, smaller sample size, lower item loadings, fewer items per factor, and increased model complexity would all be expected to worsen the performance of ML (Forero & Maydeu-Olivares, 2009; Gagné & Hancock, 2006; Marsh et al., 1998; Moshagen & Musch, 2014).

Empirical power was lower in models with sparse items on one or both factors. In these models item loadings, intercepts for sparse items, and the correlation between factors were substantial, true effects. Of these, item loadings and the correlation between factors are particularly meaningful in practice. Non-significant factor loadings for indicators of a latent construct would be very troubling in practice; these items would typically be removed (e.g., Kline, 1994). Decreased power was a result of the large increase of variability in the estimates and associated increase in standard errors for ML estimation of models with sparse indicators.

Comparing Bayesian Estimation to ML for Sparse Items

In baseline comparisons the performance of Bayesian estimation matched ML estimation using both diffuse and moderately informative prior specifications. These findings are consistent with theory that Bayesian estimation and ML estimation are generally equivalent using flat priors, and also that prior information is only negligibly influential given sufficient information in the data (Gelman et al., 2013). These findings also demonstrate that, in these conditions, there should be no statistical disadvantage to choosing Bayesian estimation over ML for IFA models, even if sparseness is not problematic. Although not studied here, Bayesian estimation may also be useful as an alternative estimator to ML for high-dimensional models and for assessing model fit (Béguin & Glas, 2001; Edwards, 2010).

Results comparing Bayesian and ML estimation in conditions where ML estimation performed poorly showed that by including prior information Bayesian estimation could provide improved efficiency and empirical power and eliminate extreme estimates. The moderately informative priors (specifying high probability that item intercepts and loadings were less than 7 in magnitude, loadings were positive, and the correlation between factors was positive) used here contained a sufficient amount of information to limit extreme estimates. However, flat priors lead to a complete lack of convergence for Bayesian estimation with sparse items.

In some sense, the comparison between ML and Bayesian estimation was not completely fair, for example the ML solutions were not constrained to have positive factor loadings, and it is possible to fix large estimates to a boundary value. Constraints were not placed on the ML solutions here because (1) these are not “true” ML solutions (Wasserman, 2005), and (2) standard errors are not available for fixed estimates. However, note from Tables 2 and 3 that very few of the extreme values observed were negative factor loadings.

The bias in some parameter estimates resulting from Bayesian estimation with moderately informative priors had a notable pattern. The correlation estimate was underestimated, loadings for items that were not sparse were overestimated, and factor loadings for items that were sparse were slightly underestimated. This pattern attributes relatively higher weight to non-sparse items; it is also interesting because ML estimation was more likely to yield extreme estimates (loadings and intercepts) for sparse items. In Bayesian estimation, less emphasis is based on unbiasedness and more emphasis is based on variance (Gelman et al., 2013, Ch. 4.5). Despite the tradeoff in bias, the overall efficiency of these parameter estimates, empirical power, and absence of extreme values were all an improvement over ML estimation.

The application to suicidality and insomnia symptoms also provides some further evidence for generalizability of these findings. This application demonstrates the beneficial “shrinkage” effect of Bayesian estimation with moderate priors across a range of item parameters and levels of endorsement.

Unique Contribution

Sparse items commonly arise in psychological research due to limited sample sizes and rare behaviors. This targeted simulation adds to the prior literature (Forero & Maydeu-Olivarez,

2009; Moshagen & Musch, 2013, Wirth & Edwards, 2007) that currently available estimation methods may not be appropriate in conditions characterized by sparseness. This means that researchers may be forced to drop sparse items and that some research questions involving sparse items cannot be asked using currently available methods.

Bayesian estimation for IFA models has been demonstrated previously (Albert & Chib, 1993; Béguin & Glas, 2001; Edwards, 2010; Patz & Junker, 1999, Song & Lee, 2002, 2012; Lee & Tang, 2006). This work builds on the prior research in three ways. First, this is the first study to use Bayesian estimation for IFA models with sparse indicators. Previous studies have motivated the use of Bayesian estimation for reasons such as estimating high-dimensional models (Edwards, 2010) or for advantages for testing hypotheses related to fit (Béguin & Glas, 2001). Second, previous research using Bayesian estimation for IFA models used relatively flat prior distributions and did not incorporate prior information to stabilize parameter estimates as I do here. Béguin & Glas (2001) examine different prior distributions as a sensitivity analysis, and Edwards (2010) incorporated minimal prior information to aid convergence, but this study is the first of my knowledge to utilize prior information about the expected range for parameter estimates in psychology to stabilize estimates for IFA. Third, previous research disseminating Bayesian estimation of IFA models used a combination of Gibbs and Metropolis-Hastings MCMC algorithms that were difficult to implement, requiring implementation using custom programming (e.g. Edwards, 2010, Patz & Junker, 1999) and offered less flexibility for model and prior specification. I demonstrate Bayesian estimation using Stan (Stan Development Team, 2015) which offers relatively efficient computation as well as flexible prior and model specification.¹³

In this study, I show that using prior information for Bayesian estimation of IFA models with sparse indicators can help stabilize estimates and improve power compared to ML. This provides a new tool for researchers to address the limitations of currently available estimation methods for a challenging problem that often arises in psychological research.

