Received: 1 March 2023 | Accepted: 17 April 2023

DOI: 10.1111/bmsp.12314

the british psychological society
promoting excellence in psychology

ARTICLE

Assessment of generalised Bayesian structural equation models for continuous and binary data

Konstantinos Vamvourellis | **Konstantinos Kalogeropoulos** | **Irini Moustaki**

Department of Statistics, London School of Economics, London, UK

Correspondence
Konstantinos Kalogeropoulos, Department of Statistics, London School of Economics, London, UK.
Email: k.kalogeropoulos@lse.ac.uk

Abstract

The paper proposes a novel model assessment paradigm aiming to address shortcoming of posterior predictive p -values, which provide the default metric of fit for Bayesian structural equation modelling (BSEM). The model framework presented in the paper focuses on the approximate zero approach (*Psychological Methods*, **17**, 2012, 313), which involves formulating certain parameters (such as factor loadings) to be approximately zero through the use of informative priors, instead of explicitly setting them to zero. The introduced model assessment procedure monitors the out-of-sample predictive performance of the fitted model, and together with a list of guidelines we provide, one can investigate whether the hypothesised model is supported by the data. We incorporate scoring rules and cross-validation to supplement existing model assessment metrics for BSEM. The proposed tools can be applied to models for both continuous and binary data. The modelling of categorical and non-normally distributed continuous data is facilitated with the introduction of an item-individual random effect. We study the performance of the proposed methodology via simulation experiments as well as real data on the 'Big-5' personality scale and the Fagerstrom test for nicotine dependence.

KEYWORDS

Bayesian model assessment, cross-validation, factor analysis, scoring rules

1 | INTRODUCTION

Structural equation modelling (SEM) is a general framework for testing research hypotheses arising in psychology and social sciences in general (Bollen, 1989). Initial inference methods for SEM have mostly been frequentist, but recently their Bayesian counterpart has gained popularity (e.g. Dunson et al., 2005; Kaplan, 2014; Merkle & Rosseel, 2015; Scheines et al., 1999; Van De Schoot et al., 2017).

In this paper, we focus on the Bayesian structural equation modelling (BSEM) framework introduced by Muthén and Asparouhov (2012). Depending on the substantive theory to be tested, structural equation models impose restrictions on some model parameters. Usually, some parameters are set to zero and thus not estimated at all (e.g. cross-loadings, error correlations, regression coefficients). Muthén and Asparouhov (2012) suggested treating such parameters as approximate rather than exact zero by assigning informative priors on them that place a large mass around zero; we will refer to this approach as the *approximate zero framework*. The introduction of such informative priors is convenient in situations where there are concerns regarding the fit of the exact zero model. More specifically, it allows using the model as an exploratory tool to identify the source of model misfit. An alternative option for such a task is the use of modification indices (e.g. MacCallum et al., 1992). More specifically, a modification index measures the improvement in model fit that would result if a previously omitted parameter were to be freely estimated. This can often lead to a model unsupported by the hypothesised substantive theory. Moreover, the greedy nature of the procedure does not guarantee convergence to an optimal model (Asparouhov et al., 2015; Muthén & Asparouhov, 2012; Stromeyer et al., 2015). On the other hand, the approximate zero approach provides information on model modification in one go, while the hypothesised theory is reflected clearly via the priors on the loadings. A related approach is to use Bayesian model searching, which could be implemented using spike and slab priors and stochastic search variable selection (see e.g. Lu et al., 2016).

Furthermore, model fit assessment is challenging in its own right, but it becomes even more complicated under the approximate zero framework. Specifically, it is unclear whether fitting such models can result in good fit indices, even when they are incorrectly specified (Stromeyer et al., 2015). In terms of specific model fit indices, several approaches exist in the literature, with the posterior predictive p -values (PPP) Meng (1994) being the most widely used. However, concerns have been raised regarding their suitability in this framework (e.g. Hoijtink & van de Schoot, 2018). Special consideration must be given to the choice of prior distributions for the model parameters, which could affect the PPP performance (Liang, 2020; MacCallum et al., 2012; Van Erp et al., 2018). Perhaps a more fundamental question is whether priors should be set on the basis of fit indices, rather than formal Bayesian model choice quantities, such as the Bayes factor. Nevertheless, the Bayes factor requires calculating the model evidence, or else marginal likelihood (Gelman et al., 2017), which can be quite a challenging task, especially in models with latent variables (e.g. Lopes & West, 2004; Vitoratou et al., 2014). Moreover, the Bayes factor is a relative measure and therefore does not directly address the question of whether a model fits the data well. Asparouhov et al. (2015) suggest avoiding the use of the approximate zero model to reach binary decisions on goodness of fit but instead use it as an exploratory tool leaving the choice up to the subject matter experts.

In this paper, we focus on improving the BSEM framework introduced by Muthén and Asparouhov (2012) in two aspects. First, we introduce an item-individual random effect term to allow for error correlations in the measurement model for data other than continuous such as binary data using the logit link. Second, we focus on the task of assessing model fit under the approximate zero framework (Asparouhov & Muthén, 2021a; Garnier-Villarreal & Jorgensen, 2020). We develop a decision framework that monitors the out-of-sample predictive performance to explore model misfit and assess the severity of the lack of fit. The proposed framework combines fit and out-of-sample predictive performance indices from different models.

Our approach aims to reach a middle ground between exploring lack of fit and assessing its severity. This is done by developing a decision framework that monitors the out-of-sample predictive performance to explore model misfit (i.e. validity of the hypothesised theory). The proposed

decision framework uses collectively fit indices and scoring rules via cross-validation to examine whether the approximate zero parameters are picking up random noise rather than systematic patterns in the data. From a machine learning viewpoint, cross-validation is one of the standard tools to guard against overfit while at the same time ensuring a good fit. The advantages of using cross-validation to measure model performance have been noted in the SEM context (e.g. Browne, 2000; MacCallum et al., 1992; Stromeyer et al., 2015, recommendation 4). An intuitive argument in favour of cross-validation is that if a measurement scale does not generalise well in parts of the existing data, it is highly unlikely that it will in future data. Merkle et al. (2019) utilise the deviance information criterion, watanabe akaike information criterion, and pareto smoothed importance sampling indices, which can be considered as approximations of cross-validation. While not always accurate (e.g. Plummer, 2008), these indices can still be effectively employed within our framework, serving as substitutes for cross-validated scoring rules. This is particularly advantageous since they can be readily computed using software tools like Stan and its associated packages. From a Bayesian viewpoint, cross-validation, when combined with the log posterior predictive scoring rule, has tight connections with the model evidence and is less sensitive to priors (Fong & Holmes, 2020). The proposed framework is developed by combining fit and out-of-sample predictive performance indices from different models.

The paper is structured as follows. Section 2 introduces the generalised Bayesian SEM framework. In Section 3, we present the methodology for assessing Bayesian SEM models. Section 4 consists of several simulation experiments that illustrate and evaluate the proposed methodology. In Section 5, we apply the methodology to two real-world datasets: the British Household Panel Survey (BHPS) to examine the 'Big 5' personality factors and the Fagerstrom Test for Nicotine Dependence (FTND). Finally, Section 6 provides a conclusion with relevant discussion and potential extensions. The code for this work is available in the accompanying repository 'bayes-sem' hosted on github.

2 | GENERALISED FRAMEWORK FOR BAYESIAN SEM

2.1 | Factor model specification

Suppose there are p observed correlated variables (items) denoted by $\mathbf{y} = (y_1, \ldots, y_p)$. One wants to find a set of continuous latent variables (factors) $\mathbf{z} = (z_1, \ldots, z_q)$, fewer in number than the observed variables, that contain essentially the same information. The factors are supposed to account for the dependencies among the observed variables in the sense that if the factors are held fixed, the observed variables would be independent (conditional independence assumption). Categorical variables (binary and ordinal) can be accommodated in the same framework as for continuous data by assuming that the categorical responses are manifestations of underlying (latent) continuous variables denoted by $\mathbf{y}^* = (y_1^*, \ldots, y_p^*)$. When continuous variables are analysed, $y_j = y_j^*$, $(j = 1, \ldots, p)$. The classical linear factor analysis model (*measurement model*) is

$$\mathbf{y}_i^* = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{1}$$

where $\boldsymbol{\alpha}$ is a $p \times 1$ vector of intercept parameters, $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings, and n is the sample size. The vector of latent variables $\mathbf{z}_i$ has a normal distribution, $\mathbf{z}_i \sim N_q(0, \boldsymbol{\Phi})$, where the covariance matrix $\boldsymbol{\Phi}$ is either unstructured or defined by a parametric model that relates latent variables to each other and observed covariates (*structural model*). The $\boldsymbol{\epsilon}_i$ are error terms assumed to be independent of each other and of the $\mathbf{z}_i$.

For binary data, the connection between the observed binary variable y_j and the underlying variable y_j^* is $y_j = \boldsymbol{I}(y_j^* > 0)$.

1https://github.com/bayesways/bayes-sem/.

More specifically, for a binary item j and individual i, the probability of success (positive) response is given by

$$P(y_{ij} = 1|\mathbf{z}_i) = P(y_{ij}^* > 0|\mathbf{z}_i) = P(\alpha_j + \Lambda_j \mathbf{z}_i + \epsilon_{ij} > 0|\mathbf{z}_i) = P(\epsilon_{ij} < \alpha_j + \Lambda_j \mathbf{z}_i|\mathbf{z}_i) = F(\alpha_j + \Lambda_j \mathbf{z}_i), \quad (2)$$

where F stands for the cumulative distribution function (CDF) of $\boldsymbol{\epsilon}_{ij}$. Finally, the model becomes

$$F^{-1}\{P(y_{ij} = 1)\} = \alpha_j + \Lambda_j \mathbf{z}_i, \quad (3)$$

where F^{-1} is the inverse of the CDF, also known as the link between the probability of success and the linear predictor. Specific choices for the distribution of the error term $\boldsymbol{\epsilon}_i$ lead to the following well-known models:

$$\boldsymbol{\epsilon}_i \sim \begin{cases} N(0_p, \boldsymbol{\Psi}), \quad \boldsymbol{\Psi} = \text{diag}(\psi_1^2, \ldots, \psi_p^2), & \text{if } \mathbf{y}_i \text{ is continuous,} \\ N(0_p, \boldsymbol{\Psi}), \quad \boldsymbol{\Psi} = I_p, & \text{if } \mathbf{y}_i \text{ is binary and } F^{-1} \text{ is the inverse CDF of the normal,} \\ \prod_{j=1}^{J} \text{Logistic}(0, \pi^2/3), & \text{if } \mathbf{y}_i \text{ is binary and } F^{-1} \text{ is the inverse CDF of the logistic,} \end{cases} \quad (4)$$

where 0_p is a p-dimensional vector of zeros, and I_p denotes the identity matrix of dimension p. The inverse CDF of the normal and the logistic distributions are known as the probit and logit links respectively. In the case of continuous or categorical items with the probit link, the marginal distribution of $\mathbf{y}_i^*$ is

$$\mathbf{y}_i^* \sim N(\boldsymbol{\alpha}, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}). \quad (5)$$

However, such an expression is not available for the logit model.