Recommendations for Applied Researchers

For well-determined models in moderate to large samples with moderate sparseness, it may not be necessary to take a Bayesian approach. However the above results and previous research suggest that Bayesian estimation should not give results inferior to ML estimation, if done correctly. If sparseness is not an issue, results should be comparable using either estimator. However if a research question requires modeling sparse data, a Bayesian estimation approach can be useful to stabilize estimates and increase statistical power by incorporating reasonable prior information. For some research questions (e.g. illicit drug use, early alcohol use), investing the time and effort to take a Bayesian estimation approach may maximize a researcher's ability to draw inferences from data that is exceedingly difficult to collect.

In practice it may be difficult to determine if ML estimates are “untrustworthy” – since extreme estimates may appear in converged solutions. However, if ML estimates appear

¹³The sampling efficiency for these conditions could be further improved using suggestions in the User's Guide (see Ch. 21, Stan Development Team, 2015), but the syntax becomes less intuitive and general.

unreasonable, this suggests that the researcher has prior information about parameter estimates (deeming estimates unreasonable requires knowledge about what is reasonable), which could be incorporated into a Bayesian specification. Prior information can include the expected direction of factor loadings, correlations between factors, and ranges of parameter estimates. The moderately informative prior I suggest here I believe is reasonable for a variety of applications, but ultimately this choice will depend on knowledge of the content area. Bayesian estimation with flat priors will offer no benefit over ML estimation for IFA models with sparse data.

Applied researchers should also be aware of difficulties in MCMC estimation. Specifically, sign indeterminacy is an issue for the scaling demonstrated here if prior information does not restrict the sign of the latent factors, and posterior inference may require post-processing to an interpretable solution. It is important to monitor convergence diagnostics including plots, and statistics such as the effective sample size and potential scale reduction factor. The method of Bayesian estimation demonstrated here may be adapted for many different models with different types of indicators, numbers of items or factors, and by incorporating predictors, and could also be extended to more comprehensive structural equation models. There are many helpful resources available specifically for Bayesian analysis using Stan (e.g., Stan Development Team, 2015; Gelman et al., 2013), which also includes an active online users group.

Limitations and Future Directions

Though a limited simulation study was used for this demonstration, it is possible to predict how results would extend to many different conditions based on theory. For example, with smaller factor loadings or fewer items, ML estimation would be expected to perform worse, and the benefit of Bayesian estimation may be greater, however this would still depend on the incorporation of sufficient prior information. The impact of extreme intercepts will vary by factor loading condition; with higher factor loadings the impact of extreme intercepts will be minimized, however note that marginal endorsement level and item loading are confounded (i.e., the same intercept yields different endorsement rates for different values of λ). Research for IFA models with continuous indicators (Gagné & Hancock, 2006; Marsh et al., 1998), also shows that stronger factor loadings improve the quality of solutions in finite samples, in terms of convergence and parameter estimate efficiency.

As in any line of enquiry, the present work answers some questions while also raising new ones. There are several remaining questions to be addressed by future research. Most notably, the models considered here were correctly specified. It is important to investigate if the use of priors to stabilize estimates could also have the unintended consequence of masking model misspecification. Related to this, an important line of research will be to investigate the utility of Bayesian methods for assessing model fit (e.g. posterior predictive checks, Bayesian model selection) in these models. Currently, methods for assessing model fit using ML estimation are limited. Limited-information methods for estimation provide tests for model fit, but are less appropriate for modeling sparse data (Wirth & Edwards, 2007; Rhemtulla et al., 2012).

Extending these results to polytomous items with more response categories also raises a number of interesting issues. Polytomous items contain more information than binary items; however they require estimating additional intercepts and the potential for varied mechanisms and patterns of sparseness raises additional complexity. With multiple ordered intercept categories, it is also necessary to constrain their order in the priors and estimation. Despite these unanswered questions, the present work provides an alternative to improve estimation for models with sparse endorsement.

Acknowledgments

I am grateful to Kiara Timpano and Brad Schmidt for providing the data used in the empirical example. I want to thank Patrick Curran, Amy Herring, David Thissen, Kenneth Bollen, Daniel Bauer, and Andrea Hussong for their feedback on this research and two anonymous reviewers for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the author alone, and endorsement by the author's institutions or the NIDA is not intended and should not be inferred.

Funding: This work was partially supported by Grant F31DA035523 from the National Institute on Drug Abuse.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

This research was supported in part by the National Institute on Drug Abuse of the National Institutes of Health under award number F31DA035523. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health. Portions of this work were previously reported as part of the author's doctoral dissertation. I am grateful to Drs. Patrick Curran, Amy Herring, David Thissen, Kenneth Bollen, Daniel Bauer, and Andrea Hussong from the University of North Carolina at Chapel Hill for their valuable advice and feedback on this research.

APPENDIX A. EXAMPLE MPLUS PROGRAM FOR IFA

```
DATA: FILE IS data1.dat;
VARIABLE: NAMES ARE dep1-dep5 drug1-drug5;
categorical = dep1-dep5 drug1-drug5;
model:
    [dep1$1*0];
    [dep2$1*0];
    [dep3$1*0];
    [dep4$1*0];
    [dep5$1*0];
    [drug1$1*0];
    [drug2$1*0];
    [drug3$1*0];
    [drug4$1*0];
    [drug5$1*0];
dep by dep1* dep2* dep3* dep4* dep5*;
dep@1;
    [dep@0];
drug by drug1* drug2* drug3* drug4* drug5*;
drug@1;
```

```

[drug@0];
dep with drug;
analysis:
estimator=ml;

```

APPENDIX B. STAN PROGRAM FOR IFA – MODERATE PRIORS

```

# 2-factor IFA for binary items
data {
  int<lower=0> N; // number of ppl
  int<lower=0> K; // number of items
  int y[N,K]; // Y matrix of P items for N ppl
}
transformed data {
  vector[2] mu;
  for (i in 1:2) mu[i] <- 0;
}
parameters {
  vector[2] eta[N]; // eta for each person
  real nu[K]; // int for item k
  real <lower=0> lambda[K]; // loading item k
  real <lower=0, upper=1> cov;
}
transformed parameters {
  matrix[2,2] sigma;
  sigma[1,1]<-1; sigma[1,2]<-cov;
  sigma[2,1]<-cov; sigma[2,2]<-1;
}
model {
  nu~ normal(0,3.57);
  lambda~normal(0,3.57);
  eta ~ multi_normal(mu,sigma);
  for (n in 1:N){
    for (k in 1:5){
      y[n,k]~ bernoulli_logit(-nu[k]+lambda[k]*eta[n,1]);
    }
    for (k in 6:K){
      y[n,k]~ bernoulli_logit(-nu[k]+lambda[k]*eta[n,2]);
    }
  }
}

```

References

- Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*. 1993; 88(422):669–679. DOI: 10.1080/01621459.1993.10476321
- Anderson JC, Gerbing DW. The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*. 1984; 49:155–173. DOI: 10.1007/bf02294170
- Béguin A, Glas C. MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*. 2001; 66(4):541–561. DOI: 10.1007/bf02296195
- Bentler PM, Chou C. Practical issues in structural modeling. *Sociological Methods & Research*. 1987; 16(1):78–117. DOI: 10.1177/0049124187016001004
- Berger JO, Bernardo JM. Ordered group reference priors with application to the multinomial problem. *Biometrika*. 1992; 79(1):25. doi: 10.1093/biomet/79.1.25
- Bollen KA, Bauldry S. Model identification and computer algebra. *Sociological Methods & Research*. 2010; 39(2):127–156. DOI: 10.1177/0049124110366238 [PubMed: 21769158]
- Bollen KA, Maydeu-Olivares A. A Polychoric Instrumental Variable (PIV) Estimator for Structural Equation Models with Categorical Variables. *Psychometrika*. 2007; 72(3):309–326. DOI: 10.1007/s11336-007-9006-3
- Cai L. A two-tier full-information item factor analysis model with applications. *Psychometrika*. 2010a; 75(4):581–612. DOI: 10.1007/s11336-010-9178-0
- Cai L. High-dimensional exploratory item factor analysis by a MetropolisHastings Robbins-Monro algorithm. *Psychometrika*. 2010b; 75:33–57. DOI: 10.1007/s11336-009-9136-x
- Carpenter B, Gelman A, Hoffman M, Lee D, et al. Stan: A probabilistic programming language. *Journal of Statistical Software*. 2017; 76(1)doi: 10.18637/jss.v076.i01
- ChassinL, , PressonC, , Il-ChoY, , LeeM, , MacyJ. Developmental Factors in Addiction: Methodological Considerations. In: MacKillopJ, , de WithH, editorsThe Wiley-Blackwell Handbook of Addiction PsychopharmacologyWiley-Blackwell; Oxford, UK: 2013
- Collins L, Schafer J, Kam C. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*. 2001; 6(4):330–351. DOI: 10.1037//1082-989x.6.4.330 [PubMed: 11778676]
- Depaoli S. The impact of inaccurate “informative” priors for growth parameters in bayesian growth mixture modeling. *Structural Equation Modeling*. 2014; 21(2):239–252. DOI: 10.1080/10705511.2014.882686
- Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid monte carlo. *Physics Letters B*. 1987; 195(2):216–222. DOI: 10.1016/0370-2693(87)91197-x
- Dunson DB, Dinse GE. Bayesian incidence analysis of animal tumorigenicity data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. 2001; 50(2):125–141. DOI: 10.1111/1467-9876.00224
- Edwards MC. A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*. 2010; 75:474–497. DOI: 10.1007/s11336-010-9161-9
- ForeroCarlos G, , Maydeu-OlivaresAlberto. Estimation of IRT Graded Response Models: Limited versus Full Information Methods. *Psychological Methods*. 2009; 14(3):275–299. DOI: 10.1037/a0015825 [PubMed: 19719362]
- Gagné PE, Hancock GR. Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*. 2006; 41:65–83. DOI: 10.1207/s15327906mbr4101_5 [PubMed: 26788895]
- Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*. 1990; 85(409):398. doi: 10.1080/01621459.1990.10476213
- GelmanA, , CarlinJB, , SternHS, , DunsonDB, , VehtariA, , RubinDB. Bayesian Data AnalysisChapman & Hall; 2013
- Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science*. 1992; 7(4):457–472. DOI: 10.1214/ss/1177011136