The model defined by (1) and (4) applies to confirmatory factor analysis (CFA) and exploratory factor analysis (EFA), and the differences between them are expressed in terms of restrictions on the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$. CFA postulates certain relationships among the observed and latent variables assuming a prespecified pattern for the model parameters (factor loadings, residual variances). For example, this is achieved by setting several elements of $\boldsymbol{\Lambda}$ to zero that are referred to as cross-loadings. EFA analyses a set of correlated observed variables without knowing in advance either the number of factors that are required to explain their interrelationships or their meaning or labelling. Finally, SEM adds an additional model (structural model) to the measurement models defined under the CFA by modelling the relationships among the latent variables according to a hypothesised theory. The assumption of conditional independence (common to EFA and CFA) is equivalent to setting the off-diagonal terms in the covariance of the $\boldsymbol{\epsilon}$, also known as error correlations, to zero.

2.2 | Generalised Bayesian model framework

The Bayesian SEM approach, introduced in Muthén and Asparouhov (2012) under the approximate zero framework, is capable of handling continuous items. Furthermore, it can be extended to binary and ordinal data using the probit specification.

We propose a distinct model specification that enables the modelling of categorical data with different links, such as the logit, and also allows the estimation of item-individual residuals. In our paper, we focus on the binary case; however, our findings can be extended to ordinal data as well. To achieve this, we expand model (1) by separating $\boldsymbol{\epsilon}_i$ into two parts: $\mathbf{u}_i$, a p-dimensional vector of random effects with a non-diagonal covariance matrix $\boldsymbol{\Omega}$, and $\mathbf{e}_i$, an error term with a diagonal covariance matrix $\boldsymbol{\Psi}$. The model is as follows:

$$\mathbf{y}_i^* = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\mathbf{z}_i + \mathbf{u}_i + \mathbf{e}_i, \tag{6}$$

The item-individual specific random effect, $\mathbf{u}_i$, is designed to capture associations among the variables that are relatively small in magnitude, beyond those explained by the vector of latent variables $\mathbf{z}_i$. These associations may be due to question wording, method effect, or other factors. Additionally, the cross-loadings $\boldsymbol{\Lambda}$ in (6) and (7) are non-zero parameters that are assigned informative priors centred around zero, such as $N(0, 0.01)$. For continuous normally distributed data, Model (6) coincides with the model proposed in Muthén and Asparouhov (2012) and can be written

$$\mathbf{y}_i^* \sim N(\boldsymbol{\alpha}, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Omega} + \boldsymbol{\Psi}), \quad i = 1, \ldots, n, \tag{7}$$

where $\mathbf{u}_i \sim N(0, \boldsymbol{\Omega})$, $\mathbf{z}_i \sim N(0, \boldsymbol{\Phi})$, and $\mathbf{e}_i \sim N(0, \boldsymbol{\Psi})$. Compared to the model in (5), the aim is to move from a diagonal matrix $\boldsymbol{\Psi}$ to an almost diagonal one, $\boldsymbol{\Psi} + \boldsymbol{\Omega}$, so that the error correlations are not substantial. This can be achieved by assigning an informative prior on the non-diagonal $\boldsymbol{\Omega}$ to ensure that its magnitude is low compared to $\boldsymbol{\Psi}$.

The generalised framework of (6) provides several extensions. It is possible to define the approximate zero model for logistic models by assuming $e_{ij} \sim \text{Logistic}(0, \pi^2/3)$ (see Section 2.2.2 for details). Other distributions (e.g. t-distribution, non-normal) can also be assumed for $\mathbf{e}_i$ and $\mathbf{u}_i$. Setting $\boldsymbol{\Phi} = I_q$ in (6) leads to the EFA model. However, fitting such a model with Markov Chain Monte Carlo (MCMC) may be challenging, as discussed subsequently (see also Conti et al., 2014; Erosheva & Curtis, 2017; Frühwirth-Schnatter & Lopes, 2018; Lopes & West, 2004, for some relevant Bayesian EFA schemes).

Inference is carried by adopting a fully Bayesian framework, which requires assigning priors on all the model parameters $\boldsymbol{\theta}$, denoted by $\pi(\boldsymbol{\theta})$, and proceeding based on their posterior given the data $\boldsymbol{Y} = \{\mathbf{y}_i\}_{i=1}^n$, denoted by $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$, obtained via Bayes theorem. A key feature of the approximate zero framework is that the priors on the cross-loadings given in $\boldsymbol{\Lambda}$ and the error covariances of $\boldsymbol{\Omega}$ are informative and point to zero. In the next section, we discuss in detail the model and prior specifications for continuous and binary data.

2.2.1 | Models and priors for continuous normally distributed data

The model in (6) originates from the following specification:

$$\begin{cases} \mathbf{y}_i = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\mathbf{z}_i + \mathbf{u}_i + \mathbf{e}_i \\ \mathbf{z}_i \sim N(0, \boldsymbol{\Phi}) \\ \mathbf{e}_i \sim N(0, \boldsymbol{\Psi}) \\ \mathbf{u}_i \sim N(0, \boldsymbol{\Omega}) \end{cases} \tag{8}$$

As mentioned earlier, it is possible to assign non-normal distributions to $\mathbf{e}_i$, $\mathbf{u}_i$, and even $\mathbf{z}_i$. However, if we assume that all of these variables follow a normal distribution, the following augmentation is equivalent:

$$\begin{cases} \mathbf{y}_i|\mathbf{u}_i \sim N\left(\boldsymbol{\alpha} + \mathbf{u}_i,\ \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}\right) \\ \mathbf{u}_i \sim N(0, \boldsymbol{\Omega}). \end{cases} \tag{9}$$

In the model, the non-diagonal covariance matrix $\boldsymbol{\Omega}$, which represents the error correlations, is assigned the inverse Wishart distribution with an identity scale matrix and $p + 6$ degrees of freedom. This choice reflects prior beliefs that the residual covariances are close to zero. This prior specification is similar to the approach taken in the work by Muthén and Asparouhov (2012), and you can refer to

Appendix A.1 of this paper for more details. As for $\mathbf{\Phi}$, it is specified as a covariance matrix. Additionally, for identification purposes, the primary loadings in the $\mathbf{\Lambda}$ matrix are set to 1.

The choice of prior distributions for the elements of the $\mathbf{\Lambda}$ matrix depends on whether they are considered as cross-loadings or free parameters in the hypothesised model. For the cross-loadings, a common approach, as used in Muthén and Asparouhov (2012), is to assign normal distributions with a mean of zero and a variance of .01. However, assigning large variance normal priors to the remaining parameters of $\mathbf{\Lambda}$ can lead to issues such as Lindley's paradox Lindley (1957). To address this issue, unit information priors Kass and Wasserman (1996) are often used. Unit information priors set the prior variances in a way that corresponds to the information from a single observation point. In the context of EFA, Lopes and West (2004) and Ghosh and Dunson (2009) proposed the following unit information priors:

$$\Lambda_{ij} \sim N(0, \psi_j^2) \tag{10}$$

where ψ_j^2 are the idiosyncratic variances of the diagonal matrix $\mathbf{\Psi}$ that are treated as unknown parameters. However, assigning the foregoing priors may lead to issues when ψ_j^2 values are quite small because it is crucial to differentiate them from the prior variance of .01 used for the cross-loadings. Therefore, a fixed value may be used for the prior variance of the free elements of $\mathbf{\Lambda}$, based on preliminary estimates.

Regarding the diagonal matrix $\mathbf{\Psi}$, independent inverse gamma priors, introduced in Frühwirth-Schnatter and Lopes (2018) and used in Conti et al. (2014), can be assigned on each ψ_j^2. The hyperparameters of these inverse gamma priors are set to give very small weight to Heywood cases. Specifically, the prior on the idiosyncratic variance is

$$\psi_j^2 \sim \text{InvGamma}(c_0, (c_0 - 1)/(S_y^{-1})_{jj}),$$

where S_y is the empirical covariance matrix, and c_0 is a constant chosen by the researcher to limit the probability of running into Heywood issues. Heywood problems arise when

$$1/\psi_j^2 \geq (S_y^{-1})_{jj}.$$

Following Frühwirth-Schnatter and Lopes (2018) and Conti et al. (2014), the constant c_0 can be chosen to keep the prior probability of such an event quite small. For our application, we chose $c_0 = 2.5$. This data-dependent prior incorporates minimal information, but it helps avoid identification and MCMC convergence issues associated with Heywood problems. Appendix A.2 presents a sensitivity analysis confirming that the results obtained using the data-dependent prior are practically identical to those obtained using several data-independent priors.

Finally, large variance normal priors are assigned on the $\boldsymbol{\alpha}$ parameters. Specifically, we use the following wide prior normal in every analysis that follows: $\boldsymbol{\alpha} \sim N(0, 10^2)$.

2.2.2 | Models and priors for binary data

The model for binary data using the underlying variables $y_{ij}^*, (j = 1 \ldots, p)$ is

$$\begin{cases} y_{ij} = \mathcal{I}(y_{ij}^* > 0), \\ \mathbf{y}_i^* = \boldsymbol{\alpha} + \mathbf{\Lambda}\mathbf{z}_i + \mathbf{u}_i + \mathbf{e}_i, \\ \mathbf{e}_i \sim \prod_{j=1}^{p} \text{Logistic}(0, \pi^2/3) \quad \text{or} \quad \prod_{j=1}^{p} N(0,1) \\ \mathbf{z}_i \sim N(0, \mathbf{\Phi}) \\ \mathbf{u}_i \sim N(0, \mathbf{\Omega}). \end{cases}$$

In the model specification described above, the $\mathbf{e}_i$ terms correspond to the logistic and probit specifications, which are the most frequently used models. However, other distribution choices are also possible. The expressions provided above can be simplified by integrating out the $\mathbf{e}_i$ terms to obtain

$$\begin{cases} \mathbf{y}_i \sim \prod_{j=1}^{p} \text{Bernoulli}\left(\pi_{ij}(\eta_{ij})\right) \\ \pi_{ij}(\eta_{ij}) = \sigma(\eta_{ij}) \text{ or } \pi_{ij}(\eta_{ij}) = F(\eta_{ij}), \ \eta_{ij} = [\boldsymbol{\eta}_i]_j \\ \boldsymbol{\eta}_i := \boldsymbol{\alpha} + \boldsymbol{\Lambda}\mathbf{z}_i + \mathbf{u}_i, \\ \mathbf{z}_i \sim N(0, \boldsymbol{\Phi}) \\ \mathbf{u}_i \sim N(0, \boldsymbol{\Omega}) \end{cases} \tag{11}$$

where $\sigma(x) = [1 + \exp(-x)]^{-1}$ denotes the sigmoid function that leads to the logit model, whereas $F(\cdot)$ denotes the cumulative distribution function of the standard normal distribution that leads to the probit model. It is important to note that under the framework, the distribution of $\mathbf{u}_i$ and even $\mathbf{z}_i$ need not be normal; this assumption was made only for exposition purposes. When assuming that the $\mathbf{u}_i$ terms are indeed normal, the amount of data augmentation can be further reduced by using the following equivalent formulation:

$$\begin{cases} \mathbf{y}_i \sim \prod_{j=1}^{p} \text{Bernoulli}\left(\pi_{ij}(\eta_{ij})\right), \\ \pi_{ij}(\eta_{ij}) = \sigma(\eta_{ij}) \text{ or } \pi_{ij}(\eta_{ij}) = F(\eta_{ij}), \\ \boldsymbol{\eta}_i \sim N(\boldsymbol{\alpha}, \ \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Omega}). \end{cases} \tag{12}$$

In the simulation experiment and real-world examples, the formulations of (11) and (12) were used as they are more convenient in the context of MCMC for models based on the logit link.

In terms of interpretation, it is interesting to note that the proposed model extends the two-parameter logistic item response theory (IRT) model by allowing for an item-individual random effect in addition to the standard individual latent variable $\mathbf{z}_i$. The probability of a correct response to item j by individual i can be written

$$\frac{1}{1 + \exp\left(-\left[\boldsymbol{\alpha} + \boldsymbol{\Lambda}\mathbf{z}_i\right]_j - u_{ij}\right)}.$$

Similar priors can be assigned as in the case of continuous data. Regarding the elements of the $\boldsymbol{\Lambda}$ matrix that are not approximate zero, unit information priors can be used. In the case of the two-parameter IRT (one-factor) model, this translates to a $N(0, 4)$ prior Vitoratou et al. (2014).

2.3 | Overview of models and their estimation

In this section, our focus is on four models that play a crucial role in the model assessment methodology developed within this paper. These models have been defined within the framework established thus far, and we will now delve into the specifics of their implementation, providing comprehensive details and engaging in insightful discussions. The four models under consideration are as follows:

- Exact zero (EZ) model. This is a CFA model and serves as the initial reference point for our analysis. It is defined by Equations (1) and (4), with all cross-loadings in $\boldsymbol{\Lambda}$ set to zero.
- Approximate zero (AZ) model: This model was initially introduced by Muthén and Asparouhov (2012) and is further extended in this paper. In this extension, an item-individual random effect is introduced

in the linear predictor, allowing for the modelling of categorical data using link functions other than probit. Moreover, this model has the potential to detect outliers (Model (6)). When assuming normal distributions for $\mathbf{u}_i$, $\mathbf{e}_i$, and $\mathbf{z}_i$, the model simplifies to (7). One crucial aspect of the AZ model is that the cross-loadings in $\mathbf{\Lambda}$ are no longer fixed at zero. It is important to note that this model should only be used in the Bayesian sense, as informative priors on $\mathbf{\Omega}$ and the cross-loadings in $\mathbf{\Lambda}$ are essential for estimation.

- Exploratory factor analysis model (EFA): This model represents the standard EFA approach and is defined by Equations (1) and (4). In this model, low informative priors are assigned to all the components of $\mathbf{\Lambda}$, and $\mathbf{\Phi}$ is set as the identity matrix, $\mathbf{\Phi} = I$.
- EFA model with item-individual random effects (EFA-C): This model follows the same structure as the EFA model, but instead of using Equation (4), it uses Equation (6). By incorporating item dependencies conditional on the extracted independent factors, this approach allows for greater dimension reduction compared to traditional EFA models. The stricter assumption of conditional independence could necessitate the inclusion of additional factors in the model.

In terms of implementation, various MCMC schemes can be employed, as discussed for example in Edwards (2010). When the $\mathbf{e}_i$, $\mathbf{u}_i$, and $\mathbf{z}_i$ are assumed to follow a normal distribution, Gibbs samplers may be formed, as described in Geweke and Zhou (1996) and Chib and Greenberg (1998). For more general models, Pólya-Gamma augmentation was proposed by Polson et al. (2013), Jiang and Templin (2019) and Asparouhov and Muthén (2021b) also provide a framework for constructing Gibbs samplers.

In this paper, we recommend the use of Hamiltonian Monte Carlo (HMC; Neal , 2011) as it covers all the models discussed and allows for flexibility in choosing priors. HMC can be implemented using programming frameworks such as Stan (Carpenter et al., 2017), which are well suited for assessing the convergence properties of MCMC schemes. This is an essential task when working with the models proposed in this paper. Higher-level software like blavaan Merkle and Rosseel (2015) also support HMC. For comprehensive code and implementation details, we refer readers to the code repository for this work hosted on github at 'bayes-sem'.

Fitting the EZ model in Stan is generally straightforward, although different parametrisations can be considered to improve MCMC performance and stability. For instance, one option is to set one loading of each factor in $\mathbf{\Lambda}$ to one and use a full covariance matrix for $\mathbf{\Phi}$. Alternatively, the leading cross-loadings can be constrained to be positive, and a correlation matrix can be employed for $\mathbf{\Phi}$.

To summarise, when the EZ model does not perform well, it is crucial to find an appropriate benchmark to assess the performance of the AZ model. As discussed in more detail in the next section, the EFA and EFA-C models can serve as benchmarks in such cases. However, fitting EFA models using MCMC can be challenging due to issues like rotational indeterminacy. The challenge arises because the likelihood is typically specified in terms of $\mathbf{\Lambda}\mathbf{\Lambda}^T$, while the interest often lies in $\mathbf{\Lambda}$ itself. The lower triangular set of restrictions (see e.g. Geweke & Zhou, 1996) ensures a well-defined mapping between these matrices but introduces order dependencies among the observed variables. The choice of the first q variables becomes influential (Carvalho et al., 2008). Alternative schemes proposed by Conti et al. (2014) and Frühwirth-Schnatter & Lopes (2018) and Bhattacharya and Dunson (2011) provide ways to overcome these restrictions and can be used to identify the number of factors within a single MCMC run. However, as noted by Bhattacharya and Dunson (2011), when tasks such as choosing the number of factors or assessing the predictive performance are the focus, monitoring $\mathbf{\Lambda}\mathbf{\Lambda}^T$ can be sufficient and avoids rotational issues. In such cases, restrictions on $\mathbf{\Lambda}$ can be omitted, as long as there are no convergence and mixing issues in the MCMC samples of $\mathbf{\Lambda}\mathbf{\Lambda}^T$. In this paper, the EFA and EFA-C models are only used to establish benchmarks for predictive performance. Therefore, focusing on $\mathbf{\Lambda}\mathbf{\Lambda}^T$ is sufficient for the purposes of this study.

3 | MODEL ASSESSMENT

In this section, we present a model assessment framework that integrates fit indices and cross-validation to identify overfitting. The objective is to enhance the evaluation of model performance by combining PPP values (or similar indices) with scoring rules. This approach aims to achieve a good fit while avoiding overfitting. The suggested procedure involves computing these metrics for various models, including the EZ and AZ models, as well as the EFA and EFA-C models, with an equal number of factors. We will provide a detailed explanation of the proposed indices and subsequently outline our suggested procedure, along with guidelines and recommendations.

3.1 | Assessing goodness of fit with PPP values

Posterior predictive p-values are commonly used to assess model fit in Bayesian SEM. This approach utilizes a discrepancy function, denoted by $D(\boldsymbol{Y}, \boldsymbol{\theta})$, which quantifies the deviation between the fitted model and the observed data. For continuous data, where the model is specified using Equations (5) and (7), the likelihood ratio test (LRT) function is typically employed as the discrepancy function (e.g. Scheines et al., 1999). The LRT function compares the estimated model (denoted as the H_0 hypothesis) to an unconstrained variance-covariance matrix model (denoted as the H_1 hypothesis), also known as the saturated model (perfect fit). The expression for $D(\boldsymbol{Y}, \boldsymbol{\theta})$ is as follows:

$$\mathrm{LR}\left[S, \Sigma(\theta)\right] = (n-1)\left\{\log\left|\Sigma(\theta)\right| + \mathrm{tr}\left[S\Sigma^{-1}(\theta)\right] - \log|S| - p\right\},$$

where S and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ are the sample and model-implied variance-covariance matrix respectively, and $\boldsymbol{\theta}$ represents population-based parameters such as $\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Omega}$, and $\boldsymbol{\Psi}$. Furthermore, $|\cdot|$ and $\mathrm{tr}(\cdot)$ denote the determinant and trace of a matrix respectively. If the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is used in (14), then $\mathrm{LR}[\cdot]$ becomes a statistic. However, if $\boldsymbol{\theta}$ is unknown, then $\mathrm{LR}[\cdot]$ can be viewed as a metric. Given the discrepancy function $D(\boldsymbol{Y}, \boldsymbol{\theta}_m)$ defined in (14), an appropriate MCMC algorithm, and M posterior draws, the PPP value can be computed as follows:

1. At each (or some) of the MCMC samples $\boldsymbol{\theta}_m$, $m = 1, \ldots, M$, do the following:
   a. Compute $D(\boldsymbol{Y}, \boldsymbol{\theta}_m)$.
   b. Draw $\tilde{Y}$ having the same size as $\boldsymbol{Y}$, from the likelihood function $f(\boldsymbol{Y}|\boldsymbol{\theta}_m)$ of the implied model in Equation (5) or (7) and using the current value $\boldsymbol{\theta}_m$.
   c. Calculate $D(\tilde{Y}, \boldsymbol{\theta}_m)$ and $d_m = \mathcal{I}\left[D(\boldsymbol{Y}, \boldsymbol{\theta}_m) < D(\tilde{Y}, \boldsymbol{\theta}_m)\right]$, where $\mathcal{I}[\cdot]$ is an indicator function.
2. Return PPP$= \frac{1}{M}\sum_{m=1}^{M} d_m$.

For binary data, we use a different formulation of the model than the one given in Equations (11) and (12) in order to define a suitable metric for PPP values based on the probabilities of each response pattern. Let us assume we have p binary items, resulting in $R = 2^p$ possible response patterns denoted by $\{\mathbf{y}_r\}_{r=1}^{R}$, with corresponding observed frequencies denoted by O_r, where $r = 1, \ldots, R$. The probability of a response pattern, based on the logistic model with $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Omega})$, and the assumption of conditional independence given $\boldsymbol{\chi}$ and $\boldsymbol{u}$, is given by

$$\pi_r(\theta) = \int \prod_{j=1}^{p} \mathrm{Bernoulli}\left\{[\mathbf{y}_r]_j | \sigma([\boldsymbol{\eta}]_j)\right\} f(\mathbf{z}) f(\mathbf{u}) d\mathbf{z} d\mathbf{u},$$

where $\mathrm{Bernoulli}\left\{[\mathbf{y}_r]_j | \sigma([\boldsymbol{\eta}]_j)\right\}$ denotes the Bernoulli probability mass function for $[\mathbf{y}_r]_j$, the jth item component of the response pattern r, with success probability $\sigma([\boldsymbol{\eta}]_j)$. $\boldsymbol{\eta}$ is as defined earlier, i.e. $\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\mathbf{z} + \mathbf{u}$.

The latent component **z** and **u** are integrated out, e.g. using Monte Carlo. Marginalising out the latent components reduces the Monte Carlo error and is considered good practice in such settings, as demonstrated in Merkle et al. (2019) for calculating information criteria. In the context of PPP values, a commonly used discrepancy measure, (e.g. Sinharay, 2005), is the G^2 statistic given by

$$D(\boldsymbol{Y}, \theta) = \sum_{r=1}^{R} O_r \log\left(\frac{O_r}{n\pi_r(\theta)}\right).$$

It is important to note that in the presence of observed frequencies O_r and the response pattern probabilities $\pi_r(\theta)$ obtained from Equation (15), an alternative formulation of the model can be based on the multinomial distribution:

$$\left(O_1, \ldots, O_R\right) \sim \text{Multinomial}\left[n, \pi_1(\theta), \ldots, \pi_R(\theta)\right].$$

For a given θ, such as a sample draw from the posterior distribution, (16) can be interpreted as the likelihood ratio between the model in (17) and the saturated version of it, where each $\pi_r(\theta)$ is a separate unknown parameter that can be estimated by O_r/n. Given a set of MCMC samples from the posterior distribution, the PPP value can be calculated following the same steps outlined earlier for the case of continuous data.