- Ghosh J, Dunson DB. Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics*. 2009; 18(2):306–320. DOI: 10.1198/jcgs.2009.07145 [PubMed: 23997568]
- Hallquist M, Wiley J. MplusAutomation: Automating Mplus Model Estimation and Interpretation 2014R package version 0.6-3 <https://CRAN.R-project.org/package=MplusAutomation>
- Hoffman M, Gelman A. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*. 2014; 15:1593–1623.
- Hussong AM, Flora DB, Curran PJ, Chassin LA, Zucker RA. Defining risk heterogeneity for internalizing symptoms among children of alcoholic parents. *Development and Psychopathology*. 2008; 20(1):165–193. DOI: 10.1017/s0954579408000084 [PubMed: 18211733]
- Jacobucci R, Grimm K, McArdle J. Regularized Structural Equation Modeling; *Structural Equation Modeling: A Multidisciplinary Journal* 2016; 112
- Kass RE, Wasserman L. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*. 1996; 91(435):1343–1370. DOI: 10.1080/01621459.1996.10477003
- Lee S, Song X. Basic and advanced bayesian structural equation modeling: With applications in the medical and behavioral sciences GB: Wiley; 2012
- Lee SY, Tang NS. Bayesian analysis of structural equation models with mixed exponential family and ordered categorical data. *British Journal of Mathematical and Statistical Psychology*. 2006; 59:151–172. DOI: 10.1348/000711005x81403 [PubMed: 16709284]
- Loken E. Identifiability constraints and the shape of the likelihood in confirmatory factor models. *Structural Equation Modeling*. 2005; 12:232–244.
- Marsh HW, Hau K, Balla JR, Grayson D. Is more ever too much? the number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*. 1998; 33(2):181–220. DOI: 10.1207/s15327906mbr3302_1 [PubMed: 26771883]
- Maxwell Scott E. The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*. 2004; 9(2):147–163. DOI: 10.1037/1082-989x.9.2.147 [PubMed: 15137886]
- Mislevy RJ. Bayes modal estimation in item response models. *Psychometrika*. 1986; 51(2):177–195. DOI: 10.1007/bf02293979
- Moshagen M, Musch J. Sample Size Requirements of the Robust Weighted Least Squares Estimator. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. 2014; 10(2):60–70. DOI: 10.1027/1614-2241/a000068
- Muthén B, du Toit SHC, Spisic D. Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. 1997 Unpublished manuscript.
- Muthén LK, Muthén BO. *Mplus User's Guide* Seventh. Los Angeles, CA: Muthén & Muthén; 2015
- Neal R. MCMC Using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo*. 2011; doi: 10.1201/b10905-6
- Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*. 1979; 44:443–460. DOI: 10.1007/bf02296207
- Park T, Casella G. The bayesian lasso. *Journal of the American Statistical Association*. 2008; 103(482):681–686. DOI: 10.1198/016214508000000337
- Patz RJ, Junker BW. A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*. 1999; 24(2):146–178. DOI: 10.2307/1165199
- Peddada SD, Dinse GE, Kissling GE. Incorporating historical control data when comparing tumor incidence rates. *Journal of the American Statistical Association*. 2007; 102(480):1212–1220. DOI: 10.1198/016214506000001356 [PubMed: 20396669]
- Rhemtulla M, Brosseau-Liard PÉ, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*. 2012; 17:354–373. DOI: 10.1037/a0029315 [PubMed: 22799625]