It is important to note that PPP values are not equivalent to traditional p-values and are not directly linked to the concept of type I error. Instead, they are considered fit indices that provide information about the goodness of fit of a model. In terms of interpreting PPP values, the discussion in Muthén and Asparouhov (2012) is followed. A model with a PPP value around .5 is generally regarded as an excellent fit. It is less clear how low a PPP value should be to indicate poor fit, but thresholds of .1 and .05 are commonly used.

The discrepancy function used in this framework assesses the overall fit of the model. However, alternative discrepancy functions can be employed to evaluate the fit on lower-order margins. For categorical data, one approach is to compute chi-squared-type residuals on univariate, bivariate, and trivariate margins as described in Jöreskog and Moustaki (2001). Additionally, the work on limited information test statistics such as the M_2 test statistic proposed by Maydeu-Olivares and Joe (2005) can be used. These discrepancy measures can be valuable in detecting model misfit in pairs or triples of items. Exploring these discrepancies within the proposed framework could be a subject of future research.

3.2 | Scoring rules in SEM via cross-validation

As previously stated, evaluating the out-of-sample predictive performance of each model is crucial to complement its fit assessment. Although factor analysis models are not primarily designed for prediction, the concept of predictive inference can be useful in evaluating model fit and detecting overfitting. The focus is on assessing a model's ability to predict new data that were not utilised for parameter estimation. To achieve this, we divide individuals into two distinct groups: (i) the training sample, denoted by $\boldsymbol{Y}^{\text{tr}}$, is used to estimate the model parameters through the posterior distribution $\pi(\theta|\boldsymbol{Y}^{\text{tr}})$, and (ii) the test sample, denoted by $\boldsymbol{Y}^{\text{te}}$, is employed to assess the accuracy of the previously estimated model's forecasts.

More specifically, the predictions for the unseen data are represented by a distribution denoted by $b(\boldsymbol{Y}^{\text{te}}|\boldsymbol{Y}^{\text{tr}})$, which can be compared as a whole to the actual test data $\boldsymbol{Y}^{\text{te}}$. In the frequentist case, one option for such a predictive distribution is $f(\boldsymbol{Y}^{\text{te}}|\widehat{\theta}^{\text{tr}})$, where $f(\cdot)$ denotes the likelihood function, and $\widehat{\theta}^{\text{tr}}$ is the MLE obtained from $\boldsymbol{Y}^{\text{tr}}$. In the Bayesian framework, the standard choice is the posterior predictive distribution:

$$f(\boldsymbol{Y}^{\text{te}}|\boldsymbol{Y}^{\text{tr}}) = \int f(\boldsymbol{Y}^{\text{te}}|\theta)\pi(\theta|\boldsymbol{Y}^{\text{tr}})d\theta.$$

To assess the quality of these distributions, scoring rules can be employed, as indices with small values typically indicate good performance (e.g. Dawid & Musio, 2014; Gneiting & Raftery, 2007). One commonly used scoring rule is the log score, which assesses the quality of a predictive distribution $h(\boldsymbol{Y}^{\text{te}})$ and is defined as

$$LS(\boldsymbol{Y}^{\text{te}}) = -\log h(\boldsymbol{Y}^{\text{te}}).$$

The log score belongs to a class of scoring rules that possess the desired property of being strictly proper. Strict propriety ensures that the optimal model among the ones being considered will be uniquely identified. Specifically, the score of this optimal model will be strictly lower than the scores of the other models. If it is smaller than or equal to those scores, we have a proper scoring rule instead of a strictly proper one.

The log score can be viewed as a natural extension of the goodness-of-fit criterion based on the likelihood ratio test statistic for prediction assessment. Let's consider a structural equation model, defined by (1) or (6) and (4), and suppose we want to compare it against the saturated model. For instance, in the case of continuous data, the saturated model is denoted by $\boldsymbol{Y}^{\text{te}} \sim N(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ represents an unconstrained variance-covariance matrix. Using $f^{\text{SEM}}(\cdot)$ and $f^{S}(\cdot)$ to denote the density functions of the SEM and saturated models respectively, the difference between the two log scores can be expressed as

$$-\log\left[\frac{f^{\text{SEM}}(\boldsymbol{Y}^{\text{te}}|\widehat{\theta}^{\text{tr}})}{f^{S}(\boldsymbol{Y}^{\text{te}}|\widehat{\alpha}^{\text{tr}}, \widehat{\Sigma}^{\text{tr}})}\right].$$

The preceding expression may be viewed as the likelihood ratio test statistic based on point parameter estimates from the training data $\boldsymbol{Y}^{\text{tr}}$, but evaluated on the unseen test data $\boldsymbol{Y}^{\text{te}}$.

Note that in (20), the predictive distributions do not account for the uncertainty in the parameter estimates, which can be substantial for small training sample sizes. The Bayesian framework accounts for this source of uncertainty in a natural way via the posterior predictive distribution (18).

In many cases, including most of the Bayesian models discussed in this paper, the log score and the posterior are not available in closed form. Instead, we rely on samples from the predictive distribution to approximate it. A commonly used approach is the mixtures-of-parameters (MP) approximation, which has been employed in various studies (see table 1 of the supplementary material in Krüger et al., 2021).

For the MP approximation for a test data point Y_i^{te}, we use a Monte Carlo approximation on the conditional predictive density $f(Y_i^{\text{te}}|\boldsymbol{\Theta})$, which is required in closed form, given samples $\boldsymbol{\Theta}_{m=1}^{M}$ from the posterior $\boldsymbol{\pi}(\boldsymbol{\Theta}|\boldsymbol{Y}^{\text{tr}})$

$$LS(Y_i^{\text{te}}) = -\log\int f(Y_i^{\text{te}}|\boldsymbol{\theta})\boldsymbol{\pi}(\boldsymbol{\theta}|\boldsymbol{Y}^{\text{tr}})d\boldsymbol{\theta} \approx -\log\left\{\frac{1}{M}\sum_{m=1}^{M} f(Y_i^{\text{te}}|\boldsymbol{\theta}_m)\right\}.$$

For continuous data, $f(Y_i^{\text{te}}|\boldsymbol{\theta}_m)$ is typically a normal probability density function, as illustrated in Section 2; see for example Equations (5) and (7).

For binary data, it is possible to compute the log score via the alternative model formulation based on frequency patterns given in (17). The model in this case can be defined based on the observed frequencies $(O_1, \ldots, O_R)$ and the model-based $\boldsymbol{\pi}_r(\boldsymbol{\theta})$, obtained from Equation (15), via the multinomial distribution.

We can therefore write

$$\text{LS}(\mathbf{O}^{\text{te}}) = -\log f(\mathbf{O}^{\text{te}}|\boldsymbol{\pi}^{\text{tr}}) = -\log\left[c\prod_{r=1}^{R}\left[\boldsymbol{\pi}_r(\boldsymbol{\theta})^{\text{tr}}\right]^{O_r^{\text{te}}}\right] = -\sum_{r=1}^{R} O_r^{\text{te}}\log\boldsymbol{\pi}_r(\boldsymbol{\theta})^{\text{tr}} + c,$$

where c represents a constant. It is important to note that the log score and G^2 differ only by a constant when used as metrics. This observation essentially confirms the argument made earlier in Equation (20) regarding the connection of the likelihood ratio test and the log score.

Until now, we have assumed a single split between the training and test data. However, to mitigate the impact of specific data splits, cross-validation can be utilised. It is worth mentioning that both the calculation of PPP values and scoring rules are based on the posterior predictive distribution. However, there is a fundamental distinction between these two approaches. PPP values rely on the posterior distribution conditioned on the entire dataset, and the prediction is made using the same dataset. On the other hand, the scoring rules approach conditions the posterior on a subset of the data (training sample) and makes predictions using the complementary set (test sample).

3.3 | Model assessment of fit and predictive performance indices

Our procedure consists of two main components: assessing goodness of fit, as is commonly done in current practice, and evaluating out-of-sample predictive performance. To assess goodness of fit, we follow the well-established procedure of comparing the fit of the hypothesised model (in our framework, the EZ model) to that of the unconstrained model or saturated model. This comparison can be achieved by examining the PPP values of the EZ model. If the PPP value is satisfactory, we recommend processing with the hypothesised model without further investigation.

However, if the PPP value of the EZ model is unsatisfactory while the AZ model's PPP value is satisfactory, the researcher should not hastily support the hypothesised model. This is because the satisfactory PPP value of the AZ model could be the result of overfitting the data. In the context of SEM, overfitting is defined as follows. If the AZ model demonstrates better goodness of fit compared to its EZ counterpart but performs poorly in terms of out-of-sample predictive performance, then it is considered to have overfit the data. In other words, if the improved goodness of fit of the AZ model does not capture systematic patterns in the data, it would be of limited use in predicting unseen data. Additionally, the slight increase in model complexity of the AZ model may negatively impact prediction compared to the corresponding EZ model. Therefore, our proposed framework emphasises parsimony alongside goodness of fit. As per our framework, if the AZ model performs worse than the EZ model in terms of the relevant scoring rule, then there is little support for the hypothesised model in the data.

Now let us consider the scenario where the AZ model outperforms the corresponding EZ model in terms of both PPP value and predictive performance, as measured by the relevant scoring rule. In this case, we advise caution and recommend conducting further checks. It is possible that the AZ model is merely compensating for a poorly specified EZ model, but there may be other EZ or AZ models that predict even better. If the poor fit of the EZ model is primarily due to minor cross-loadings or error correlations, an AZ model that captures these parameters should exhibit good performance. However, if the EZ model lacks other systematic patterns that are captured by the AZ model, the improvement offered by the AZ model may be limited. Ultimately, the key question we aim to answer is whether the predictive performance of each model is sufficiently good. Therefore, it is crucial to establish a benchmark when comparing predictive performances. While the performance of the unconstrained saturated model is commonly used as benchmark for assessing goodness of fit, it may not be suitable for assessing predictive performance (MacCallum et al., 1992). The saturated model has substantially higher complexity (larger number of parameters) than the hypothesised models. When two models with different numbers of parameters exhibit similar in-sample performance, the model with fewer parameters typically performs better out of sample. As an alternative benchmark model in this paper, we consider the EFA model with the same number of factors as the hypothesised model, provided that it fits the data well. This EFA model has fewer parameters than the saturated model and is expected to perform well in terms of predictive performance because it can explore systematic patterns in the data without any restrictions, other than having a specified number of factors q. However, the EZ and AZ models have explicit restrictions and are often expected to be relatively close to the EFA model. Therefore, to deem

the predictive performance of the hypothesised model satisfactory, its scoring rule should be comparable to that of the EFA model selected as the benchmark.