- Rindskopf D. Structural equation models: Empirical identification, heywood cases, and related problems. *Sociological Methods & Research*. 1984; 13(1):109–119. DOI: 10.1177/0049124184013001004
- R Core TeamR: A language and environment for statistical computingR Foundation for Statistical Computing; 2015Vienna, Austria. URL <http://www.R-project.org/>
- Savalei V. What to do about zero frequency cells when estimating polychoric correlations. *Structural Equation Modeling: A Multidisciplinary Journal*. 2011; 18(2):253–273. DOI: 10.1080/10705511.2011.557339
- Song XY, Lee SY. Analysis of structural equation model with ignorable missing continuous and polytomous data. *Psychometrika*. 2002; 67(2):261–288. DOI: 10.1007/bf02294846
- Stan Development Team . Stan Modeling Language Users Guide and Reference Manual, Version 2.7.0. 2015
- Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*. 1987; 52:393–408. DOI: 10.1007/bf02294363
- Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*. 1987; 82(398):528–540. DOI: 10.1080/01621459.1987.10478458
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996; 58(1):267–288. DOI: 10.1214/10-aos375supp
- WassermanL. All of StatisticsNew York, NY: Springer Science+Business Media, Inc; 2005
- Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. *Psychological Methods*. 2007; 12:58–79. DOI: 10.1037/1082-989x.12.1.58 [PubMed: 17402812]
- Yuan K, Wu R, Bentler P. Ridge structural equation modelling with correlation matrices for ordinal and continuous data. *British Journal of Mathematical & Statistical Psychology*. 2011; 64(1): 107.doi: 10.1348/000711010x497442 [PubMed: 21506947]
- Zhu M, Lu AY. The counter-intuitive non-informative prior for the bernoulli family. *Journal of Statistics Education*. 2004; 12(2):1–10.

Figure 1. MAD for ML and Bayesian estimation using moderately informative priors for conditions with N=500 and 2% endorsement

Median absolute deviation for parameter estimates using ML and Bayesian estimation with moderate priors. Results shown with 4/5 sparse items on one factor (Left) and with 4/5 sparse items on both factors (Right). Note that the results for ML estimation include only converged solutions and results for Bayesian estimation include solutions with above threshold effective sample size for all parameters, so the solution sets do not exactly overlap.

Figure 2. RMSE for ML and Bayesian estimation using moderately informative priors for conditions with N=500 and 2% sparseness

Note that the y-axes are different between plots, due to the extremely large discrepancy in RMSE values for each condition. Root-mean-square-error for parameter estimates using ML and Bayesian estimation with moderate priors. Results shown with 4/5 sparse items on one factor (Left) and with 4/5 sparse items on both factors (Right). Note that the results for ML estimation include only converged solutions and results for Bayesian estimation include solutions with above threshold effective sample size for all parameters, so the solution sets do not exactly overlap.

Figure 3. Estimates and confidence intervals using ML estimation in full sample (N=457) and randomly split halves

Note that the y-axes are truncated for comparison with subsequent figures, and estimates and confidence intervals fall outside of displayed range. Subscripts $i1$ denote insomnia item thresholds (v) and loadings (λ), subscripts $i2$ denote suicidality items, and σ_{12} is the factor correlation.

Figure 4. Posterior median estimates and 95% intervals using Bayesian estimation with moderate priors ($\sim N(0, 3.57)$) in full sample (N=457) and randomly split halves
 Subscripts 1 denote insomnia item thresholds (v) and loadings (λ), subscripts 2 denote suicidality items, and σ_{12} is the factor correlation.

Figure 5. Posterior median estimates and 95% intervals using Bayesian estimation in full sample (N=457) across range of three priors

Subscripts 11 denote insomnia item thresholds (v) and loadings (λ), subscripts 12 denote suicidality items, and σ_{12} is the factor correlation.

Figure 6. Posterior median estimates and 95% intervals using Bayesian estimation in sub-sample (N=223) across range of three priors

Subscripts 11 denote insomnia item thresholds (v) and loadings (λ), subscripts 12 denote suicidality items, and σ_{12} is the factor correlation.

Table 1

Recovery of population generating values with N=250 using ML estimation.