It is important to exercise caution when choosing the benchmark EFA model to ensure that it provides a suitable standard for predictive performance. Choosing an over- or under-parametrised model could lead to an inappropriate benchmark and potentially underestimate or overestimate the expected predictive performance. One approach to determining the appropriate number of factors for the benchmark EFA model is to consider models with varying numbers of factors, as long as they fit the data well. To select the number of factors, models with different numbers of factors can be fit separately and their fit to the data compared. This comparison can be done using standard indices, such as model evidence, Bayesian information criterion (BIC), or other appropriate goodness-of-fit measures. Additionally, their predictive performance, as described earlier, can be evaluated using scoring rules or other relevant metrics.

Considering these factors, it is important to note that the presence of small error correlations induced by the $\mathbf{u}_i$ in the approximate zero framework can potentially benefit CFA models in terms of predictive performance. These small correlations can be interpreted as an additional minor factor. To ensure a fair and meaningful comparison between CFA and EFA models, it is reasonable to address this advantage by incorporating small error correlations in both model types. This can be achieved through the use of the exploratory factor analysis with correlated errors model (EFA-C).

Note also that, while the AZ models are more flexible than their EZ counterparts, they can still perform poorly in cases of substantial model misspecification. For instance, if there are large cross-loadings, greater than than .5, using the normal$(0, 0.01)$ prior can still result in poor performance compared to the EFA model. This comparison can therefore be used to detect misspecified models, as we illustrate in Sections 4 and 5.

We summarise the recommendations of our proposed framework in what follows and in Figure 1.

1. If the EZ model has satisfactory fit indices such as PPP values, this represents strong support for the hypothesised model.
2. If both the EZ and AZ models have poor PPP values, then there is little support for the hypothesised model. In such cases, it may be useful to use more vague priors to explore the weaknesses of the model. It would be expected in this case that the EFA models will have better predictive performance; otherwise, there may be issues in the fitting algorithms or elsewhere.
3. If the EZ model has poor performance in terms of fit indices, whereas the AZ model is satisfactory, it is essential to check the scoring rules. If the improvement offered by the AZ model is due to overfitting, it is expected that the prediction score for the AZ model will be inferior to that of the EZ one.

The predictive performance of models that overfit is therefore expected to diminish. On the other hand, a prediction score that still favours the AZ model suggests that overfitting did not occur. To check whether the predictive performance of the AZ model is good enough, comparisons with EFA-type models can be made. In cases of comparable or improved performance, there is supporting evidence for the hypothesised model.

Assessing model fit in factor analysis can indeed be challenging, considering factors such as misspecification of the latent variable distribution, item dependencies, skewed data, and non-linear predictors. In this paper, we used a discrepancy function to calculate PPP values, which provide an evaluation of the overall fit of the model for both continuous and binary data.

To complement these overall goodness-of-fit tests, one can also consider measures of fit such as residuals and limited information test statistics that detect item misfit on lower order margins, as explained in Section 3.1.

Our proposed methodology aims to tackle some of the aforementioned challenges by focusing on out-of-sample prediction performance of the model within the Bayesian modelling framework for structural equation models. This approach introduces new tools and sheds light on the limitation of relying solely on PPP values.

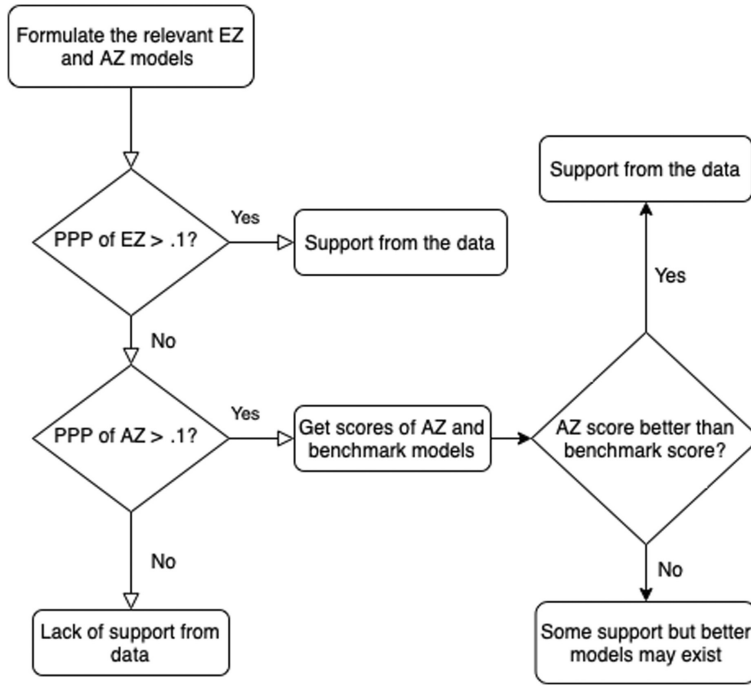

FIGURE 1 Diagram illustrating proposed model assessment procedure. The first step involves fitting the EZ model to the data and assessing its fit using conventional methods such as posterior predictive *p*-values (PPP). If the EZ model shows poor fit, the next step is to fit the AZ model and assess its fit using PPP. If the PPP of the AZ model indicates poor fit, this suggests that the data do not support the AZ model. However, if the PPP of the AZ model is not problematic, the next step is to compare the predictive performance of the AZ model with that of a benchmark EFA model using cross-validated scoring rules. If the comparison does not favour the AZ model, this suggests that there may be better models available. It is important to note that alternative methods, such as using different thresholds for PPP values or alternative fit indices, can also be incorporated into the procedure.

4 | SIMULATION EXPERIMENTS

4.1 | Setup

Simulation experiments were conducted to study the performance of the proposed models and demonstrate the assessment framework for both continuous and binary data. We focused on two cases of data generated using Equation (4): continuous and binary. For each of these two cases, three scenarios were considered when generating simulated data:

- Scenario 1: Data generated from the EZ model;
- Scenario 2: Data generated from the AZ model with small error correlations, introduced by item-individual random effects, and without cross-loadings;
- Scenario 3: Data generated from the AZ model with two non-negligible cross-loadings and without small error correlations.

The purpose of these scenarios is to illustrate three distinct situations regarding the level of support provided by the data for the hypothesised theory and to showcase the behaviour of the proposed model assessment procedure. In scenario 1, the observations are generated from the EZ model, which aligns with the hypothesised theory. The expected outcome is to indicate support from the data, confirming the alignment between the observed data and the theoretical model. In scenario 2, the presence of

small error correlations is introduced, but it should ideally have minimal impact on the hypothesised theory. Therefore, the desired outcome in this case is again an indication of support from the data. In scenario 3, significant cross-loadings are incorporated, which should indicate a lack of support from the data for the hypothesised theory.

For both continuous and binary data, we considered $p = 6$ items and $q = 2$ factors. The factor loadings used to generate the data, in each of the three scenarios, are shown in Table 1.

In the case of continuous data, the correlation between variable y_j and the hth latent variable z_h ($h = 1, \ldots, q$) can be used as a measure of the standardised effect size for the cross-loadings. For example, in scenarios 1 and 2, the cross-loadings that are equal to 1 and .8 correspond to standardised effect sizes of .71 and .64 respectively, whereas in scenario 3, the cross-loadings $\lambda_{32} = 0.6$ and $\lambda_{41} = 0.6$ correspond to standardised effect sizes of .51 and .49 respectively. For binary data, the correlation between the underlying variable y_j^* and the hth latent variable z_h can be used as the standardised effect sizes for the cross-loadings. In scenarios 1 and 2, the cross-loadings that are equal to 1 and .8 correspond to standardised effect sizes of .48 and .4 respectively under the logit model, whereas in scenario 3, the standardised effect for the cross-loadings $\lambda_{32} = 0.6$ and $\lambda_{41} = 0.6$ are both .36.

Although the data were generated under the three scenarios, the hypothesised model assumes a simple structure in which the first three items load on the first factor and the last three items load on the second factor. In other words, for the AZ model, the first three elements of the first $\boldsymbol{\Lambda}$ column and the last three of the second $\boldsymbol{\Lambda}$ column are regarded as the major parameters, whereas the other elements of $\boldsymbol{\Lambda}$ are cross-loadings. In all three scenarios, the factor correlation was set to .2, and the intercepts $\boldsymbol{\alpha}$ were all zero. For identification purposes, when estimating the models for continuous data, one loading of each factor is set to 1 and the factor variances are free parameters to be estimated. In the models for binary data, only sign restrictions are placed on one of the loadings for each factor and the corresponding factor variances are set to 1. Both choices are made to improve MCMC performance. The sample size was set to $n = 1000$ for the continuous data and $n = 2000$ for the binary data. For scenario 2, Equation (7) was used in the case of continuous data by setting the matrix $\boldsymbol{\Omega} + \boldsymbol{\Psi}$ to have ones in the diagonal and randomly drawn correlations from the normal distribution with zero mean and a standard deviation of .2, while checking also for positive definiteness. In the case of binary data, the matrix $\boldsymbol{\Omega}$ was drawn from its informative prior.

The models and priors were specified as described in Section 2, and samples from the posterior of each model were obtained using Hamiltonian MCMC programmed in Stan. For continuous data, we conducted a warm-up period of 1000 iterations followed by 2000 iterations for inference purposes. In the case of binary data, we performed 2000 iterations for warm-up and an additional 2000 per analysis. The models were run in 4 parallel chains, resulting in a total of $4 \times 2000 = 8000$ posterior draws. These settings were chosen based on our experience and were found to provide reasonable convergence and mixing properties, as confirmed by monitoring the R-hat diagnostics, (see e.g. Vehtari et al., 2021), and visually inspecting the posterior draws. If necessary, further improvements can be achieved by exploring additional options in Stan during the warm-up period. In all cases, including the real-world examples, we employed a K-fold cross-validation and aggregated the scores by summation. The size of

TABLE 1 True factor loadings used in the three simulation scenarios.

Scenario 1		Scenario 2		Scenario 3	
z_1	z_2	z_1	z_2	z_1	z_2
1	0	1	0	1	0
.8	0	.8	0	.8	0
.8	0	.8	0	.8	.6
0	1	0	1	.6	1
0	.8	0	.8	0	.8
0	.8	0	.8	0	.8

the simulated and real data used in this paper is sufficient, allowing for relatively fast computation with small values of K, while still ensuring adequately sized training and test samples. Therefore, we used $K = 3$ throughout this paper. However, it is generally recommended to explore different values of K to cross-validating the scoring rules values. As scoring rules are comparative indices, we reported the score differences between each model and the best-performing model. Specifically, we assigned a value of zero to the best model in each scenario, or the model with the smallest score.

In each scenario, we applied the proposed model assessment framework outlined in Section 3 by computing the PPP values and scoring rules for all models mentioned earlier. We then fitted and summarised these models according to the procedures described in Sections 3.1 and 3.2 and followed the recommendations outlined in Section 3.3. The following two sections provide the results of the simulation experiments conducted for continuous and binary data respectively. The aim of these experiments was to demonstrate the effectiveness of the proposed model framework and serve as a proof of concept. It is important to note that more extensive simulation studies are needed, as discussed in the subsequent sections, and are left for future research. Finally, Appendix A3 presents a simulation experiment that examines the ability of the AZ model to recover parameters when analysing binary data.

4.2 | Continuous data

Table 2 gives the log score (LS) and the PPP values for the three simulation scenarios.

In scenario 1, all models show good fit to the data based on the PPP values. However, when considering predictive performance, the EZ model outperforms the other models. This result is expected since the data were generated from the EZ model. Interestingly, the EZ model performs even better than the EFA models in terms of predictive performance because it is a more parsimonious model.

In simulation scenario 2, the EZ and EFA models exhibit poor fit according to their PPP values, which is expected since these models assume zero error correlations. However, the AZ and EFA-C models, which allow for small non-zero error correlations, fit the data well. The question arises whether the improved fit of the AZ model is due to fitting noise or overfitting, as defined in recommendation 2 of Section 3.3. If the AZ model were overfitting the data, we would not expect to see improved performance over the EZ model, as we see here. To investigate further, we can compare the predictive performance of the AZ model against other models using recommendation 3 of Section 3.3. The AZ model shows competitive predictive performance compared to the EFA models, with a log score similar to that of EFA-C. This provides strong support for the AZ model and, consequently, the hypothesised model. This conclusion is reasonable in the SEM context since poor fit is due to error correlations that are usually linked with observation error rather than factor loading misspecifications. PPP values alone would not have been sufficient to reach this conclusion.

In scenario 3, the EZ model exhibits poor fit to the data, which aligns with our expectations. However, all other models have PPP values around .5, indicating reasonable fit. The key question is whether the AZ model overfits the data and what conclusions we can draw about the hypothesised

TABLE 2 Simulation results for continuous data.

	Scenario 1		Scenario 2		Scenario 3	
Model	**PPP**	**LS**	**PPP**	**LS**	**PPP**	**LS**
EZ	.66	0	.03	–	.00	–
AZ	.51	2.42	.47	0	.53	1.37
EFA	.62	1.14	.17	4.47	.59	0
EFA-C	.53	2.49	.48	.34	.56	1.02

Note: PPP values and sum of log scores (LS) obtained from three-fold cross-validation for the relevant models. In each scenario, the best-performing model is assigned a log score of 0, and the differences from this model are reported for the other models. The log scores are not reported when PPP values are <.1.

model. To address these questions, we compare the AZ model to the benchmark EFA model, which in this case is the EFA model with higher predictive performance, as the data were simulated without error correlations. However, the AZ model does not outperform the EFA model, suggesting that an alternative hypothesised theory regarding the loading structure of the six items may be more appropriate. In fact, the theory corresponding to scenario 3 in Table 1 provides a superior model because the data were generated from it.

4.3 | Binary data

In this section, we summarise the results of the three simulation experiments for binary data. Table 3 gives the PPP values and the log scores.

The results for binary data are very similar to those for continuous data. In scenario 1, all models show good fit, as indicated by the PPP values. Additionally, the EZ model demonstrates optimal predictive performance, which is expected since the data were generated from this model. In scenario 2, the EZ model shows very poor fit caused by the additional error correlations in the simulated data, as indicated by the PPP value of .02. The other models demonstrate moderately good fit, with PPP values above .10. Similar to the continuous case, the AZ model performs well in terms of both recommendations 2 and 3 of Section 3.3, exhibiting the best predictive performance. Moving on to scenario 3, the EZ model also fails to fit the data well due to the presence of non-zero cross-loadings. On the other hand, the other models perform well, but questions arise about the validity of the hypothesised theory. Hence, we apply recommendation 3 of Section 3.3, comparing the predictive performance of the AZ model against the best-performing EFA model. In this scenario, the AZ model does not outperform the EFA model in terms of predictive performance.

4.4 | Model performance assessment

Having demonstrated the proposed model assessment framework in the previous subsections, we further examine its performance by considering the sampling variability in the data. To do this, we simulate 100 datasets for each scenario described in Section 4.1 and explore variations of the proposed framework by considering different benchmark options and existing model assessment alternatives. It is important to note that our goal is to provide a proof of concept rather than an extensive simulation, which would require significantly more effort. Therefore, we focus our attention on continuous data, as the computational time required to fit a model for binary data is considerably higher.

The standard approach to model assessment in the Bayesian context involves fitting only the EZ model and calculating a suitable goodness-of-fit index, such as the PPP value. This index is then used to either support or reject the model, a method commonly known as the 'EZ-only' approach. However, as mentioned earlier in this paper and in the literature (see e.g. Muthén & Asparouhov, 2012), this approach

TABLE 3 Simulation results for binary data.

	Scenario 1		Scenario 2		Scenario 3	
Model	**PPP**	**LS**	**PPP**	**LS**	**PPP**	**LS**
EZ	.52	0	.02	–	.00	–
AZ	.50	.68	.12	0	.52	1.90
EFA	.59	1.45	.13	.09	.45	0
EFA-C	.54	3.27	.17	.24	.50	2.96

Note: PPP values and sum of log scores (LS) obtained from three-fold cross-validation for the relevant models. In each scenario, the best-performing model is assigned a log score of 0, and the differences from this model are reported for the other models. The log scores are not reported when PPP values are <.1.

tends to over-reject models. Therefore, we propose an alternative approach that utilises the AZ model instead, referred to as the 'AZ-only' approach. It is important to note that this is not necessarily the approach suggested by Muthén and Asparouhov (2012). In fact, Asparouhov et al. (2015) suggests an alternative approach that avoids framing the problem as a binary decision. Although we do not oppose this way of thinking, we believe that our proposed methodology, as depicted in Figure 1, can provide insights into its behaviour. In our analysis, we consider the AZ-only approach, using PPP values with a threshold of .1 in all cases, and three different benchmarks of predictive performance based on EFA models. Since the data are simulated from a two-factor model, one might expect the optimal benchmark to be the two-factor EFA model (2F-EFA). However, in scenarios 2 and 3, the presence of error correlations and cross-loadings respectively may result in other EFA models performing better. Therefore, we also fit the two-factor EFA-C model (2F-EFA-C) defined earlier, as well as the three-factor EFA model (3F-EFA). Based on the cross-validated scores for these models, we construct the following three benchmarks and explore the performance of the suggested methodology with each of them:

1. (2F-EFA and 2F-EFA-C): Comparing the log score of the AZ-model with the best (smaller) log score between the 2F-EFA and 2F-EFA-C models.
2. (2F-EFA and 2F-EFA-C)+ ϵ: The same approach as in (2F-EFA and 2F-EFA-C) is used, but with the addition of a small tolerance value ϵ to the log score of each of the two benchmark models, 2F-EFA and 2F-EFA-C. The purpose of this is not necessarily to require the AZ model to be better than the benchmark models but just comparable. The constant ϵ is a positive value that introduces some tolerance. In this case, the results for $\epsilon = 0.7$ are reported, which resulted in the best performance in this particular exercise.
3. (2F-EFA, 3F-EFA): An alternative approach to account for cross-loadings or error correlations is to add an additional factor to the model. Therefore, in this case, we compare the AZ-model log score with the lowest log score between the 2F-EFA and 3F-EFA models.

For each scenario, we determine the most appropriate outcome and record the proportion of times each approach leads to that outcome. As discussed in Section 4.1, in scenario 1 the data are simulated from the EZ model, and therefore, the expected outcome is an indication of support from the data. Similarly, in scenario 2, the presence of small error correlations is not expected to provide evidence against the hypothesised theory. However, in scenario 3, where there are significant cross-loadings, the hypothesised theory should not be supported. Table 4 provides the proportion of 'correct' outcomes for each model assessment procedure across all three scenarios.

The simulation exercise clearly demonstrates that relying solely on the EZ-only approach to assess model fit can lead to over-rejection in scenarios 2 and 3. The approach indicates lack of support in almost all cases of scenarios 2 and 3, which is desirable in scenario 3 but concerning in scenario 2, where the correct decision is only reached in 3% of cases. On the other hand, the AZ-only approach accepts the hypothesised theory in all cases, resulting in complete failure in scenario 3. Therefore, the simulation exercise successfully highlights the need for a balanced approach with high accuracy. Although the benchmark of taking the minimum of 2F-EFA and 2F-EFA-C may set the bar too high in terms of predictive performance, resulting in too many instances of a lack of support being suggested, as shown

TABLE 4 Percentage of 'correct' decisions for each model assessment procedure across the three simulation scenarios.

Model assessment procedure	Scenario 1	Scenario 2	Scenario 3
EZ-only	98	3	100
AZ-only	100	100	0
(2F-EFA, 2F-EFA-C)	99	32	92
(2F-EFA, 2F-EFA-C)+ϵ	100	67	67
(2F-EFA, 3F-EFA)	100	97	78

in scenario 2, where the 'correct' decision is only reached in 32% of cases, it does offer an improvement over the EZ-only approach. The calibrated version of the benchmark provides a slight improvement, but the (2F-EFA, 3F-EFA) benchmark performs better overall. It is important to note that the proposed approach can be further improved by exploring different versions of priors, which could be investigated in future research.

5 | REAL-WORLD DATA EXAMPLES

In this section, we present our proposed model assessment framework using two real datasets. The first dataset is a widely used psychometric test known as the 'Big 5 Personality Test', which decomposes human personality into five main traits using 15 items measured on a 7-point Likert scale. The second dataset is based on the FTND and consists of six binary variables.

5.1 | Example 1: 'Big 5 Personality Test'

The data for this study were collected from the British Household Panel Survey in 2005–2006, and the sample consists of 589 female subjects aged between 50 and 55. The survey included the 'Big 5 Personality Test', which is a 15-item questionnaire assessing social behaviour and emotional state. Participants rate each item on a scale from 1 to 7, where 1 indicates 'strongly disagree' and 7 indicates 'strongly agree'. In this study, items are treated here as continuous, and the test aims to measure five major, potentially correlated, personality traits, with each trait corresponding to a factor that explains exactly three out of the 15 items.

Previous studies, such as Muthén and Asparouhov (2012), Stromeyer et al. (2015), and Asparouhov et al. (2015), analysed these data and found that the EZ model did not fit well based on various standard indices, including the PPP values. On the other hand, the AZ model gave a good fit in terms of the PPP values but had many non-zero error correlations, leading to concerns over whether the flexibility of the AZ model was capturing noise and producing a misleadingly high PPP value. Therefore, the validity of the 'Big 5' scale on these data remains uncertain. To shed more light on this question, we apply our model assessment framework and present the results in Table 5.

The findings from our analysis of the 'Big 5 Personality Test' dataset are consistent with the error correlations scenario in Section 4.2, but with a more pronounced effect. Our results confirm the poor fit of the EZ and the EFA with five factors. In contrast, both the AZ and the EFA-C model exhibit reasonably good PPP values, indicating that error correlations contribute significantly to the lack of fit. To further evaluate the question of overfit and the validity of the 'Big 5' scale, we calculate the log scores for each model. The log score of the AZ model outperforms all the other models, suggesting that it is fitting consistent patterns in the data and providing strong support for the 'Big 5' scale. The issues with the fit of the EZ model can be attributed to error correlations, which may have been caused by issues such as item wording and other factors present in the BHPS data.

TABLE 5 'Big 5' personality test data, BHPS.

Model	PPP	LS
EZ	.0	–
AZ	.23	0
EFA	.00	–
EFA-C	.38	3.49

Note: PPP values and sum of log scores (LS) obtained from three-fold cross-validation for the relevant models. The best-performing model is assigned a log score of 0, and the differences from this model are reported for the other models. The log scores are not reported when PPP values are <.1.

5.2 | Binary data: Fagerstrom test for nicotine dependence

In this section, we use data from the National Institute on Drug Abuse (study: IDA-CTN-0051) consisting of 566 patients. The data are based on the FTND (Heatherton et al., 1991), which is a test designed to measure nicotine dependence related to cigarette smoking. The FTND consists of six items that assess the amount of cigarette consumption, compulsion to use, and dependence. The original scale consists of four binary and two ordinal items for self-declared smokers:

1. FNFIRST: How soon after you wake up do you smoke your first cigarette? ['3' = Within 5 min, '2' = 6–30 min, '1' = 31–60 min, '0' = After 60 min]
2. FNGIVEUP: Which cigarette would you hate most to give up? ['1' = First one in the morning, '0' = All others]
3. FNFREQ: Do you smoke more frequently during the first hours after waking than during the rest of the day? ['1' = Yes, '0' = No]
4. FNNODAY: How many cigarettes/day do you smoke? ['0' = 10 or less, '1' = 11–20, '2' = 21–30, '3' = 31 or more]
5. FNFORBDN: Do you find it difficult to refrain from smoking in places where it is forbidden (e.g. in church, at the library, in cinema)? ['1' = Yes, '0' = No]
6. FNSICK: Do you smoke if you are so ill that you are in bed most of the day? ['1' = Yes, '0' = No].

For the purposes of our analysis, item FNFIRST was dichotomised as '1' = [3] and '0' = [0,1,2] and item FNNODAY as '1' = [2,3] and '0' = [0,1].

There is no clear mapping between the FTND scale and a CFA model, as noted in Richardson and Ratner (2005) and references therein. Richardson and Ratner (2005) fitted three models, namely a single factor, a correlated two-factor, and a two-factor model with one cross-loading, denoted by 1F, 2F-EZ, and 2F EZ-b respectively. More specifically, under the EZ model, items 1, 2, and 3 load on a 'morning' smoking factor, whereas items 4, 5, and 6 load on a 'daytime' smoking factor. The EZ-b model is specified by letting item 'FNFIRST' load on both factors. Additionally, we also considered their approximate zero versions, denoted by 1F-AZ, 2F-AZ, and 2F-AZ-b, as well as the two-factor EFA models with and without error correlations (2F-EFA and 2F-EFA-C). The results are shown in Table 6.

Based on the PPP values, we ruled out models 1F and 2F-EZ due to concerns about their fit. This raises several questions, such as whether the 2F-EZ-b model is the best or if any of the AZ models (2F-AZ or 2F-AZ-b) performs better, and which measurement scale should be used for the FTND test based on this dataset? To address these questions, we calculated cross-validated log scores for each model. Our analysis suggests that the best model is the 2F-EZ-b, which corrects the misspecifications of 2F-EZ with a single additional parameter. Moreover, the log score of 2F-EZ-b is smaller than that of

TABLE 6 FTND data, PPP values, and sum of log scores (LS) for three-fold cross validation for relevant models.

Model	PPP	LS
1F	.01	–
1F-AZ	.32	6.63
2F-EZ	.04	–
2F-AZ	.40	6.23
2F-EZ-b	.41	.00
2F-AZ-b	.44	2.01
2F-EFA	.44	2.66
2F-EFA-C	.58	2.38

Note: The best-performing model is assigned a log score of 0, and the differences from this model are reported for the other models. The log scores are not reported when PPP values are <.1. 1F and 2F refer to one- and two-factor models respectively. Models with '-b' are specified by letting item 'FNFIRST' load on both factors.

the EFA models, providing support for the scale with two correlated factors where the item 'FNFIRST' loads on both of them.

6 | DISCUSSION

In summary, the proposed methodology enhances the Bayesian SEM framework of Muthén and Asparouhov (2012) by providing tools for model exploration and assessment through the use of scoring rules combined with cross-validation. The study focuses on determining whether a hypothesised theory is supported by the data, which can be a challenging task due to the existence of multiple models, such as the EZ and AZ models, that are consistent with the theory. To address this issue, we propose incorporating a model selection step, based on a Bayesian assessment of the predictive performance, into the fit assessment procedure. If the model assessment does not provide enough support for the theory, further exploration of the model can be carried out by examining residuals, at both the individual and item levels, to identify the source of the problem. Such tasks could be facilitated by the extended modelling framework of this paper via item-individual random effects. However, this aspect is not covered in this paper and is left for future research.

The use of scoring rules has been demonstrated on simulated and real-world data, but further exploration is needed to understand the range of values indicating a good model under different settings, for example sample size, number of factors, parameter values, type of data, choice of scoring rules, number of folds or the form of cross-validation in general, and choice of benchmark model. Another important component, present in any form of Bayesian analysis, is the prior specification. While we provide a justification of the choice of priors in our paper, we acknowledge that we did not fully explore their impact on our results. Therefore, further investigation into the effect of priors on our framework is an area for future research. In general, it is difficult to disentangle the misfit caused by model misspecifications and unsuitable priors, since both affect the posterior distribution. However, it is certainly a good practice to explore different priors and evaluate their impact. In our case, the misfit is less likely to be attributed to the priors since we used either default low informative options or introduced priors on parameters that were otherwise fixed to zero. Hence, the resulting formulations are more likely to correct, and in some cases mask, likelihood misspecifications rather than introduce misfit. The main focus of our paper is the prevention of the masking of misfit issues that could result from the additional flexibility of the approximate zeros framework.

The calculations can be implemented using MCMC through standard user-friendly software like Stan and can be combined with existing packages for SEM. This provides the flexibility of using fast approximate methods, such as Variational Bayes (Kucukelbir et al., 2017), which are readily available and automated. This can be particularly useful in the analysis of categorical data, where the use of MCMC and high-dimensional latent variables can result in computationally intensive tasks, exceeding user expectations. Furthermore, Variational Bayes can be used to improve the efficiency of MCMC samplers.

The suggested methodology can be extended to ordinal and mixed data using a combination of the approaches described for continuous and binary data. For example, for a dataset with both continuous and binary variables, the goodness of fit can be monitored separately for each data type using PPP values based on the respective covariance matrices and response pattern frequencies. In terms of forecasting, log scores for each data type can be computed separately and then combined or even monitored separately. Additionally, ordinal data can be modelled using the same framework as binary data, which now involves assuming ordered thresholds for the underlying variable.

The methodology can also be extended to handle missing values. Assuming that the data are missing at random, posterior draws of the missing data can be obtained from the predictive distribution, given the posterior draws of the parameters $\boldsymbol{\theta}$ and the latent variables for each individual i, u_i, and z_i. This provides a natural multiple imputation approach for fitting the models discussed and assessing their fit via PPP values under missingness. For the prediction of unseen individuals, the posterior draws of $\boldsymbol{\theta}$ can be combined with draws from the prior distributions of u_i and z_i. However,

in the case of non-ignorable missingness, the factor models will need to be extended to incorporate the missing data mechanism (e.g. Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2017).

In addition to the aforementioned extensions, there are further possibilities for generalising the presented models. For instance, non-normal distributions can be assumed for the errors $\mathbf{e}_i$ or the random effects $\mathbf{u}_i$. It would also be interesting to explore the connection with Bayes factors that have the potential to identify parsimonious models that typically do well in cross-validation. Calculating Bayes factors can be challenging and their results may be sensitive to the choice of priors. Nevertheless, such issues can be alleviated by suitable choice of priors, as is done in this paper.

We would also like to draw a connection between the AZ model and model averaging. While it is not straightforward to perform model averaging between EFA models and EZ or AZ models, one could argue that the AZ model already incorporates a form of model averaging by allowing for models with small cross-loadings and error correlations. The EZ model can be seen as a special case of the AZ model, since the exact zeros are within the range of the approximate zero priors in the AZ model. The AZ model gives more prior weight to the EZ model but also allows models with small cross-loadings and error correlations; the prior weight on such models is determined by the approximate zero priors and gets smaller as cross-loadings and error correlations deviate from zeros. Alternative model averaging specifications, such as using spike and slab priors as in Lu et al. (2016), are also possible. However, such approaches may introduce identifiability or multi-modality issues that need to be carefully addressed.

Finally, it is important to note that the developed model assessment framework can be applied outside the Bayesian SEM context. In fact, it can be useful in situations where we need to assess the fit of a more flexible model, such as semi-parametric or non-parametric formulations (Song et al., 2013; Yang & Dunson, 2010). In such models, attaining a good fit is not always associated with a good systematic part of the model, as the flexibility in its error part can lead to overfitting. Such models arise in many scientific areas and go well beyond the SEM framework.

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest.

ORCID

Konstantinos Vamvourellis https://orcid.org/0000-0002-7355-6865
Konstantinos Kalogeropoulos https://orcid.org/0000-0002-0330-9105

REFERENCES

Asparouhov, T., & Muthén, B. (2021a). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 1–14.

Asparouhov, T., & Muthén, B. (2021b). Expanding the Bayesian structural equation, multilevel and mixture models to logit, negative-binomial, and nominal variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 622–637.

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on stromeyer et al. *Journal of Management*, *41*(6), 1561–1577.

Bhattacharya, A., & Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika*, *98*(2), 291–306.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley Series in Probability and Mathematical Statistics. Wiley.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, *103*(484), 1438–1456.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361.

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., & Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics*, *183*(1), 31–57.

Dawid, A. P., & Musio, M. (2014). Theory and applications of proper scoring rules. *METRON*, *72*(2), 169–183.

Dunson, D. B., Palomo, J., & Bollen, K. (2005). Bayesian structural equation modeling. *SAMSI# TR2005-5*.

Edwards, M. C. (2010). A Markov Chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474–497.

Erosheva, E. A., & Curtis, M. S. (2017). Dealing with reflection invariance in Bayesian factor analysis. *Psychometrika*, *82*(2), 295–307.
Fong, E., & Holmes, C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, *107*(2), 489–496.
Frühwirth-Schnatter, S., & Lopes, H. F. (2018). Parsimonious Bayesian factor analysis when the number of factors is unknown. *Unpublished Working Paper*.
Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*(1), 46.
Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555–497.
Geweke, J., & Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies*, *9*(2), 557–587.
Ghosh, J., & Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, *18*(2), 306–320.
Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.
Heatherton, T., Kozlowski, L., Frecker, R., & Fagerstrom, K. (1991). The Fagerstrom test for nicotine dependence: A revision of the Fagerstrom Tolerance Questionnaire. *British Journal of Addiction*, *86*, 1119–1127.
Hoijtink, H., & van de Schoot, R. (2018). Testing small variance priors using prior-posterior predictive *p* values. *Psychological Methods*, *23*(3), 561.
Jiang, Z., & Templin, J. (2019). Gibbs samplers for logistic item response models via the pólya–gamma distribution: A computationally efficient data-augmentation strategy. *Psychometrika*, *84*(2), 358–374.
Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*, 347–387.
Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Publications.
Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343–1370.
Krüger, F., Lerch, S., Thorarinsdottir, T., & Gneiting, T. (2021). Predictive inference based on Markov chain Monte Carlo output. *International Statistical Review*, *89*(2), 274–301.
Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, *18*(1), 430–474.
Liang, X. (2020). Prior sensitivity in Bayesian structural equation modeling for sparse factor loading structures. *Educational and Psychological Measurement*, *80*(6), 1025–1058.
Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1/2), 187–192.
Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, *14*, 41–67.
Lu, Z.-H., Chow, S.-M., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research*, *51*(4), 519–539.
MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods*, *17*(3), 340–345.
MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490.
Maydeu-Olivares, A., & Joe, H. (2005). Limited and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *6*, 1009–1020.
Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142–1160.
Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, *84*(3), 802–829.
Merkle, E. C., & Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. arXiv preprint arXiv:1511.05604.
Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A*, *163*, 445–459.
Muthén, B., & Asparouhov, T. (2012). Bayesian Structural Equation Modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335.
Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo* (vol. *2*, issue 11, p. 2). Chapman & Hall/CRC Press.
O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A*, *162*, 177–194.
Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, *9*(3), 523–539.
Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349.
Richardson, C. G., & Ratner, P. A. (2005). A confirmatory factor analysis of the Fagerstrom Test for Nicotine Dependence. *Addictive Behaviors*, *30*(4), 697–709.
Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*, 795–819.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*(1), 37–52.
Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*(4), 375–394.
Song, X.-Y., Lu, Z.-H., Cai, J.-H., & Ip, E. H.-S. (2013). A Bayesian modeling approach for generalized semiparametric structural equation models. *Psychometrika*, *78*(4), 624–647.
Stromeyer, W. R., Miller, J. W., Sriramachandramurthy, R., & DeMartino, R. (2015). The prowess and pitfalls of Bayesian structural equation modeling: Important considerations for management research. *Journal of Management*, *41*(2), 491–520.
Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239.
Van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363.
Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2), 667–718.
Vitoratou, V., Ntzoufras, I., & Moustaki, I. (2014). Marginal likelihood estimation from the metropolis output: Tips and tricks for efficient implementation in generalized linear latent variable models. *Journal of Statistical Computation and Simulation*, *84*(10), 2091–2105.
Yang, M., & Dunson, D. B. (2010). Bayesian semiparametric structural equation models with latent variables. *Psychometrika*, *75*(4), 675–693.

How to cite this article: Vamvourellis, K., Kalogeropoulos, K., & Moustaki, I. (2023). Assessment of generalised Bayesian structural equation models for continuous and binary data. *British Journal of Mathematical and Statistical Psychology*, *00*, 1–26. https://doi.org/10.1111/bmsp.12314

APPENDIX

A.1 | INVERSE WISHART

We recall here that the inverse Wishart distribution $\mathcal{IW}(D_p, d)$ is parameterised by matrix D_p of dimension $p \times p$ and d degrees of freedom, where we need $d > p + 1$ for the distribution to be well defined. The higher the value of d, the more concentrated the distribution gets around D_p. For example, if we choose $D_p = I_p$ the identity matrix of size p, then the marginal distribution of the diagonal elements will be distributed with mean $1/(d-p-1)$ and variance $2/[(d-p-1)^2(d-p-3)]$, whereas the off-diagonal elements will be distributed with mean 0 and variance $1/[(d-p)(d-p-1)^2(d-p-3)]$. Note that these expressions simplify when, for example, d is set to $p + 6$. We refer the interested reader to the appendix of Muthén and Asparouhov (2012) for more information.

A.2 | SENSITIVITY ANALYSIS FOR DATA-DEPENDENT PRIORS

We conducted a sensitivity analysis to assess the influence of data-dependent priors on the final results. To amplify the impact of the priors, we employed a relatively small sample size and simulated 200 data points from a standard two-factor model, following Simulation Scenario 1 in Section 4. We focused on fitting the EZ model and explored the use of various data-dependent and data-independent priors for the idiosyncratic variances ψ_j^2. Specifically, we employed the data-dependent prior proposed by Frühwirth-Schnatter and Lopes (2018) and Conti et al. (2014), which provides protection against Heywood cases. The prior is given by

$$\psi_j^2 \sim \text{InvGamma}(c_0, (c_0 - 1)/(S_y^{-1})_{jj}), \tag{A1}$$

where $c_0 = 2.5$ and S_y represents the sample covariance matrix. In addition to the data-dependent prior, we also considered three data-independent priors: InvGamma(0.1, 0.1), Half−Cauchy(5), and Uniform(0, 10). We obtained posterior samples from all four priors and examined the resulting posterior distributions of the free elements in the Λ matrix. Figure A1 displays the kernel density plots for the posterior of these elements. Notably, the plots indicate that the posterior density plots are nearly identical for all four priors. We observed similar results for the remaining model parameters. Based on these findings, we conclude that the choice of data-dependent prior does not significantly impact the final results of the analysis. However, it does serve the important purpose of safeguarding against Heywood cases.

A.3 | PARAMETER RECOVERY FOR AZ MODEL IN BINARY DATA CASE

To evaluate the performance of the AZ model in recovering parameters for binary data, we conducted a simulation experiment involving 100 different datasets. For each dataset, we simulated data from the EZ model and fitted the AZ model to obtain samples from its posterior distribution. Our focus was primarily on the factor loadings and the correlation between the factors. Each simulated dataset consisted of binary data, and the sample size was set to 2000. The factor loadings used to simulate the data were the same as those employed in Scenario 1 of the simulation experiments described in Section 4. The correlation between the two factors was set to .2. We assumed that the intercept parameters were all equal to 0. We employed a parameterisation where the loadings were unconstrained, while the variances of the latent factors were fixed at 1. This constraint ensured that the covariance matrix of the factors remained a correlation matrix.

In the prior specification of the AZ model, we incorporated the hypothesised theory by assuming that the first factor loaded on the first three items and the second factor loaded on the last three items. Consequently, the remaining loading parameters were treated as cross-loadings and were assigned informative priors around zero. As for the other parameters, we followed the prior specifications described earlier in the paper.

To assess the performance of parameter estimation in the AZ model, we employed informal measures based on frequentist properties of certain estimators derived from the posterior samples. These estimators included the following:

1. 95% credible intervals, which are interval estimators, by extracting the 2.5th and 97.5th percentiles from the posterior draws. These intervals provided a range of plausible values for the parameters, and we examined the coverage probability of these intervals.
2. We computed the posterior mean and the posterior median.

By analysing the coverage probability of the credible intervals and assessing the bias of the posterior mean and median, we aimed to evaluate the performance of these estimators in capturing the true parameter values. It is important to note that these summaries, namely the credible intervals, posterior mean, and posterior median, are not necessarily designed to exhibit optimal frequentist properties, even if the model fits the data well. However, if these estimators demonstrate good performance in terms of coverage and bias, it provides reassurance regarding their validity and reliability.

We examined the main parameters of interest, such as the loadings Λ and the factor correlation ϕ, in the AZ model. The results are presented in Table A1, which summarises the coverage probabilities and biases of the posterior summaries mentioned earlier. The coverage probabilities are reasonably close to .95, whereas the biases are not substantial, particularly for the posterior median. Based on these findings, we can conclude, while acknowledging the informal nature of the experiment, that there are no significant concerns regarding the recovery of the model parameters in the AZ model.

TABLE A1 True values, 95% coverage success rate, and bias of point estimators out of 100 replications, AZ model for binary data.

Parameter	True value	Coverage rate	Bias of post. mean	Bias of post. median
$\Lambda_{[1,1]}$	1.0	.94	.06	.03
$\Lambda_{[2,1]}$	.8	.96	.05	.03
$\Lambda_{[3,1]}$	.8	.94	.05	.03
$\Lambda_{[4,1]}$	.0	1.00	.00	.00
$\Lambda_{[5,1]}$	.0	1.00	.00	.00
$\Lambda_{[6,1]}$	.0	1.00	.00	.00
$\Lambda_{[1,2]}$	.0	1.00	.00	.00
$\Lambda_{[2,2]}$	.0	1.00	−.01	−.01
$\Lambda_{[3,2]}$	.0	1.00	.00	.00
$\Lambda_{[4,2]}$	1.0	.99	.03	.00
$\Lambda_{[5,2]}$	.8	.95	.06	.04
$\Lambda_{[6,2]}$	.8	.99	.03	.02
ϕ	.2	1.00	−.01	−.01

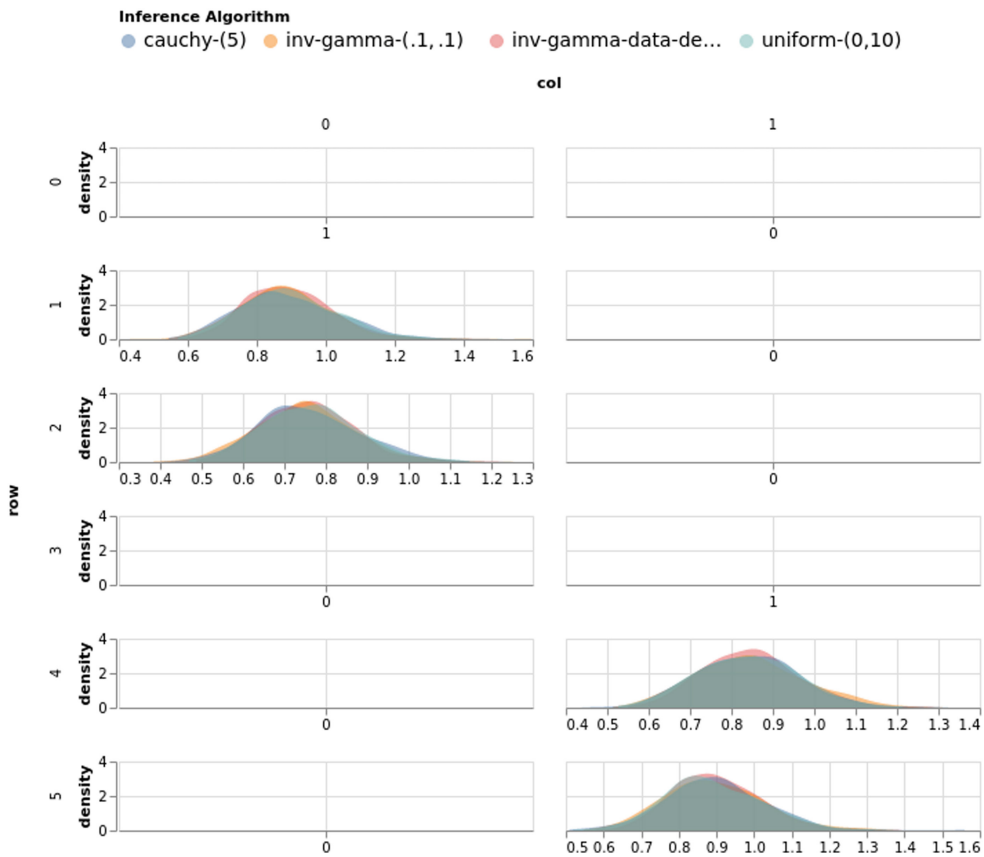

FIGURE A1 Posterior density plots of loading matrix parameters under four different prior choices. The model using a data-dependent prior (red) produces posterior density plots identical to the three other models using priors independent of the data.