No Sparse Items (R=500)											
True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	Sig
ψ_{12}	0.3	0.29	0.29	0.10	-0.01	0.10	0.06	-0.03	0.13	0.46	0.86
λ	1.5	1.54	1.50	0.35	0.04	0.35	0.21	0.74	1.05	2.18	1.00
ν	0	0.00	0.00	0.19	0.00	0.19	0.12	-0.68	-0.30	0.31	0.05
4/5; 0/5 Sparse Items, 4% endorsement (R=487, 25 with extreme values)											
ψ_{12}	0.3	0.29	0.28	0.14	-0.01	0.14	0.09	-0.17	0.08	0.53	0.62
λ	1.5	1.66	1.52	0.69	0.16	0.56	0.25	0.21	1.00	2.78	0.90
λ_{SP}	1.5	1.94	1.48	2.62	0.44	2.60	0.51	-2.79	0.39	4.39	0.46
ν	0	0.00	0.00	0.20	0.00	0.20	0.12	-1.22	-0.31	0.32	0.05
ν_{SP}	4.11	5.07	4.16	4.90	0.96	4.86	0.59	2.65	3.20	8.59	0.89
4/5; 4/5 Sparse Items, 4% endorsement (R=474, 85 with extreme values)											
ψ_{12}	0.3	0.29	0.29	0.18	-0.01	0.18	0.12	-0.41	0.02	0.57	0.44
λ	1.5	2.29	1.69	1.65	0.79	1.82	0.68	0.12	0.72	5.25	0.40
λ_{SP}	1.5	2.29	1.50	6.80	0.79	5.69	0.50	-0.71	0.47	4.97	0.49
ν	0	0.00	0.00	0.27	0.00	0.27	0.14	-1.27	-0.41	0.41	0.02
ν_{SP}	4.11	5.66	4.12	13.24	1.55	10.90	0.57	2.63	3.18	9.40	0.89
4/5; 0/5 Sparse Items, 2% endorsement (R=471, 139 with extreme values)											
ψ_{12}	0.3	0.30	0.28	0.17	0.00	0.17	0.10	-0.28	0.07	0.59	0.56
λ	1.5	1.68	1.53	0.71	0.18	0.57	0.26	0.08	0.96	2.96	0.85
λ_{SP}	1.5	4.09	1.48	34.01	2.59	25.22	0.74	-560.5	0.00	21.18	0.15
ν	0	0.00	0.00	0.20	0.00	0.20	0.13	-1.29	-0.31	0.33	0.04
ν_{SP}	4.9	12.83	5.04	76.89	7.93	57.37	0.91	3.19	3.70	49.29	0.77
4/5; 4/5 Sparse Items, 2% endorsement (R=409, 224 with extreme values)											

No Sparse Items (R=500)													
μ_{12}	0.3	0.35	0.32	0.23	0.05	0.23	0.14	-0.28	0.02	0.78	0.98	0.91	0.35
λ	1.5	2.46	1.86	1.88	0.96	2.09	0.99	-0.34	0.46	5.28	12.11	0.92	0.06
λ_{sp}	1.5	13.01	1.48	280.6	11.51	213.4	0.73	-5937	-0.12	15.31	9360	0.91	0.15
ν	0	0.01	0.01	0.28	0.01	0.28	0.15	-1.64	-0.39	0.43	1.3	0.99	0.01
ν_{sp}	4.9	47.50	5.05	670.4	42.60	515.1	0.91	2.99	3.72	43.46	21790	0.88	0.76

Note. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5th and 95th quantile estimates, 95% CI is the coverage for the 95% confidence interval, and Sig is the proportion of significant estimates. R is the number of replications with technically converged solutions.

Table 2

Recovery of population generating values with N=500 using ML estimation.

No Sparse Items (R=500)												
True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	95% CI	Sig
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.04	0.14	0.19	0.41	0.51	0.95	1.00
λ	1.5	1.53	1.51	0.23	0.03	0.15	0.85	1.18	1.93	2.66	0.96	1.00
ν	0	0.00	0.13	0.00	0.13	0.09	-0.42	-0.21	0.22	0.47	0.95	0.05
4/5; 0/5 Sparse Items, 4% endorsement (R=498, 2 with extreme values)												
ψ_{12}	0.3	0.30	0.30	0.09	0.00	0.06	-0.01	0.15	0.44	0.69	0.95	0.90
λ	1.5	1.56	1.50	0.41	0.06	0.34	0.17	0.50	1.13	2.12	0.96	0.97
λ_{SP}	1.5	1.61	1.49	0.91	0.11	0.88	0.35	-0.10	0.74	2.74	0.95	0.91
ν	0	0.00	0.00	0.13	0.00	0.13	0.09	-0.64	-0.21	0.51	0.96	0.04
ν_{SP}	4.11	4.37	4.14	1.60	0.26	1.46	0.42	2.95	3.38	5.92	0.94	0.98
4/5; 4/5 Sparse Items, 4% endorsement (R=500, 8 with extreme values)												
ψ_{12}	0.3	0.30	0.30	0.12	0.00	0.12	0.08	-0.03	0.11	0.50	0.93	0.74
λ	1.5	1.91	1.60	1.14	0.41	1.21	0.40	0.51	0.92	4.34	0.95	0.81
λ_{SP}	1.5	1.63	1.50	1.52	0.13	1.18	0.36	-0.35	0.77	2.73	0.95	0.92
ν	0	0.00	0.00	0.16	0.00	0.16	0.09	-0.68	-0.26	0.24	0.97	0.03
ν_{SP}	4.11	4.39	4.12	2.69	0.28	1.94	0.42	2.77	3.38	5.87	0.94	0.98
4/5; 0/5 Sparse Items, 2% endorsement (R=489, 44 with extreme values)												
ψ_{12}	0.3	0.29	0.28	0.11	-0.01	0.11	0.07	-0.04	0.12	0.47	0.92	0.80
λ	1.5	1.64	1.52	0.59	0.14	0.44	0.18	0.31	1.12	2.68	0.95	0.91
λ_{SP}	1.5	2.29	1.47	5.10	0.79	5.14	0.47	-1.51	0.44	4.05	0.94	0.52
ν	0	0.00	0.00	0.14	0.00	0.14	0.09	-0.65	-0.23	0.22	0.96	0.04
ν_{SP}	4.9	6.76	4.90	11.17	1.86	11.29	0.58	3.35	3.92	9.18	0.91	0.92
4/5; 4/5 Sparse Items, 2% endorsement (R=458, 82 with extreme values)												

No Sparse Items (R=500)													
μ_{12}	0.3	0.30	0.28	0.16	0.00	0.16	0.10	-0.12	0.07	0.58	0.98	0.89	0.54
λ	1.5	2.21	1.69	1.46	0.71	1.61	0.65	0.19	0.74	4.95	9.42	0.93	0.41
λ_{sp}	1.5	3.29	1.46	55.94	1.79	39.96	0.49	-2259	0.45	3.99	1751	0.94	0.50
λ	0	0.00	0.00	0.19	0.00	0.19	0.10	-1.23	-0.31	0.28	0.97	0.97	0.03
λ_{sp}	4.9	11.92	4.89	133.9	7.02	93.59	0.61	3.34	3.90	9.16	5792	0.91	0.91

Note. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5th and 95th quantile estimates, 95% CI is the coverage for the 95% confidence interval, and Sig is the proportion of significant estimates. R is the number of replications with technically converged solutions.

Recovery of population generating values for baseline condition using Bayesian estimation with flat and moderate priors

Table 3

Flat Prior N=250 (R=500)										
True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Sig
ψ_{12}	0.3	0.29	0.29	0.09	-0.01	0.09	0.06	0.14	0.45	1.00
λ	1.5	1.59	1.55	0.38	0.09	0.39	0.22	0.74	2.29	1.00
ν	0	0.00	0.19	0.00	0.19	0.12	-0.69	-0.3	0.32	0.05
Moderate Prior N=250 (R=500)										
ψ_{12}	0.3	0.29	0.28	0.09	-0.01	0.09	0.06	0.14	0.45	1.00
λ	1.5	1.59	1.55	0.38	0.09	0.39	0.22	0.75	2.28	1.00
ν	0	0.00	0.19	0.00	0.19	0.12	-0.70	-0.3	0.31	0.05
Flat Prior N=500 (R=498)										
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.14	0.19	1.00
λ	1.5	1.56	1.54	0.24	0.06	0.25	0.16	0.86	1.20	1.00
ν	0	0.00	0.13	0.00	0.13	0.09	-0.43	-0.21	0.22	0.05
Moderate Prior N=500 (R=500)										
ψ_{12}	0.3	0.30	0.30	0.06	0.00	0.06	0.04	0.14	0.19	1.00
λ	1.5	1.55	1.54	0.24	0.05	0.24	0.16	0.85	1.19	1.00
ν	0	0.00	0.13	0.00	0.13	0.09	-0.43	-0.21	0.22	0.05

Note. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5th and 95th quantile estimates, 95% CI is the confidence coverage for the 95% credible interval, and Sig is the proportion of significant estimates. R is the number of replications with effective sample size > 100 for all parameters.

Table 4

Recovery of population generating values for sparse conditions using Bayesian estimation with N=250 and moderately informative priors

4/5; 0/5 Sparse, 4% Endorsement (R=500)											
True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	Sig
ψ_{12}	0.3	0.28	0.27	0.11	-0.02	0.11	0.07	0.08	0.12	0.47	1.00
λ	1.5	1.76	1.59	0.69	0.26	0.59	0.27	0.35	1.07	3.35	1.00
λ_{SP}	1.5	1.49	1.43	0.62	-0.01	0.62	0.41	0.24	0.56	2.61	1.00
ν	0	-0.01	-0.01	0.20	-0.01	0.20	0.13	-0.98	-0.33	0.32	0.05
ν_{SP}	4.11	4.26	4.14	0.75	0.15	0.76	0.48	2.75	3.25	5.69	1.00
4/5; 4/5 Sparse, 4% Endorsement (R=497)											
ψ_{12}	0.3	0.27	0.26	0.12	-0.03	0.12	0.08	0.04	0.10	0.49	1.00
λ	1.5	2.54	2.43	1.18	1.04	1.57	0.99	0.26	0.89	4.51	1.00
λ_{SP}	1.5	1.51	1.45	0.62	0.01	0.62	0.43	0.14	0.59	2.64	1.00
ν	0	-0.03	-0.02	0.25	-0.03	0.25	0.15	-0.87	-0.44	0.37	0.05
ν_{SP}	4.11	4.23	4.09	0.74	0.12	0.75	0.46	2.66	3.25	5.67	1.00
4/5; 0/5 Sparse, 2% Endorsement (R=480)											
ψ_{12}	0.3	0.27	0.26	0.10	-0.03	0.11	0.07	0.06	0.11	0.46	1.00
λ	1.5	1.84	1.63	0.77	0.34	0.64	0.27	0.49	1.08	3.78	1.00
λ_{SP}	1.5	1.37	1.28	0.60	-0.13	0.62	0.44	0.23	0.50	2.46	1.00
ν	0	-0.01	0.00	0.20	-0.01	0.20	0.13	-1.16	-0.33	0.33	0.05
ν_{SP}	4.9	4.95	4.87	0.77	0.05	0.77	0.56	3.22	3.80	6.29	1.00
4/5; 4/5 Sparse, 2% Endorsement (R=456)											
ψ_{12}	0.3	0.26	0.25	0.11	-0.04	0.12	0.08	0.06	0.11	0.47	1.00
λ	1.5	3.01	3.16	1.10	1.51	1.87	0.80	0.35	0.96	4.56	1.00
λ_{SP}	1.5	1.37	1.30	0.59	-0.13	0.60	0.43	0.21	0.51	2.40	1.00
ν	0	-0.02	-0.02	0.28	-0.02	0.28	0.18	-1.03	-0.47	0.43	0.05
ν_{SP}	4.9	4.95	4.88	0.76	0.05	0.76	0.56	3.06	3.85	6.23	1.00

Note. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5th and 95th quantile estimates, 95% CI is the confidence coverage for the 95% credible interval, and Sig is the proportion of significant estimates. R is the number of replications with effective sample size > 100 for all parameters.

Table 5

Recovery of population generating values for sparse conditions using Bayesian estimation with N=500 and moderately informative priors

4/5; 0/5 Sparse, 4% Endorsement (R=499)											
True	Mean Est	Med Est	SD Est	Raw Bias	RMSE	MAD	Min Est	.05 Q Est	.95 Q Est	Max Est	Sig
ψ_{12}	0.3	0.28	0.28	0.09	-0.02	0.09	0.05	0.15	0.41	0.66	1.00
λ	1.5	1.63	1.54	0.41	0.13	0.39	0.18	1.16	2.44	5.00	1.00
λ_{SP}	1.5	1.49	1.46	0.54	-0.01	0.50	0.33	0.72	2.39	3.59	1.00
ν	0	0.00	-0.01	0.14	0.00	0.13	0.09	-0.63	0.21	0.44	0.05
ν_{SP}	4.11	4.19	4.09	0.68	0.08	0.62	0.38	3.39	5.36	6.89	1.00
4/5; 4/5 Sparse, 4% Endorsement (R=497)											
ψ_{12}	0.3	0.27	0.27	0.1	-0.03	0.10	0.07	0.12	0.45	0.61	1.00
λ	1.5	2.17	1.93	1.34	0.67	1.19	0.62	0.55	4.11	5.06	1.00
λ_{SP}	1.5	1.49	1.46	0.54	-0.01	0.50	0.34	0.76	2.38	3.44	1.00
ν	0	-0.01	-0.01	0.20	-0.01	0.16	0.10	-0.58	0.25	0.87	0.05
ν_{SP}	4.11	4.18	4.07	0.68	0.07	0.60	0.37	3.39	5.30	7.00	1.00
4/5; 4/5 Sparse, 2% Endorsement (R=491)											
ψ_{12}	0.3	0.26	0.26	0.09	-0.04	0.10	0.06	0.13	0.41	0.70	1.00
λ	1.5	1.76	1.57	0.68	0.26	0.50	0.19	0.41	3.52	5.27	1.00
λ_{SP}	1.5	1.39	1.35	0.56	-0.11	0.57	0.38	0.55	2.39	3.41	1.00
ν	0	-0.01	0.00	0.14	-0.01	0.14	0.09	-0.67	0.22	0.52	0.05
ν_{SP}	4.9	4.91	4.79	0.71	0.01	0.71	0.44	3.99	6.26	7.94	1.00
4/5; 4/5 Sparse, 2% Endorsement (R=466)											
ψ_{12}	0.3	0.24	0.23	0.10	-0.06	0.12	0.07	0.10	0.44	0.72	1.00
λ	1.5	2.77	2.79	1.17	1.27	1.73	0.97	0.25	4.65	5.46	1.00
λ_{SP}	1.5	1.39	1.34	0.57	-0.11	0.58	0.40	0.19	2.42	3.25	1.00
ν	0	-0.02	-0.01	0.20	-0.02	0.20	0.12	-0.67	0.30	0.63	0.07
ν_{SP}	4.9	4.93	4.80	0.74	0.03	0.74	0.47	3.34	6.34	7.97	1.00

Note. Med is the median estimate, SD Est is the empirical standard deviation of the estimate, .05 Q and .95 Q are the 5th and 95th quantile estimates, 95% CI is the confidence coverage for the 95% credible interval, and Sig is the proportion of significant estimates. R is the number of replications with effective sample size > 100 for all parameters.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript