The London School of Economics and Political Science

Bayesian Inference Methods for Latent Variable Modelling

Konstantinos Vamvourellis

A thesis submitted to the Department of Statistics
of the London School of Economics and Political Science
for the degree of Doctor of Philosophy,
London, December 2021

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 34,355 words.

Statement of co-authored work

I confirm that versions of Chapters 2, 3 and 4 were co-authored with Prof. Konstantinos Kalogeropoulos.
I confirm that a version of Chapter 2 was jointly co-authored with Prof. Irini Moustaki.

Abstract

This thesis develops novel Bayesian Inference methodology for a wide range of factor analysis models. The contributions consist of new Bayesian modelling approaches and the associated Bayesian inference methodology, novel model assessment frameworks and proposed applications of these models in clinical pharmacology. The first section focuses on developing a generalised framework for Bayesian Structural Equation Modelling (SEM) that can be applied to a variety of data types. It extends the previously available framework enhancing the available capabilities in two ways: it can handle binary and ordinal data, in addition to continuous data, which was known before, and it allows the errors to follow any distribution possible, removing the previously imposed restriction of Gaussianity. Moreover, it proposes a novel model assessment paradigm aiming to address shortcomings of posterior predictive p-values, which provide the default metric of fit for Bayesian SEM. The second section extends this framework to the sequential setting utilising techniques from the area of Sequential Monte Carlo. Sequential frameworks are powerful tools that can be used in dynamic settings where statistical inference is performed recursively on a continuous stream of data. In addition, this sequential approach is used for hypothesis testing, where it has been proven superior to the traditional null hypothesis (NHST) paradigm. Sequential Bayesian Factor (SBF) do not suffer from bias associated with the stopping rule, the practice of stopping the processing of new data only when conclusive evidence can be reached. The third section, extends these frameworks to data of mixed type, combining categorical and continuous types, to be used in clinical trial analysis where data is commonly of such mixed type. It develops a novel sequential modelling paradigm to inform regulatory evaluation of new drugs in real time while incorporating all the data available as well as clinical weights of importance relative to patient outcomes.

Acknowledgements

First and foremost, my gratitude goes to my advisor Prof. Kostas Kalogeropoulos for being my fearless guide throughout the doctoral programme. I will be forever grateful to Kostas for giving me inspiration and support to complete this thesis. Special thanks go to Prof. Irini Moustaki who has been a true joy to have met and worked with, Prof. Larry Phillips for providing good questions and conversations, and to Prof. Pauline Barrier for welcoming me warmly to the program.

I cannot thank enough my partner and soon-to-be wife, Kate, for encouraging me to pursue this PhD and supporting me throughout this process. Speaking of people I cannot thank enough, my family. I could not have made it without the love and support of my parents, Millie and Takis, and my brother Dimitris. I would also like to thank all my friends that gave me more than just a little help all these years in order to reach this milestone today. And of course, I am grateful to Jon, Jill and Judd for their unending encouragement. Special thanks go to my office mates on the 7th floor at the LSE Stats department for being a fun, friendly, and supportive bunch. An honourable mention goes to my dog, Moby, for being my faithful companion throughout the writing of the present thesis.

This work would not have been possible without the tireless work of Penny and all of the administrative staff at the Stats department. Thank you for all your hard work.

Finally I would like to thank all my teachers and mentors, academic and otherwise, who are too many to mention by name, but no less important in my development. Thank you to each one of you.

Contents

Abstract **5**

Acknowledgements **7**

1 Background and Summary **1**

1.1 Summary 1

1.2 Background 3

2 Generalised Bayesian Structural Equation Models **20**

2.1 Introduction 20

2.2 Generalised framework for Bayesian SEM 24

2.2.1 Model specification 24

2.2.2 Generalised Bayesian model framework 26

2.2.3 Overview of the models and their estimation procedure 34

2.3 Model assessment 37

2.3.1 Assessing goodness of fit with PPP values 38

2.3.2 Scoring rules and cross validation in SEM 40

2.3.3 Scoring rules for continuous and normally distributed data 43

2.3.4 Scoring rules for binary and ordinal data 44

2.3.5 Model assessment with fit and predictive performance indices 44

2.4 Simulation experiments 49

2.4.1 Setup 49

2.4.2 Continuous data 51

2.4.3 Binary data 52

2.4.4 Parameter recovery for the AZ model in the binary data case. 53

2.5 Real-world data examples 55

2.5.1 Example 1: 'Big 5 Personality Test' 55

2.5.2 Binary Data: Fagerstrom Test for Nicotine Dependence 56

2.6 Discussion 58

3 Sequential Bayesian Inference for Factor Analysis 61

3.1 Introduction 61

3.2 Sequential Monte Carlo for Factor Analysis Based on Continuous Data 64

3.2.1 Model and Priors 64

3.2.2 Sequential Algorithm 66

3.3 Sequential Monte Carlo methods for Binary data 70

3.3.1 Model, Priors and MCMC Scheme 70

3.3.2 Sequential Scheme 71

3.4 Applications 73

3.4.1 Continuous Data Simulations 74

3.4.2 Parameter Estimation for Binary Data 76

3.4.3 Sequential Model Choice 83

3.4.4 Application: Big 5 Personality Test - British Household Panel Survey . . 93

3.5 Discussion 97

4 Bayesian Benefit Risk Analysis **100**

4.1 Introduction 100

4.2 Factor Models for Multi-criteria Decision Analysis 104

4.2.1 Multi-criteria Decision Analysis Score and Data 104

4.2.2 Factor Analysis for Mixed Type Data 105

4.3 Model Assessment 110

4.3.1 PPP Values 111

4.3.2 Scoring Rules 113

4.4 Sequential Monte Carlo 114

4.4.1 IBIS Algorithm for Benefit Risk Analysis 115

4.5 Rosiglitazone Case Study 117

4.5.1 Data and MCDA Setup 118

4.5.2 Model Choice 119

4.5.3 MCDA Scores 123

4.5.4 Sequential Analysis 123

4.6 Discussion 126

Appendices **129**

A Generalised Bayesian Structural Equation Models **130**

A.1 Inverse Wishart 130

A.2 Sensitivity analysis for data-dependent priors 131

B Sequential Latent Variable Modelling **133**

B.1 HMC Implementation Details 133

B.2 IBIS with Laplace Approximation 133

Bibliography **136**

References **137**

List of Tables

2.1 True factor loadings used in the three simulation scenarios. 49

2.2 Simulation Results for Continuous Data. PPP values and sum of variogram scores of 3-fold cross validation for the relevant models. For each scenario, the best model has 0 variogram score and the differences from it are reported for the other models. 51

2.3 Simulation Results for Binary Data. PPP values and sum of variogram scores of 3-fold cross validation for the relevant models. For each scenario, the best model has log score equal to 0 and the differences from it are reported for the other models. 53

2.4 True values, 95% coverage success rate and bias of point estimators out of 100 replications, AZ model for binary data. 54

2.5 'Big 5' personality test data, BHPS. PPP values and sum of variogram scores of 3-fold cross validation for the relevant models. For each scenario, the best model has 0 variogram score and the differences from it are reported for the other models. 56

2.6 PPP values and and sum of log scores of 3-fold cross validation for the relevant models. The models with '-b' refer to the measurement model with the first item loading to both factors. The best model had log score equal to 0 and the differences from it are reported for the other models. 58

3.1 True factor loadings used in continuous data simulation. 88

3.2 Loading structure of all three models, x represents a free parameter, ~ 0 represents a free parameter with a prior distribution concentrated around 0. 88

3.3 Log Marginal Likelihood or Log Model Evidence for candidate models in simulation Scenario 1, at the final point $i = 200$. We highlight model EZ with the highest value. 90

3.4 Log Bayes Factor for candidate models in simulation Scenario 1, at the final point $i = 200$. The table values represent the log ratio of the model on the top row divided by the model on the column. For example the $LBF(\text{EZ/AZ}) = 0.3$. 90

3.5 Log Marginal Likelihood or Log Model Evidence for candidate models in simulation Scenario 2, at the final point $i = 200$. We highlight model AZ with the highest value. 91

3.6 Log Bayes Factor for candidate models in simulation Scenario 2, at the final point $i = 200$. The table values represent the log ratio of the model on the top row divided by the model on the column. For example the $LBF(\text{EZ/AZ}) = -9.8$. 91

3.7 Posterior mean estimates of loading values for the Big 5 dataset at the final point of inference. We present the values at 1 decimal point. 97

4.1 Outcomes of interest and MCDA parameters. 119

4.2 Hypothesised factor loading structure: the first factor z_1 loads onto the efficacy variables, the first two items, and the second factor z_2 loads onto the risk variables, last four items. 120

4.3 Summary of out of sample predictive performance for all candidate models using scoring rules. Sum of variogram and log scores of 3-fold cross validation for continuous and binary data respectively. 121

4.4 Model Fit Metrics for the best performing models. 122

4.5 True values, 95% coverage success rate and bias of point estimators out of 100 replications. 122

4.6 Schematic sequential schedule of synthetic re-ordering of the original clinical trial data. 124

List of Figures

3.1 Real data values marked in red dots overlaid with the 95% credible intervals for Λ, the loading matrix parameters in Continuous Scenario 1. 76

3.2 Real data values marked in red dots overlaid with the 95% credible intervals for α, the intercept parameters in Continuous Scenario 1. 76

3.3 Real data values marked in red dots overlaid with the 95% credible intervals for Φ, the factor covariance parameters in Continuous Scenario 1. 77

3.4 Real data values marked in red dots overlaid with the 95% credible intervals for diag(Θ) the diagonal elements of the residual covariance parameters in Continuous Scenario 1. 77

3.5 Posterior Draws for the Loading matrix Λ in Continuous Scenario 1, for both IBIS and batch MCMC methodologies. 78

3.6 Posterior Draws for the Intercept parameters α in Continuous Scenario 1, for both IBIS and batch MCMC methodologies. 79

3.7 Posterior Draws for the Factor Covariance matrix Φ in Continuous Scenario 1, for both IBIS and batch MCMC methodologies. 79

3.8 Posterior Draws for the diagonal of the covariance matrix of the residual Errors diag(Θ) in Continuous Scenario 1, for both IBIS and batch MCMC methodologies. 80

3.9 Real data values marked in red dots overlaid with the 95% credible intervals for the loading matrix parameters Λ in the Binary Scenario 1 data simulation. . . . 82

3.10 Real data values marked in red dots overlaid with the 95% credible intervals for the intercept parameters α in the Binary Scenario 1 data simulation. 83

3.11 Posterior Draws for the Loading matrix Λ in for EZ model in the binary data simulation, for both IBIS and batch MCMC methodologies. 84

3.12 Posterior Draws for the intercept parameters α of EZ model in the binary data simulation, for both IBIS and batch MCMC methodologies. 85

3.13 Bayes Factors for EFA models of 1, 2 and 3 factors respectively in Scenario 1. We see that EFA2 outperforms both of the other models, which is an indication that the right number of factors is 2. 93

3.14 Bayes Factor values of 3 candidate models, in Scenario 1 where the data was generated from a structure with zero cross loadings. We expect EZ to be the model of choice since it matches the structure of the data generation model the best of all the candidate models. 94

3.15 Bayes Factors for EFA models of 1, 2 and 3 factors respectively in Scenario 2. We see that EFA2 outperforms both of the other models, which is an indication that the right number of factors is 2. However, because of model misspecification the EFA model with an additional factor, EFA3, remains competitive. 94

3.16 Bayes Factor values of the 4 candidate models, in Scenario 2 where the data was generated from a structure that included cross loadings. We expect AZ to be the model of choice since it matches the structure of the data generation model the best of all the candidate models. 95

3.17 Bayes Factor values of the 4 candidate models. Based on prior research we expect AZ to be the model of choice. 96

3.18 Posterior density plots for the loadings parameter of the AZ model fit to the Big5 dataset. 98

4.1 MCDA Scores at the end of the sequential run. Avandia (AVM) scores substantially higher than the other two treatments, Rosiglitazone (RSG) and Metformin (MET). 123

4.2 Sequentially updated probabilities of $P(s_{\text{AVM}} > s_{\text{MET}})$ and $P(s_{\text{AVM}} > s_{\text{RSG}})$. We see tha the probabilities converge to 1 within the first 300 patients. A dynamic trial could have concluded early or could have assigned the remaining patients to AVM given that it is considered a better treatment based on MCDA scores. . 125

4.3 Sequentially updated probabilities of $P(s_{\text{RSG}} > s_{\text{MET}})$. We see that the comparison is inconclusive as the probability fluctuates around 0.5. Under a dynamic clinical trial it would be possible to monitor these probabilities and allocate patients away from treatment arms that results have converged early, such as the AVM-RSG and AVM-MET pair comparisons, and into groups that require more data to become conclusive, such as RSG-MET presented here. 126

4.4 Sequentially updated posterior mean (lines) and the 95% central quantiles (shade bands) of the posterior distributions for the final MCDA population scores s_r for each treatment $r \in \{\text{AVM, MET, RSG}\}$, at each point in the trial. 127

A.1 Posterior density plots of the loading matrix parameters under 4 different prior choices. The model using a data-dependent prior (red) produces identical posterior density plots as three other models using priors independent of the data. . 132

Chapter 1

Background and Summary

1.1 Summary

This research thesis concerns Bayesian factor modelling methodologies for a variety of data types, including continuous and categorical. The frameworks presented cover sequential and non-sequential settings. All models are implemented and applied in simulated and real world datasets, and the accompanying code is openly accessible on github. The final part of the thesis employs the frameworks proposed to analyse data from a real clinical pharmacology setting.

The thesis consists of three successive research papers.

The first paper is presented in Chapter 2, titled 'Generalised Bayesian Structural Equation Modelling'. There we propose a generalised framework for Bayesian Structural Equation Modelling (SEM) that can be applied to a variety of data types. It extends previously suggested models by Muthén and Asparouhov (2012) and can handle continuous, binary, and ordinal data. Moreover, we propose a novel model assessment paradigm aiming to address shortcomings of posterior predictive $p-$values, which provide the default metric of fit for Bayesian SEM. We incorporate scoring rules and cross-validation to supplement existing model assessment metrics for Bayesian SEM. The methodology is illustrated in continuous and categorical data examples via simulation experiments as well as real-world applications on the 'Big-5' personality scale

and the Fagerstrom test for nicotine dependence.

The second paper is presented in Chapter 3, titled 'Sequential Bayesian Inference for Factor Analysis'. There we extend the modelling techniques presented in Chapter 2, to the sequential setting. We develop an efficient Bayesian sequential inference framework for factor analysis models observed via various data types, such as continuous, binary and ordinal data. In the continuous data case, where it is possible to marginalise over the latent factors, the proposed methodology tailors the Iterated Batch Importance Sampling (IBIS) of Chopin (2002) to handle such models and we incorporate Hamiltonian Markov Chain Monte Carlo. For binary and ordinal data, we develop an efficient IBIS scheme to handle the parameters and latent factors, combining with Laplace or Variational Bayes approximations. The methodology can be used in the context of sequential hypothesis testing via Bayes factors, which are known to have advantages over traditional null hypothesis testing. Moreover, the developed sequential framework offers multiple benefits even in non-sequential cases, by providing posterior distribution, model evidence and scoring rules (under the prequential framework) in one go, and by offering a more robust alternative computational scheme to Markov Chain Monte Carlo that can be useful in problematic target distributions.

The third paper is presented in Chapter 4, titled 'Bayesian Benefit Risk Analysis'. In this final chapter we develop a benefit risk analysis framework to compare three different treatments for type 2 diabetes. We analyse a proprietary data set of a clinical trial conducted to understand the benefits and risks of the proposed treatments. We employ the Multi-Criteria Decision Analysis (MCDA) framework that facilitates decision making amidst multiple competing actions. To create a holistic modelling framework we combine MCDA with factor models that can accommodate data of mixed type, as is the case in most clinical trials. Typically efficacy variables (benefits) are continuous items, whereas adverse effects (risks) are categorical items. To accommodate such cases we tailor the modelling framework of Chapter 2 and also provide a framework for model choice in this setting. Additionally, we propose a sequential clinical trial paradigm, using the methodologies developed in Chapter 3. This proposed framework offers multiple benefits: (i) it allows us to recursively update model estimates with each new subject receiving the treatment; (ii) it permits stopping the exposure as soon as the research

requirements are satisfied reducing unnecessary further exposure to undesirable treatments; (iii) it allows us to assign treatment groups dynamically based on research objectives. We demonstrate these benefits by applying it to the dataset at hand and highlight what could have been done differently under the sequential setting.

1.2 Background

Factor Analysis

Factor Analysis is a modelling technique involving latent variables often used in social sciences, psychometric and econometric theories among other fields. It provides a method to study variables of interest that are not directly measurable. The classical factor model equation is

$$\mathbf{y} = \alpha + \Lambda \mathbf{z} + \boldsymbol{\epsilon} \tag{1.1}$$

where $\mathbf{y}$ is the observed data, $\mathbf{z}$ are the latent variables and $(\alpha, \Lambda, \boldsymbol{\epsilon})$ are the model parameters to be estimated. For one example, consider that $\mathbf{z}$ is the mathematical ability of high school students that is never directly observed, rather it is indirectly deduced in relation to measurable and quantifiable variables, denoted by $\mathbf{y}$, such as test scores. Another example could be a social attitude, such as introversion, which is never observed as a measurable quantity, but rather can only be expressed via a set of observable actions or survey responses. Factor modelling is a principled methodology that allows us to gain insights and quantify such latent variables of interests, by analysing the associated observable items.

The dimension of the latent variables is typically much smaller than the dimension of the observed items, so that a single latent variable is associated to more than one observed variable. This way factor models also function as a dimensionality reduction technique aiming at condensing a dataset of observations by expressing its structure using a lower dimensional construct. In summary, factor analysis offers a way to determine whether a small number of latent

variables can account for the dependencies among the observed variables. Note that the dual goals of dimensionality reduction and gaining insights into latent variables are by no means mutually exclusive, and in fact real practical applications of factor analysis typically employ both angles in the course of the data analysis.

Latent variable models is the general umbrella term for all models that involve latent variables, variables that are never directly observed, but rather deduced via their connection to observed variables. When the analysis utilises categorical latent variables, it is called Latent Class Analysis for categorical observations and Latent Profile Analysis or Mixture Models for continuous observations. On the other hand when the analysis utilises continuous latent variables, it is termed Factor Analysis when the observations are also continuous and Latent Trait Analysis when the observations are categorical. This thesis concentrates on the latter two types of factor analysis, where the latent variables are continuous.

Factor Analysis is concerned with uncovering common latent sources of influence (*factors*) among the observed variables (*manifest variables*). In cases where we observe a dependency between two variables it is often useful to examine if that is because of a common dependency on a third variable. If such third variables were directly observed then regression analysis would be the natural modelling framework to understand the correlation structure. When they are not directly observed, factor analysis can be used to determine whether any correlation among the observed variables can be explained due to common dependency on a smaller set of latent variables, the factors. These factors could be convenient constructs that provide a possible way to explain the dependencies or they can be hypothesised quantities of interest that are anticipated due to previous research. The latter often arises in case of data sets from surveys, designed specifically to study such latent quantities of interest.

Factor analysis is usually grouped in two main categories, Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). Informally, EFA is the process of uncovering a compressed, lower dimension, representation of the data. More formally, EFA "can be construed as an inductivist method designed to discover an optimal set of factors that accounts for the covariation among the items" (Skrondal and Rabe-Hesketh, 2004). EFA aims at discovering the

factor structure, i.e. which observed variables/items load to which latent factors. This means that loading parameters Λ are free but then scores $\mathbf{z}$ are assumed independent for identifiability reasons (more on this later on). Confirmatory Factor Analysis (CFA) aims at examining a hypothesised structure. Skrondal and Rabe-Hesketh (2004) defines it as "a hypothecist procedure designed to test hypotheses about the relationships between items and factors, where the number and interpretation of the factors are given in advance. Hence, in the confirmatory mode, particular parameters are set to prescribed values". Usually the scores $\mathbf{z}$ are allowed to be correlated (or with unit diagonal) and restrictions (zeros) are placed on the loading parameters Λ. A generalisation of the CFA framework for testing research hypotheses arising in psychology and social sciences is Structural Equation Modelling (SEM). It is used to model directional (regression coefficients) and non-directional (correlations) relationships among latent variables (structural model) which are identified by observed variables (measurement model) (Bollen, 1989). The theory of factor analysis extends to SEM models so while we choose to give examples mostly in factor analysis, the methodologies developed in this thesis extend to SEM models as well.

Classical Model Assessment

Although ideally we would measure the absolute fit of a model, in practice we can usually only measure the relative fit of one model versus a set of pre-specified alternative models. Such comparisons can be achieved using relative fit criteria. Conventional hypothesis testing can be used to compare nested models. Specifically, likelihood ratio tests compare a constrained model M_0 with k_0 parameters (null) to an expanded model M_1 (alternative) with k extra parameters $k_1 = k_0 + k$. Nested models can also be compared using Wald or Lagrange Multiplier Score tests. Assuming certain regularity conditions, under the null, all three tests are asymptotically distributed as chi-squared χ^2_k with k degrees of freedom (Cox and Hinkley, 1979; Buse, 1982; Engle, 1984), but do have meaningful differences in the finite sample case (Hauck and Donner, 1977; Pawitan, 2001).

Such conventional tests, aside from their limited applicability, have a number of weaknesses that

make them unsuitable in many cases. Asymptotic results hold under a number of conditions that are often not satisfied in practice. For example, if the sample size N is large compared with 2^p (all the possible patterns of the categorical data), the expected frequency for each response pattern is likely to be large enough to carry out a valid chi-squared (χ^2) test to compare the observed and expected frequencies. As 2^p becomes large, relative to N, the contingency tables become too sparse and no test can be done, see Bartholomew, Knott, and Moustaki (2011) section 4.9. Another weakness of the conventional testing approach is that the power of hypothesis tests depends on sample size. While acknowledging that more observations imply more information, it nevertheless appears difficult to base model assessment actions on this criterion. This point can be made clear by considering a situation with a very large number of observations. Here, we expect all but extremely complicated models to be rejected and we are left with 'models' which merely mirror the particular data set at hand. On the other hand, if few observations are available, we expect that oversimplifications tend to be retained (Skrondal and Rabe-Hesketh, 2004). Finally, significance probabilities can point to contradictory directions and/or disagree with the evidence, even for the unrealistic case of solely two nested models (Berger and Sellke, 1987; Berger and Delampady, 1987).

There are no similar tests for model comparisons between non-nested models. This is particularly relevant to the EFA context because most models of interest in exploratory analysis are non-nested. Hence there is no agreed upon way to do EFA in practice, leaving unanswered one of the most basic questions of EFA, determining the appropriate number of factors. Skrondal and Rabe-Hesketh (2004) explain the status quo for continuous data as follows. "An exploratory factor analysis usually proceeds through the following rather ad hoc steps. First the number of factors is determined based on a principal component analysis of the correlation matrix. The number of factors is typically chosen to be equal to the number of eigenvalues that are larger than one, the so-called Kaiser-Guttmann criterion. Sometimes, however, a so-called scree-plot is used where the eigenvalues are plotted against their rank and the number of factors is indicated by the 'elbow' of the curve (Cattell, 1966)". For categorical data there are fewer options, mainly information criteria metrics, as we describe later on.

Model Assessment for CFA

Informally CFA describes the scenario of having a single theory that is to be tested against the data. For example the theory could postulate the existence of two factors, not necessarily independent, and each factor is associated with distinct sets of measured variables. It is natural to ask if the data support this theory and by how much. In practice this takes the form of comparing the fit of the tested model to a benchmark model where certain parameters are set to pre-specified values, typically zero. In some test cases, the benchmark model can also be the saturated model for a given set of variables. Such comparisons to a baseline are considered a measure of absolute fit within the context of factor analysis. Any hypothesis tests discussed above can be used, since the models are usually nested in this case, with the likelihood ratio test being the most popular. In covariance structure modelling the null hypothesis corresponds to a restricted model and the alternative to the empirical covariance matrix[1]. Such methods are usually applied to detect misspecification but cannot detect the omission of variables.

Another interesting non-nested scenario of CFA is the multi-group analysis, which occurs when we want to compare the fit of the model to two or more sets of data. For example if we test the above theory in data collected from two different countries, it might be natural to ask if there are meaningful deviations between the two countries. There are usually likelihood ratio tests for this cases but when the number of factors is even moderately large (bigger than 3) it becomes subject to big approximation errors. The Bayesian model choice, and posterior predictive checks in particular, provide an alternative measure of absolute fit.

Goodness of fit indices

One method to circumvent the limitation of the hypothesis tests mentioned above, is to use one of the standardised forms of the chi-squared statistic known as goodness of fit indices (GFI). GFIs are easy to use; they are standardised between 0 and 1, 0 being the worst fit possible

[1]It's worth noting that that the status of null and alternative hypothesis is reversed in this case, compared to the standard framework for statistical testing as the researcher desires to retain the null hypothesis in favour of the alternative.

and 1 being the best. Also, indices tend to be less sensitive to departures from the ideal model conditions or different sample sizes. GFIs compare the incremental fit of a model over a baseline (dictated by the choice of GFI), measured as the discrepancy between the sample and the estimated covariance matrices. There is a wide range of GFIs, we refer the interested reader example to Bartholomew et al. (2011) for a comprehensive list. That is more of a con than a pro, as there is often no principled way to choose among the many options. Two representative examples are the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). The CFI is a relative fit index; it compares the chi-squared statistic of the fitted model with the chi-squared of a baseline model, usually the independence model (the degrees of freedom of each model are taken into account). The RMSEA is calculated without an explicit baseline using the chi-squared statistics and the degrees of freedom of the fitted model. In practice using such indices can be problematic. Aside from having to make an arbitrary choice, GFI are heavily affected by the baseline model they use (Sobel and Bohrnstedt, 1985) among other issues (Skrondal and Rabe-Hesketh, 2004).

Information Criteria

Practitioners can also use conventional information criteria when choosing between models. These metrics are easy to calculate and interpret. *Deviance*, twice the difference in log-likelihood between a model and the saturated model, gives rise to the deviance information criterion (DIC). Related criteria are the Bayesian (BIC) and the Akaike information criteria (AIC), which in addition to approximating deviance they also penalise model complexity. For categorical data specifically, another easy-to-use metric is the Pearson X^2 statistic. Despite limitations these are widely used in model selection, mainly because of a lack of alternative, especially for non-nested models.

Posterior Predictive Checks

Posterior predictive checks (Rubin, 1984; Meng, 1994b; Gelman et al., 1996) is the method of comparing the observed data with data that the model generated. The comparison can be done visually, or quantitatively with diagnostic measures of discrepancy between the two datasets. It is left to the researcher to decide what constitutes acceptable levels of discrepancy. Posterior predictive checks, as the name suggests, compare models only at the prediction level. As such, it can be used to compare any set of models, no matter how simple or complex. For examples of using posterior predictive checks for IRT see Sinharay et al. (2006); Levy et al. (2009).

Bayesian model choice can be a powerful alternative, especially for the non-nested model comparison. Methods to compute marginal likelihood for Bayesian model comparison include reversible jump MCMC (Green, 1995) and Bridge Sampling (Gelman and Meng, 1998). When the latent variables are Gaussian another method is Integrated Nested Laplace approximations (Rue, Martino, and Chopin, 2009). All these methods require additional work beyond the task of inference and can be computationally involved, especially for non-trivial models. As we will see later, sequential inference methods offer an attractive alternative because they produce estimates of the marginal likelihood as a byproduct of inference. However, the existing sequential methodologies do not cover factor analysis models, especially for categorical data. As we will discuss later, this thesis addresses the issue of sequential inference for the generalised latent variable models including the case of categorical data.

Overall, the existing methodologies for model assessment are not able to provide clear guidance on how to proceed in the case when a confirmatory model does not exhibit 'good fit'. One possible reason for the lack of fit is that the hypothesised structure is inadequate to capture the variability of the data. A second possibility is that the structure is reflected too strictly in the model, causing the available testing apparatus to reject the model. A third possibility is that there is imperfect measurement, such as measurement error misspesification, unrelated to the hypothesised structure that is causing the model to be rejected. This situation arises often due to wording differences between items in survey data. In the absence of conclusive

testing mechanism researchers use modification indices by which they free one by one different parameters of the model to see which changes result in a better fit. This technique however is not advisable as it is prone to capitalising on chance and arriving at false conclusions. The alternative of freeing all parameters is not feasible either, since it would result in an over parameterised EFA model. These challenges apply equally to the continuous and categorical data cases. Hence, developing a comprehensive model assessment framework for CFA and SEM models remains an elusive task in factor analysis.

Identifiability Challenges

Allowing for latent variable in a model is a useful abstraction but fitting them to data poses a challenge. The observations are usually not enough to identify uniquely the latent values. The issue boils down to the simple fact that the likelihood is invariant under certain transformations of the latent variables, and hence the likelihood has multiple maxima. Although most identifiability issues can be seen as illustration of the same underlying phenomenon, one finds them under a plethora of names in the literature. In categorical latent variables the issue is often referred to as *label switching* or *label ordering* problem; see section 22.3 in Gelman et al. (2014), Betancourt (2017), section 11.3.1 in Murphy (2012), Jasra et al. (2005). In continuous latent variables the non-identifiability is usually referred to as *non-uniqueness* or *rotational indeterminacy*; see section 12.1.3 in Murphy (2012), Lopes and West (2004), section 2.11 in Bartholomew et al. (2011). Note that the lack of identifiability is a theoretical issue and as such, it cannot be resolved by investing more computational power.

Non-identifiability usually manifests itself in a multimodal posterior distribution, as we explain in more detail in the examples that follow. From a practical point of view, finding any one of the modes suffices for most cases. As a result, MLE techniques that are designed to find local maxima and purposefully avoid exploring other modes could yield satisfactory answers without the need to artificially constraint the parameter space. Bayesian inference on the other hand, is more susceptible to these problems because it is designed to explore all the available space for modes. All random walk algorithms suffer, to varying degree, when the space they

are trying to explore is multimodal. That's why LVMs present a bigger challenge for Bayesian inference. As Bartholomew et al. (2011) put it "However, the implications of rotation and label switching have not been thoroughly investigated under the Bayesian framework and workable solutions still need to be given. As with rotation, any random permutation of the set of factors leaves the joint distribution of the observed variables unchanged, which often leads to multi-modal posterior distributions for the parameters of interest. To our knowledge no research has been published that addresses label switching in latent variable models. For the classical factor analysis model, restrictions similar to those applied for model identifiability have usually been adopted as a practice for avoiding label switching." The authors continue "in a Bayesian estimation framework, rotational indeterminacy often leads to computational problems such as non-convergence or label switching among the factors in the multidimensional case. Lopes and West (2004) suggested constraints for the classical factor analysis model, but no systematic work has been done for the factor analysis model for categorical manifest variables. Re-parameterisations of the model are in some cases necessary to avoid lack of convergence when large discrimination parameters occur."

Unidentifiability in IRT models

The *label switching* issue can be more easily understood in the case of discrete latent classes, called mixture models. There are multiple possible labellings of the latent classes and they are indistinguishable from the perspective of likelihood. As a result each possible labelling ($k!$ possibilities for k classes) has to be equally likely which creates a multimodal posterior distribution. The case of continuous latent factors suffers from a generalisation of the label switching problem. At a high level, the likelihood is invariant to a broader range of transformations of the factors. Any orthogonal transformation, not just permutations, can be shown to leave the likelihood unchanged, and hence the latent factors can only be identified "up to orthogonal transformation". Let's focus on the case of IRT models with p binary items and k latent variables as an example. Each response variable y_j, conditionally on the latent vector, follows a Bernoulli($\pi_j(\mathbf{z})$) distribution with success probability $\pi_j(\mathbf{z})$ which, for the logit link, is given by

$$\text{logit}(\pi_j(\mathbf{z})) = \alpha_j + \sum_{\ell \neq 1}^{k} \Lambda_{j\ell}\, z_\ell, \quad \text{for} \quad j = 1, \ldots p\ . \tag{1.2}$$

Under this approach, assuming a random sample of N individuals with y_{ij} responses on p items (i.e. $i = 1, \ldots, N$ and $j = 1, \ldots, p$) the model is written

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(\pi_j(z_i)) \\ \upsilon\{\pi_j(z_i)\} &= \alpha_j + \sum_{\ell \neq 1}^{k} \Lambda_{j\ell}\, z_{i\ell}, \\ z_{i\ell} &\sim N(0,1) \end{aligned}$$

where $z_i = (z_{i1}, \ldots, z_{ik})$ is the vector of latent variables for i-th individual,and $\pi_j(z_i)$ is the success probability of i subject and j item, and υ is the link used.

The linear predictor of the GLLVM can be re-written in a matrix form by

$$\upsilon\{\mu(\mathbf{z})\} = \mathbf{A} + \mathbf{z}\boldsymbol{\Lambda} \tag{1.3}$$

where $\mu(\mathbf{z})$ is a $N \times p$ matrix with elements $\mu_{ij}(\mathbf{z})$, $\boldsymbol{A} = \boldsymbol{\alpha}' \otimes \mathbf{1}_N$ is a $N \times p$ matrix with elements $A_{ij} = \alpha_j$ and $\mathbf{z}$ is $N \times k$ matrix with elements $z_{i\ell}$. Any orthogonal transformation of the factors in (1.3) will leave the model unchanged since

$$\upsilon\{\mu(\mathbf{z})\} = \mathbf{A} + \mathbf{z}\boldsymbol{U}\boldsymbol{U}'\boldsymbol{\Lambda}.$$

Hence, there are infinite possible solutions with respect to the model slope parameters, $\mathbf{B}^* = \boldsymbol{U}'\boldsymbol{\Lambda}$. Note that permutations are orthogonal transformations and hence label switching is a special case of the non-uniqueness.

The so-called *non-uniqueness* or *rotational indeterminacy problem* is dealt with the use of constraints imposed on the model parameters to ensure a unique solution; for a detailed solution

see Bartholomew et al. (2011) section 2.11. One of the first publications on the topic was by Anderson and Rubin (1956) who give sufficient or necessary conditions to eliminate the problem of rotation in factor analysis models. Ever since, several solutions have been proposed by imposing zero or non-zero constraints on the loadings matrix; see in Lawley and Maxwell (1963); Jöreskog (1969); Dunn (1973); Steiger (1979); Algina (1980) among others.

Geometrically, multiplying Λ by an orthogonal matrix $\boldsymbol{U}$ is like rotating $\mathbf{z}$ before generating $\mathbf{y}$, which leaves the likelihood unchanged. Recall that Λ is a $p \times k$ matrix where $k < p$, and hence the space of all matrices that can act on Λ has dimension k^2. To ensure unique solution we have to remove from our search space all the orthogonal ones, there are $k(k-1)/2$ of them. Some methods to achieve this is restricting Λ to be orthonormal, or just lower triangular. In the latter case the total number of parameters can be calculated to be $p+pk-k(k-1)/2$ which is equal to the number of identifiable parameters. One caveat of this method is that the interpretation of the latent factors is dependent on the ordering of the factors (Fruhwirth-Schnatter and Lopes, 2010).

For further readings on identifiability see Williams (1978); Elffers et al. (1978); Bekker (1986); Sato (1991); Browne (2001) and references within. In particular, the specific side effects of constraining the parameter space to achieve identifiability is an active area of research. One of the most recent works on this topic is by Fruhwirth-Schnatter and Lopes (2010) where we find the following: "A common way of dealing with these problems is to constrain the upper triangular part of Λ to be zero and to assume that main diagonal elements of Λ are strictly positive, i.e. $\Lambda_{jj} > 0$ for all $j = 1, \ldots, k$, see for example Geweke and Zhou (1996a). This constraint simultaneously prevents factor rotation and identifies the sign of the elements of z_i and Λ, however it is generally too restrictive. It induces an order dependence among the responses and makes the appropriate choice of the first r response variables an important modelling decision (Carvalho, Chang, Lucas, Nevins, Wang, and West, 2008a). Well-known difficulties arise if one of the true factor loadings Λ_{jj} is equal to or close to 0, see for example Lopes and West (2004)."

Classical Inference for Factor Analysis

The standard way to estimate the parameters of a GLLVM is with the Expectation-Maximisation (EM) algorithm. This algorithm exploits the fact that it is easy to compute the MLE of the parameters if we observed the values of the latent variables, often in closed form. The EM is an iterative algorithm that alternates between inferring the missing values of the latent variables (step E), and then optimising the likelihood of the parameters given the latent values that were just filled in (step M). By interweaving these two kind of steps at each iteration the EM algorithm monotonically increases the log-likelihood of the observed data, until it reaches a local maximum, see Murphy (2012) section 11.4.7 for a proof. On the bright side EM is often the fastest method available, especially if we can make use of closed form solutions for the steps. On the other hand, it often converges to local maximum, not a global one.

Marginalising out the latent variables to write down the likelihood for GLLVMs is sometime possible. Moustaki and Knott (2000) work out the general gradient formulas for the continuous latent variables case where the distribution of the observed variable belongs to the exponential family. For some special cases of these models there are closed formulas; for all the rest the authors recommend a Newton-Raphson method to approximate the necessary quantities. For a contrast between the two methods discussed so far for the probit regression model see section 9.4.1 (gradient based method) and section 11.4.6 (E-M formulas) in Murphy (2012).

A notable LVM is one where both observed and latent variables are Gaussian. A common objective for such models is to decompose the observable variables by expressing them in terms of the latent factors. We can constrain the covariance matrix of the factors according to research objective. For example we can enforce the factors to explain all the dependence between the observable variables by constraining the covariance matrix to be diagonal (factors are independent). Two common analyses along these lines are the PCA and PPCA and those can be fit in a variety of ways, including with EM. For mode details see chapter 12 in Murphy (2012), Tipping and Bishop (1999), and section 14.7.1 in Friedman et al. (2001).

Bayesian Inference for Factor Analysis

Bayesian inference for factor analysis has mainly been based on Markov Chain Monte Carlo methods. For continuous data and continuous latent variables the posterior distribution formulas have been derived in previous work (Lee, 2007; Arminger and Muthén, 1998). The methodologies available are typically based on Gibbs sampling, using the derived conditional distributions. User-friendly implementation of these methods are available in a few commercial and open source projects such as 'Bugs', 'WinBugs' and 'Mplus', and 'blavaan'.

For binary data and continuous latent variables Bayesian inference is more challenging (Lee, Song, and Cai, 2010). Cowles (1996) proposed multivariate extensions of the algorithm by for estimating univariate probit regression model and Albert and Chib (1993) proposed a similar methodology. For the multivariate probit model, Patz and Junker (1999a) proposed a Metropolis within Gibbs approach that used the conditional posterior distributions for the parameters that are available, and a Metropolis step for the rest. MCMC Gibbs for the probit model was also developed by Chib and Greenberg (1998a) and then also by Ansari and Jedidi (2000), Lee and Song (2003), and Talhouk, Doucet, and Murphy (2012). All these schemes use conjugate priors have been developed in previous papers (Lee and Song, 2010). Currently such schemes are available for the probit model but not for the logit. Ordinal data with continuous latent variables are studied as an extension of the probit model (Johnson and Albert, 2006). Implementation for Latent Trait models is rare, with 'Mplus' providing a limited number of models based on the probit link only. For an example of a Gibbs sampler for latent trait models with binary data see Geweke and Zhou (1996b) and Lopes and West (2004).

Bayesian Model Choice and Priors

Bayesian theory suggests that we should choose the model with the highest marginal likelihood, see for example section 8.4.2 in Skrondal and Rabe-Hesketh (2004) for a relevant discussion. As an alternative to computing the marginal likelihood it is common to use approximations such as AIC or BIC. In addition to the potential difficulties involved in the computation of the

marginal likelihood, another issue that needs to be addressed is the sensitivity on priors. We turn our attention to that problem next. We illustrate this point continuing with the example of the IRT model from equation (1.2).

A latent variable model involving N individuals, p binary items and k factors, entails $(N+p)\times k+p$ parameters. Naturally, all parameters are considered a-priori independent and the prior has the general structure

$$\pi(\theta,\mathbf{z})=\prod_{i=1}^{N}\prod_{\ell=1}^{k}\pi(z_{i\ell})\times\prod_{j=1}^{p}\pi(\alpha_j)\times\prod_{j=1}^{p}\prod_{\ell=1}^{k}\pi(\Lambda_{j\ell})\ . \tag{1.4}$$

In classical factor analysis models it is common to assume the latent variables the latent variables are typically assumed to be a-priori distributed as independent standard normal distributions, that is

$$\pi(z_{i\ell})=N(0,1). \tag{1.5}$$

With respect to the item parameters, four criteria were considered in the construction of the corresponding priors. In particular, the item priors should

a) be low informative to express prior ignorance or indifference,

b) impose constraints in order to achieve unique solution,

c) be suitable for Bayesian model comparison, in a way that is takes into consideration the Lindley-Bartlett paradox (Lindley, 1957; Bartlett, 1957), and

d) be easily generalisable to other members of the GLLVM.

Those four criteria will now be discussed in detail for the multidimensional logistic factor model for binary data. For the normal ogive model which implements the probit response function, conjugate priors exist, which facilitate the implementation of the Bayesian approach. In particular, normal priors are used for the α parameter, truncated normal priors for the Λ parameter (restricted to be positive) and finally beta priors are implemented for the guessing

parameter, in the 3-PL model see for instance Sahu (2002). The positivity constraint is imposed to address the rotation problem in the uni-dimensional case. Similar priors are used in models with a multilevel structure (Fox and Glas, 2001; Janssen et al., 2000; Glas and Meijer, 2003; Beguin and Glas, 2001; Sheng, 2008).

In the case of the logistic IRT models, there are no priors available leading to conjugate forms. The effect of priors on the parameter estimation of the logistic IRT models is assessed in Gifford and Swaminathan (1990). Beguin and Glas (2001) also examine the effects of different prior distributions on parameter recovery. Typically normal priors, $N(0, \sigma^2_{\alpha_j})$, are used for the α parameter and $LN(0, \sigma^2_{\Lambda_j})$ for the Λ parameter in the univariate case, with large prior variances to express lack of prior information (Sinharay, 2005a, 2006; Kang and Cohen, 2007; Kim and Bolt, 2007; Patz and Junker, 1999a, 1999b).

Sequential Monte Carlo

Sequential Monte Carlo refers to a wide network of inferential techniques that are used broadly in the sciences and engineering fields. This field was given birth with the discovery of Particle Filters in 1993 (Gordon, Salmond, and Smith, 1993) and the subsequent study of the "filtering problem". Informally, for a time series of discrete observations y_t and a related time series of latent state z_t, the filtering problem refers to the question of inferring z_t from noisy observations of y_t. One of the most celebrated solutions to this problem was given by Kalman (1960) for the case of linear Gaussian state-space models. Since their introduction, Particle Filters research focused on expanding the inferential framework around state-space models and solving more general cases of the filtering problem for non-linear and non-Gaussian systems.

The model parameters governing the relationship between observations and latent space in traditional state-space models were considered static and known, and the focus was on learning the state variables. The existing parameter estimation techniques based on Markov Chain Monte Carlo, were not easily compatible to the sequential framework. One of the most influential works was the introduction of Particle Markov Chain Monte Carlo frameworks for sequential inference by Doucet, De Freitas, Gordon, et al. (2001). That line of research was developed in

parallel to that of Particle Filters' research until more recently when the two merged, becoming known as Sequential Monte Carlo methods.

The goal of a sequential scheme is to recursively explore the sequence of posterior distributions

$$\pi_0(\theta) = p(\theta), \quad \pi_t(\theta) = p(\theta|y_{1:t}), \quad t \geq 1 \tag{1.6}$$

where θ denotes all the unknown parameters in our model. The sequential inference paradigm approximates recursively these posterior distributions and the model evidence $p(y_{1:t})$ (Del Moral, Doucet, and Jasra, 2006).

Although the inclusion of latent variables makes factor modelling a natural fit for SMC methods, this area remains unexplored to the best of our knowledge. In Factor Analysis when it is possible to marginalise out the latent variables the natural sequential scheme of choice would be Iterative Batch Importance Sampling (IBIS) (Chopin, 2002). For the rest of the cases, which are the majority, inference in factor models is a harder problem. For example the IBIS algorithm is not directly applicable for factor models for categorical data, where the latent variables cannot be integrated out. The main framework for sequential Bayesian inference of latent variable models is the SMC^2 scheme of (Chopin, Jacob, and Papaspiliopoulos, 2012) which combines previously available methodologies of IBIS and the pseudo-marginal framework of Andrieu and Roberts (2009), and applies to cases of Markov dependent latent variables. While we recognise the contributions of this research work, as we explain later on, we find that the needs of Sequential Factor Analysis are not adequately met by any of the existing frameworks.

Sequential Monte Carlo research is a powerful network of techniques that has been successfully applied in a variety of tasks. Although it has not been extensively used in the study of factor models, we find that it can provide solutions to a number of areas of active research including identifiability issues around inference, model assessment in confirmatory analysis, and determining the right number of factors. Furthermore, the sequential paradigm opens up the possibility of dynamic or online inference for factor analysis. Separately, the sequential paradigm can be particularly useful in experimental designs and alleviate issues of the traditional hypothesis

testing framework.

Chapter 2

Generalised Bayesian Structural Equation Models

2.1 Introduction

Structural Equation Modelling (SEM) is a general framework for testing research hypotheses arising in psychology and social sciences in general. It is used to model directional (regression coefficients) and non-directional (correlations) relationships among latent variables (structural model) which are identified by observed variables (measurement model) (Bollen, 1989). Initial inference methods for SEM have mostly been frequentist, but recently their Bayesian counterpart has gained popularity (see e.g. Scheines et al., 1999a; Dunson et al., 2005; Kaplan, 2014; Merkle and Rosseel, 2015; Van De Schoot et al., 2017). Bayesian SEM offers several potential advantages. It has been shown to perform well with small sample sizes (Depaoli and Clifton, 2015) or small number of groups in hierarchical modelling (Hox et al., 2012), and can help resolve issues of inadmissible estimates (Can et al., 2015). Moreover, it provides computationally efficient schemes for models with large numbers of latent variables (Lüdtke et al., 2013; Oravecz et al., 2011), and can quantify uncertainty via credible intervals of quantities that are functions of parameters such as reliability (Geldhof et al., 2014) or indirect effects (Yuan and MacKinnon, 2009). It can also provide a unified framework for handling missing data (Lee and

Xia, 2008) and semi-parametric models (e.g. Yang and Dunson, 2010; Song et al., 2013).

In this paper, we focus on the Bayesian SEM framework introduced by Muthén and Asparouhov (2012). Structural equation models impose some kind of restrictions on the number of parameters to be estimated. 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 the approximate zero framework was intended to address a number of issues with classical SEM. First and foremost, it aimed to correct the fact that classical null-hypothesis testing framework over-rejects valid hypothesised structures due to differences that, from a research point of view, may be small, negligible, or irrelevant. Second, it provided researchers with a modelling framework that can be used as an alternative, or additional, diagnostic of goodness of fit and help investigate what aspects of the model are causing it to have bad fit. Third, by virtue of being a Bayesian approach, it also alleviates estimation issues common under the maximum likelihood framework, such as non-covergence or Heywood cases. Using such informative priors, instead of exact zero assumptions, 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). However, the approximate zero framework offers several advantages, see for example the discussion in Stromeyer et al. (2015), and its implementation is possible for the case of continuous and normally distributed data in the Mplus package (Muthén and Muthén, 2017). 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 (Muthén and Asparouhov, 2012; Asparouhov, Muthén, and Morin, 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. To our knowledge, the approximate

zero approach offers the only Bayesian alternative to the modification indices. In theory, one can approach the same problem via Bayesian model searching using e.g. spike and slab priors and stochastic search variable selection but this is a substantially more challenging approach; see for example Lu et al. (2016).

In this paper, we aim to improve upon two aspects of the approximate zero framework. First, the existing framework covers models for continuous data and relies on the normality assumption. Here, we propose a generalised approximate zero framework that can accommodate more distributions, such as the logistic which is commonly employed in item response theory (IRT). This is achieved by introducing in the measurement model item-individual random effects. The model by Muthén and Asparouhov (2012) becomes then a special case of the proposed framework.

Second, we focus on the task of assessing model fit under the approximate zero framework (Garnier-Villarreal and Jorgensen, 2020; Asparouhov and Muthén, 2020). Several approaches exist in the literature, with posterior predictive p-values (PPP) (Meng, 1994a) being the most widely used. However, concerns have been raised regarding their suitability in this framework (see e.g. Stromeyer et al., 2015; Hoijtink and van de Schoot, 2018a). Special consideration has to be given to the choice of prior distributions for the model parameters, which can potentially affect the PPP performance (MacCallum et al., 2012; Van Erp et al., 2018; Liang, 2020). 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 and 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.

Our approach aims towards reaching 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). Our 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; see, for example, MacCallum et al. (1992), Browne (2000) and recommendation 4 of Stromeyer et al. (2015). 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) use the DIC and WAIC indices that can be viewed as approximate versions of cross-validation (Gelman et al., 2014), although their resulting approximation is not always satisfactory; see for example Plummer (2008). 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 and Holmes, 2020). Our proposed framework is developed by combining fit and out of sample predictive performance indices from different models.

The paper is structured as follows. In Section 2, we define the proposed generalised Bayesian SEM framework, illustrate how existing Bayesian SEM formulations are special cases, and provide examples of new models. Section 3 introduces the framework for assessing Bayesian SEM models. In Section 4, we illustrate and assess our methodology through several simulation experiments. Section 5 presents the analyses of two real applications: the first is a standard example on examining the 'Big 5' personality factors on data from the 2005-06 British Household Panel Survey (BHPS), whereas the second example is on the Fagerstrom Test for Nicotine Dependence (FTND). Finally, Section 6 concludes with some relevant discussion and extensions. The code for this work is available in the accompanying repository 'bayes-sem' hosted on github[1].

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

2.2 Generalised framework for Bayesian SEM

2.2.1 Model specification

We use a unified Bayesian framework that encompasses models for categorical and continuous responses. Suppose there are p observed variables (items) denoted by $\mathbf{y} = (y_1, \ldots, y_p)$ and that their associations are explained by k continuous latent variables (factors) denoted by $\mathbf{z} = (z_1, \ldots, z_k)$. 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:

$$\mathbf{y}_i^* = \alpha + \Lambda \mathbf{z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{2.1}$$

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

For binary data, the connection between the observed binary variable y_j and the underlying variable y_j^* is $y_j = \mathcal{I}(y_j^* > 0)$. Similarly, for an ordinal variable with m_j categories, $y_j = a$ if $\tau_{a-1}^{(j)} < y_j^* \leq \tau_a^{(j)}, a = 1, \ldots, m_j$ where $\tau_0^{(j)} = \infty, \tau_1^{(j)} < \tau_2^{(j)} < \ldots < t_{m_j-1}^{(j)}, \tau_{m_j}^{(j)} = +\infty$. More specifically, for a binary item j and individual i, the probability of success (positive) response is given by:

$$P(y_{ij} = 1 \mid \mathbf{z}_i) = P(y_{ij}^* > 0 \mid \mathbf{z}_i) = P(\alpha_j + \Lambda_j \mathbf{z}_i + \epsilon_{ij} > 0 \mid \mathbf{z}_i) = P(\epsilon_{ij} < \alpha_j + \Lambda_j \mathbf{z}_i \mid \mathbf{z}_i) = F(\alpha_j + \Lambda_j \mathbf{z}_i), \tag{2.2}$$

where F stands for the cumulative distribution function (CDF) of ϵ_{ij}. Finally, the model

becomes:

$$F^{-1}\{P(y_{ij}=1)\} = \alpha_j + \Lambda_j \mathbf{z}_i, \tag{2.3}$$

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:

$$\boldsymbol{\epsilon}_i \sim \begin{cases} N(0_p, \Psi), \quad \Psi = \text{diag}(\psi_1^2, \ldots, \psi_p^2), & \text{if } \mathbf{y}_i \text{ is continuous} \\ N(0_p, \Psi), \quad \Psi = I_p, & \text{if } \mathbf{y}_i \text{ is binary \& } F^{-1} \text{ is the inverse CDF of the normal} \\ \prod_{j=1}^{J} \text{Logistic}(0, \pi^2/3), & \text{if } \mathbf{y}_i \text{ is binary \& } F^{-1} \text{ is the inverse CDF of the logistic,} \end{cases} \tag{2.4}$$

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 case of continuous or categorical items with the probit link, the marginal distribution of $\mathbf{y}_i^*$ is:

$$\mathbf{y}_i^* \sim N(\alpha, \Lambda\Phi\Lambda^T + \Psi). \tag{2.5}$$

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

The model defined in (2.1) and (2.4) 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 Λ and Φ. CFA is a method used to verify the factor structure of a set of observed variables. This is achieved by setting several elements of Λ to zero that are referred to as cross-loadings. EFA, on the other hand, uses a much more flexible Λ by only placing identifiability restrictions on it, and sets $\Phi = I_k$. An assumption, which is common to both approaches, is the conditional independence of the variables given the factors. This is equivalent to setting the off-diagonal terms in the covariance of the $\boldsymbol{\epsilon}$, also known as error correlations, to zero.

2.2.2 Generalised Bayesian model framework

The Bayesian SEM approach introduced in Muthén and Asparouhov (2012) (approximate zero framework), mostly covers continuous items, and it can potentially be extended to binary and ordinal data under the probit specification.

The AZ model was introduced in (Muthén and Asparouhov, 2012) to relax the exact zero conditions so that it better reflects a hypothesised substantive theory and better serves the goal of confirming it. These relaxation of the exact zero assumption was used in two aspects of the model, the cross-loadings and the error correlations.

The relaxation of the cross-loadings is done by freeing up the cross-loadings that would be otherwise set to zero, but constraining them near zero by assigning them a highly informative prior distribution around zero.

The relaxation of the error correlation assumption, also known as local independence, can be done in a few different ways. One way is via the covariance matrix according to which the residual errors are distributed. Under the exact zero assumptions, that matrix is a strictly diagonal matrix. Under the approximate zero framework that matrix can be assumed to be 'approximately' diagonal, by assigning it an informative prior distribution with a lot of mass around a diagonal matrix. One common option for such a distribution is the Inverse Wishart using a diagonal matrix and high degrees of freedom. Another way is using the model we propose here, a more general specification that includes the logistic model for categorical data and for which the model by Muthén and Asparouhov (2012) is a special case.

Model (2.1) is extended by adding an item-individual specific random effect u_{ij} giving

$$\mathbf{y}_i^* = \alpha + \Lambda \mathbf{z}_i + \mathbf{u}_i + \mathbf{e}_i, \tag{2.6}$$

where $\boldsymbol{\epsilon}_i$ in (2.1) is split into $\mathbf{u}_i$, a p-dimensional vector of random effects with a non-diagonal covariance matrix Ω, and $\mathbf{e}_i$, an error term with a diagonal covariance matrix Ψ^*. The $\mathbf{u}_i$ terms aim to capture associations among the items beyond those explained by the vector of

latent variables $\mathbf{z}_i$. Those associations can be due to question wording, method effect, etc. Furthermore, the item-individual specific random effects u_{ij} can be seen as an additional residual term that provides model diagnostics information for the detection of two-way outliers (e.g. leaked items and cheating behaviour in educational testing or secondary response strategies employed by some of the respondents to some of the items). Moreover, we note that now, contrary to equations (2.1) and (2.4), the cross loadings Λ in (2.6) and (2.7) are non-zero parameters that are assigned informative priors centred around zero, e.g. $N(0, 0.01)$.

For continuous normally distributed data, Model (2.6) becomes the model proposed in Muthén and Asparouhov (2012) written as

$$\mathbf{y}_i^* \sim N(\alpha, \Lambda\Phi\Lambda^T + \Omega + \Psi^*), \quad i = 1, \ldots, n, \tag{2.7}$$

where $\mathbf{u}_i \sim N(0, \Omega)$, $\mathbf{z}_i \sim N(0, \Phi)$ and $\mathbf{e}_i \sim N(0, \Psi^*)$.

Under this new model defined in equations (2.6) and (2.7), we can relax the local independence assumption by introducing the item-individual random effect u_{ij}, that aim to capture associations between the items beyond those explained already by the latent variables. These random effects have to be small in magnitude and in particular have to be smaller than the rest of the residual error e because they are meant to capture correlation among the items, not residual errors of individual items. Furthermore, the error correlations need to be small for the model to stay close to hypothesised theory and for identifiability reasons, as mentioned before. Hence, the prior distribution of the u should be concentrated around zero to degree that ensures that u are smaller than e, and have a non-diagonal covariance matrix. In particular, a Normal distribution $N(0, \Omega)$ would work, where Ω is a full covariance matrix with an appropriate informative prior. As mentioned earlier, it is essential that the overall amount of error correlations is not substantial and this can be achieved by ensuring that the impact of Ω is low compared to Ψ^*. One approach to make this more specific is to use an estimate of diagonal matrix Ψ^* under the same model without error correlations, and then set the prior of Ω to favour low values for its diagonal elements compared to the estimate. For example, in the applications considered in this paper, the elements of the Ψ matrix were all estimated to be relatively closed to one.

The Inverse Wishart distribution with identity scale matrix and $p + 6$ degrees, also used in (Muthén and Asparouhov, 2012), may therefore provide a reasonable choice. Under this prior, the diagonal elements of Ω have mean 0.2 and standard deviation 0.163 (see appendix A for explicit formulae for the prior mean and variance and note that p cancels out), hence the prior probability mass is concentrated well below the Ψ estimates.

Note that in this case the residual errors e are assumed to be distributed according to the covariance matrix Ψ^*, which is constrained to be diagonal, so that all correlation among the errors are channelled through the random effect u_{ij}.

As it will be discussed in Section 2.2.3, it is essential to assign an informative prior on Ω. A reasonable choice that favours diagonal matrices is the Inverse Wishart, which is also used in Muthén and Asparouhov (2012). This choice of prior can be thought of as controlling the magnitude of model flexibility the researcher is willing to allow for capturing the effects of external factors on measurement, such as question wording. Hence, it is important when setting this prior to ensure that the $\mathbf{u}_i$s are of lower magnitude than the $\mathbf{e}_i$s.

The generalised framework of (2.6) provides several extensions. It is now possible to define the approximate zero model for logistic models by assuming $e_{ij} \sim \text{Logistic}(0, \pi^2/3)$ (see Section 2.2.2.2 for details). Other distributions (e.g. t-distribution, non-Normal) can also be assumed for $\mathbf{e}_i$ and $\mathbf{u}_i$s. Setting $\Phi = I_k$ in (2.6) leads to the EFA model, nevertheless fitting such a model with MCMC may be challenging as we discuss later on; see also (Lopes and West, 2004; Erosheva and Curtis, 2017; Frühwirth-Schnatter and Lopes, 2018; Conti, Frühwirth-Schnatter, Heckman, and Piatek, 2014) for some relevant Bayesian EFA schemes.

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

2.2.2.1 Model and priors for continuous normally distributed data

The model in (2.6) originates from the specification below:

$$\begin{cases} \mathbf{y}_i = \alpha + \Lambda \mathbf{z}_i + \mathbf{u}_i + \mathbf{e}_i \\ \mathbf{z}_i \sim N(0, \Phi) \\ \mathbf{e}_i \sim N(0, \Psi^*) \\ \mathbf{u}_i \sim N(0, \Omega) \end{cases} \tag{2.8}$$

Nevertheless, as mentioned earlier, non-Normal distributions can be assigned on $\mathbf{e}_i$s, $\mathbf{u}_i$s and even $\mathbf{z}_i$. In the case where all these are assumed to be Normal, the following augmentation is also equivalent:

$$\begin{cases} \mathbf{y}_i \mid \mathbf{u}_i \sim N\big(\alpha + \mathbf{u}_i,\, \Lambda\Phi\Lambda^T + \Psi^*\big) \\ \mathbf{u}_i \sim N(0, \Omega). \end{cases} \tag{2.9}$$

Regarding priors, we begin with Ω, the non-diagonal covariance matrix that introduces the error correlations. As mentioned earlier, it is essential that the overall amount of error correlations is not substantial and this can be achieved by ensuring that the impact of Ω is low compared to Ψ^*. One approach to make this more specific is to use an estimate of diagonal matrix Ψ under the same model without error correlations, and then set the prior of Ω to favour low values for its diagonal elements compared to the estimate. For example, in the applications considered in this paper, the elements of the Ψ matrix were all estimated to be relatively closed to one. The Inverse Wishart distribution with identity scale matrix and $p + 6$ degrees, also used in (Muthén and Asparouhov, 2012), may therefore provide a reasonable choice. Under this prior, the diagonal elements of Ω have mean 0.2 and standard deviation 0.163 (see appendix A.1 for explicit formulae for the prior mean and variance and note that p cancels out), hence the prior probability mass is concentrated well below the Ψ estimates.

Regarding Λ and Φ there are generally two parametrisations to ensure identifiability. Under the first one, Φ is a full covariance matrix and the leading loadings in Λ for each factor are

set to one. In this case, an Inverse Wishart prior with relatively low amount of information, compared to the prior of Ω, is sought for Φ, e.g. the Inverse Wishart with the identity as the scale matrix and $p+4$ degrees of freedom or lower. Under the second formulation, the leading loadings in Λ for each factor are just constrained to be positive and Φ is a correlation matrix. It is also possible to remove these positivity constraints and therefore assign Normal priors to all of the elements of Λ; this formulation may be viewed as a special case of the parameter expansion suggested in Ghosh and Dunson (2009) for EFA. To ensure identifiability under this formulation, post-processing should be applied on the MCMC output. More specifically, the posterior samples of the columns corresponding to each factor should be multiplied by -1 if the relevant leading loading is negative, otherwise they are left as they are. In our experience, this option results in more efficient MCMC performance, in terms of mixing and convergence, when running the model in Stan. Regarding the correlation matrix Φ, the LKJ prior, introduced in Lewandowski et al. (2009), can be used.

The variances of the Normal priors assigned on the elements of Λ depend on whether these are regarded as cross-loadings or free parameters according to the hypothesised model. The cross loadings are assigned Normal distributions with zero mean and a variance of 0.01 as in Muthén and Asparouhov (2012), whereas the remaining parameters of Λ require some extra attention. A frequently used option is to assign large variance Normal priors, but this can lead to issues such as Lindley's paradox (Lindley, 1957). One way to guard against such problems is to use unit information priors (Kass and Wasserman, 1996). The main idea behind unit information priors is to avoid the very large prior variances causing the paradox, by setting them so that they correspond to information from a single observation point. Lopes and West (2004) and Ghosh and Dunson (2009), in the context of EFA, recommend the following unit information priors:

$$\Lambda_{ij} \sim N(0, \psi_j^2) \tag{2.10}$$

where ψ_j^2 are the idiosyncratic variances of the diagonal matrix Ψ that are treated as unknown parameters. Note, however, that the above priors may cause problems in cases where the ψ_j^2s are quite small as it is essential to differentiate from the prior variance of 0.01 used for the cross

loadings. For this reason, a fixed value may be used instead for the prior variance of the free elements of Λ, based on preliminary estimates of them, or even the value of one if the items are on similar scales.

Regarding the diagonal matrix Ψ, 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 hyper-parameters of these Inverse Gamma priors are set in a way so that Heywood cases are given very small prior weight. More specifically, the prior given to 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 that the researcher can choose in order to limit the probability of running into Heywood issues that arise when

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

Following Frühwirth-Schnatter and Lopes (2018); Conti et al. (2014), the constant c_0 can be chosen such that the prior probability of the event above is quite small. In the data considered in this paper, the value of $c_0 = 2.5$ was chosen on that basis. This is a data-dependent prior but the impact incorporates a minimal amount of information and it also helps avoid identification and MCMC convergence issues that are associated with Heywood problems. To back this up we also conducted a sensitivity analysis that is presented in appendix A.2. The results using the chosen data-dependent prior were practically identical with those obtained using several data-independent priors.

Finally, large variance Normal priors are assigned on the α parameters. In every analysis that follows we use the following wide prior Normal, $\alpha \sim N(0, 10^2)$.

2.2.2.2 Model and priors for binary and ordinal data

The model for binary data using the underlying variables $y^*_{ij}, (j = 1 \ldots, p)$ can be written as

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

In the above models the $\mathbf{e}_i$s correspond to the logistic and probit specifications that are the most frequently used models, although other choices of distributions are also possible. The above expressions may be simplified by integrating out the $\mathbf{e}_i$s and obtain

$$\begin{cases} \mathbf{y}_i \sim \prod_{j=1}^p \text{Bernoulli}\big(\pi_{ij}(\eta_{ij})\big) \\ \pi_{ij}(\eta_{ij}) = \sigma(\eta_{ij}) \;\text{ or }\; \pi_{ij}(\eta_{ij}) = \Phi(\eta_{ij}), \;\; \eta_{ij} = [\boldsymbol{\eta}_i]_j \\ \boldsymbol{\eta}_i \coloneqq \alpha + \Lambda \mathbf{z}_i + \mathbf{u}_i, \\ \mathbf{z}_i \sim N(0, \Phi) \\ \mathbf{u}_i \sim N(0, \Omega) \end{cases} \tag{2.11}$$

where $\sigma(\cdot)$ denotes the sigmoid function and leads to the logit model, whereas $\Phi(\cdot)$ denotes the cumulative density function of the standard Normal distribution and leads to the probit model. Note that the distribution of $\mathbf{u}_i$s, and even $\mathbf{z}_i$s, need not be Normal under the framework, this was only done for exposition purposes. In the cases where $\mathbf{u}_i$s are indeed assumed to be Normal, the amount of data augmentation can be reduced further by the following equivalent

formulation

$$\begin{cases} \mathbf{y}_i \sim \prod_{j=1}^{p} \text{Bernoulli}\big(\pi_{ij}(\eta_{ij})\big) \\ \pi_{ij}(\eta_{ij}) = \sigma(\eta_{ij}) \;\text{ or }\; \pi_{ij}(\eta_{ij}) = \Phi(\eta_{ij}) \\ \boldsymbol{\eta}_i \sim N(\alpha,\, \Lambda\Phi\Lambda^T + \Omega). \end{cases} \tag{2.12}$$

In the simulation experiment and real-world examples the formulations of (2.11) and (2.12) 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 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 as

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

Similarly to the binary case, to model an ordinal observed variable y_j with m_j categories, we assume the existence of an underlying continuous variable y_j^* so that $y_j = a$ if $\tau_{a-1}^{(j)} < y_j^* \leq \tau_a^{(j)}, a = 1, \ldots, m_j$.

The multinomial model is assumed to be:

$$y_{ij} \sim \prod_{s=1}^{m_j} \pi_{j,s}(\boldsymbol{\eta})^{y_{j,s}}$$

where $y_{j,s} = 1$ if the response y_{ij} is in category s and 0 otherwise, $\pi_{ij,s}(\eta_{ij}) = (\gamma_{ij,s}(\eta_{ij}) - \gamma_{ij,s-1}(\eta_{ij}))$ and $\gamma_{ij,s}(\eta_{ij})$ is a cumulative probability of a response in category s or lower to item

y_j. Furthermore,

$$\begin{cases} \mathbf{y}_i \mid \boldsymbol{\eta}_i \sim \prod_{j=1}^{p} \text{Multinomial}(\pi_{ij}(\eta_{ij})) \\ \gamma_{ij}(\eta_{ij}) = \sigma(\eta_{ij}) \ \text{ or } \ \gamma_{ij}(\eta_{ij}) = \Phi(\eta_{ij}) \\ \boldsymbol{\eta}_i = \tau + \Lambda \mathbf{z}_i + \mathbf{u}_i, \\ \mathbf{z}_i \sim N(0, \Phi) \\ \mathbf{u}_i \sim N(0, \Omega). \end{cases} \tag{2.13}$$

The parameters τ are unknown parameters also referred to as 'cut-points' on the logistic, probit or other scale, where $\tau_0^{(j)} = \infty, \tau_1^{(j)} < \tau_2^{(j)} < \ldots < t_{m_j-1}^{(j)}, \tau_{m_j}^{(j)} = +\infty$.

Similar priors can be assigned as in the case of continuous data. Regarding the elements of the Λ matrix that are not approximate zero, unit information priors can be used. In the case of the 2PL IRT model this translates to a $N(0, 4)$ prior (Vitoratou et al., 2014).

2.2.3 Overview of the models and their estimation procedure

In this Section, we highlight some models, within the framework defined so far, that are essential for the methodology developed in this paper. We then provide details and discussion regarding their implementation. These models are defined below:

- The exact zero (EZ) model. This is the standard structural equation model and provides the starting point in the analysis considered here. It is defined by equations (2.1) and (2.4) with the cross-loadings in Λ being fixed to zero.

- The approximate zero (AZ) model. This is the model first introduced in Muthén and Asparouhov (2012) and generalised in this paper. In its general form it is defined in equation (2.6). In the case of normally distributed $\mathbf{u}_i$s, $\mathbf{e}_i$s and $\mathbf{z}_i$s, is simplified to (2.7). An important feature is that the cross-loadings in Λ are no longer being fixed to zero. It is a model to be used only in the Bayesian sense, as the informative priors on the Ω and on the cross-loadings in Λ are essential.

- The exploratory factor analysis model (EFA). It is the standard EFA model, defined here by equations (2.1) and (2.4) where low informative priors are assigned to all the components of Λ and $\Phi = I$.

- The EFA model with item-individual random effects (EFA-C). It is defined as the EFA model but with equation (2.6) instead of (2.4). This approach to EFA allows for a small amount of item dependencies conditional on the extracted independent factors. That model specification might result in greater amount of dimension reduction than EFA, since the stricter assumption of conditional independence could require a model with additional factors.

In terms of implementation, it is generally possible to use MCMC and several schemes can be used, (see e.g. Edwards, 2010). In cases where the $\mathbf{e}_i$s, $\mathbf{u}_i$s and $\mathbf{z}_i$s are all assumed to be Normal, Gibbs samplers may be formed, (see e.g. Geweke and Zhou, 1996b; Chib and Greenberg, 1998b), and the model may also be fitted with standard software such MPlus. Nevertheless, if any of these are assumed to be non-Normal, e.g. logistic models, different software and MCMC algorithms are needed. In this paper, we recommend the use of Hamiltonian Monte Carlo (HMC) (Neal, 2011), as it covers all cases. HMC is a Markov Chain Monte Carlo (MCMC) technique for Bayesian inference that updates all the parameters simultaneously. It utilises information from the gradient of the log-posterior via the Hamiltonian equations in order to provide an efficient Markov chain with good mixing and convergence properties. The user is referred to Neal (2011) for more details. HMC can be implemented with the help of programming frameworks such as Stan (Carpenter et al., 2017). It is supported by high-level software packages such as 'blavaan' (Merkle and Rosseel, 2015), PyStan or RStan which are the Python and R language interfaces of the Stan language respectively. In this work, we chose to implement all model inference using HMC and the Stan language to take advantage of the generality of the HMC methodology and the software ecosystem built around the Stan language. For a complete repository of the code and further implementation details we refer the interested reader to the code repository for this work hosted on github at 'bayes-sem'[2]. We

[2]https://github.com/bayesways/bayes-sem/

note of course that it is possible to construct tailor made MCMC schemes for specific models by taking advantage of potential simplifications in their structure; the Gibbs sampler for the fully Normal case provides such an example.

Fitting the EZ model in Stan is generally straightforward although we note that it may be useful to consider different parametrisations to improve MCMC performance and stability. For example, one may set the leading loadings in Λ to one and consider a full covariance matrix Φ or just restrict the leading cross loadings to be positive and consider a correlation matrix for Φ.

While the EZ model is identifiable both under the frequentist and Bayesian framework, this is not the case for the AZ model that, as mentioned earlier is to be approached in a Bayesian manner. The AZ model was introduced in Muthén and Asparouhov (2012) to relax the exact zero conditions so that it better reflects a hypothesised substantive theory and better serves the goal of confirming it. The cross-loadings or error correlations that would be constrained to zero under the EZ model, become free parameters in the AZ model but with highly informative priors centred at zero. Usually, there is information in the data to identify some of those parameters but not all of them. Hence, if all those parameters were freed and a frequentist model was adopted one would run into identifiability issues. As we typically do not know which of these parameters to free, an alternative is to adopt the AZ model, where identifiability is less of a concern given the informative priors that are concentrated around zero. Hence, AZ models essentially satisfy two goals: i.) the model stays close to the substantive hypothesis, only replacing the exact zero assumptions, used in traditional SEM modelling, with approximate zero, and ii.) it protects against any identifiability issues since the prior contains enough information to guide the inference algorithm to completion.

In cases where the EZ model does not perform well it is essential to find an appropriate benchmark to assess the performance of the AZ model. As discussed in more detail in the next section, such benchmarks can be provided by the EFA and EFA-C models. In general, fitting EFA models using MCMC can be a challenging task, due to issues such as rotational indeterminacy. The problem lies in the fact that the likelihood is specified in terms of $\Lambda\Lambda^T$ but

often interest lies instead on Λ. The lower triangular set of restrictions (see e.g. Geweke and Zhou, 1996b) ensures the mapping between those matrices is well defined, but introduces order dependence among the observed variables. The choice of the first k variables, which is an important modelling decision (Carvalho, Chang, Lucas, Nevins, Wang, and West, 2008b), thus becomes influential. The schemes of Conti et al. (2014); Frühwirth-Schnatter and Lopes (2018); Bhattacharya and Dunson (2011) provide an alternative to setting these restrictions and can also be used to identify the number of factors in a single MCMC run. However, as also noted in Bhattacharya and Dunson (2011), for a number of tasks such as choosing the number of factors or assessing the predictive performance, there is no need to focus on Λ, but on $\Lambda\Lambda^T$ instead which is free of rotational issues. In such cases, the restrictions on Λ can be omitted as long as there are no MCMC convergence and mixing issues on the $\Lambda\Lambda^T$ elements. As described in the next section, EFA and EFA-C models are only used in this paper to establish a benchmark for their predictive performance, hence focusing on $\Lambda\Lambda^T$ is sufficient. It is important to note here that this does not apply for the SEM driven EZ and AZ models, where we are also interested on the Λ elements and the fit of the model. But for these models, the restrictions implied by the hypothesised SEM ensure that the elements of Λ are free of rotational issues. The number of factors of the EFA and EFA-C models can either be matched to that of the EZ model or, alternatively, models with different number of factors can be fit separately and compared. The comparison can be done by standard indices, such as the model evidence, BIC, etc., or via the model assessment framework introduced in this paper and presented in the next section. Drawing inference on Λ in the EFA context, in addition to $\Lambda\Lambda^T$, remains an interesting and challenging problem, especially in the case of binary data and the presence of item-individual random effects. But, as it is beyond the scope of the paper, it is left for future research.

2.3 Model assessment

In this section, we introduce a model assessment framework that collectively uses fit indices and cross-validation to detect overfit. The aim is to complement PPP values, or other similar indices, with scoring rules to evaluate the prediction extracted from the model. The aim is to

achieve a good fit and avoid overfit. The suggested procedure involves calculating these metrics for the EZ and AZ models as well as the EFA and EFA-C models with the same number of factors. We begin by presenting the proposed indices in detail, and finally provide our suggested procedure along with some guidelines and recommendations.

2.3.1 Assessing goodness of fit with PPP values

PPP values are perhaps the most frequently used method to assess model fit in the Bayesian SEM framework. Posterior predictive checking relies on a discrepancy function denoted by $D(\boldsymbol{Y}, \theta)$ that quantifies how far the fitted model is from the data. For continuous data, the discrepancy function used here is the likelihood ratio test (LRT) function (see e.g. Scheines et al., 1999a) comparing the estimated model (H_0 hypothesis), and the unconstrained variance-covariance matrix model (H_1 hypothesis). The unconstrained model is also known as the saturated model (perfect fit). $D(\boldsymbol{Y}, \theta)$ is given by:

$$\mathrm{LR}\left[S, \Sigma(\theta)\right] = (n-1)\left\{\log|\Sigma(\theta)| + \mathrm{tr}\left[S\Sigma^{-1}(\theta)\right] - \log|S| - p\right\}, \tag{2.14}$$

where S and $\Sigma(\theta)$ are the sample and model implied variance-covariance matrix respectively. Furthermore, $|\cdot|$, $\mathrm{tr}(\cdot)$ denote the determinant and trace of a matrix respectively. For example, if the maximum likelihood estimate (MLE) of θ, is plugged in (2.14), then $\mathrm{LR}\left[\cdot\right]$ is a statistic, but if θ is unknown then $\mathrm{LR}\left[\cdot\right]$ may be viewed as a metric. Given the discrepancy function $D(\boldsymbol{Y}, \theta_m)$ defined in (2.14), a suitable MCMC algorithm and M posterior draws, the PPP value is computed as follows:

1. At each (or some) of the MCMC samples θ_m, $m = 1, \ldots, M$, do the following:

   (a) Compute $D(\boldsymbol{Y}, \theta_m)$.

   (b) Draw $\widetilde{\mathbf{Y}}$ having the same size as $\boldsymbol{Y}$, from the likelihood function $f(\boldsymbol{Y}|\theta_m)$ of the implied model in Equation (2.5) or (2.7) and using the current value θ_m.

(c) Calculate $D(\widetilde{\mathbf{Y}}, \theta_m)$ and $d_m = \mathcal{I}\big[D(\boldsymbol{Y}, \theta_m) < D(\widetilde{\mathbf{Y}}, \theta_m)\big]$, where $\mathcal{I}\big[\,\cdot\,\big]$ is an indicator function.

2. Return PPP= $\frac{1}{M}\sum_{m=1}^{M} d_m$.

In the case of binary and ordinal data, the model is written as the probability of a response pattern. For p binary items, there are $R = 2^p$ 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 a parameter vector θ, and the assumption of conditional independence given $\mathbf{z}$ and $\mathbf{u}$ is:

$$\pi_r(\theta) = \int \prod_{j=1}^{p} \text{Bernoulli}\left\{[\mathbf{y}_r]_j | \sigma([\boldsymbol{\eta}]_j)\right\} f(\mathbf{z}) f(\mathbf{u}) d\mathbf{z} d\mathbf{u}, \tag{2.15}$$

where Bernoulli$\{y|\pi\}$ denotes the Bernoulli probability mass function for a binary observation $\mathbf{y}$ and probability of success π, $\boldsymbol{\eta}$ is as defined in Section 2.2.2, i.e. $\boldsymbol{\eta} = \alpha + \Lambda\mathbf{z} + \mathbf{u}$, and $\mathbf{z}$ and $\mathbf{u}$ are the latent components in the implied model. The integral in (2.15) can be approximated using Monte Carlo. Similar expressions can also be obtained for the probit specification.

An equivalent model can now be defined for the observed frequencies $(O_1, \ldots, O_R)$ given the model-based $\pi_r(\theta)$s via the Multinomial distribution

$$(O_1, \ldots, O_R) \sim \text{Multinomial}\left[n, \pi_1(\theta), \ldots, \pi_R(\theta)\right]. \tag{2.16}$$

In the context of PPP values, a frequently used discrepancy measure, (see e.g. Sinharay, 2005b), is the G^2 statistic given by

$$D(\boldsymbol{Y}, \theta) = \sum_{r=1}^{R} O_r \log\left(\frac{O_r}{n\pi_r(\theta)}\right). \tag{2.17}$$

For a given θ, e.g. a sample draw from the posterior, (2.17) can be derived from the likelihood ratio between the model in (2.16) and the saturated version of it where each $\pi_r(\theta)$ is replaced by O_r/n. Given M MCMC samples from the posterior, the PPP value is then calculated following the steps given above for continuous data. PPP values are not p-values and therefore are

not necessarily connected with the relevant type I error argument. Instead, they are regarded merely as fit indices. In terms of criteria on the PPP values, we follow the relevant discussion in Muthén and Asparouhov (2012). As such, the fit of a model with a PPP value around 0.5 is regarded as excellent. It is generally not clear how low a PPP value should be to warrant poor fit but usually this threshold is set to 0.1 or 0.05.

The discrepancy function used here checks the overall fit of the model. Other discrepancy functions can be used that check the fit on lower order margins. In the case of categorical data, one can compute chi-square type residuals (see e.g. Jöreskog and Moustaki, 2001) on the univariate, bivariate and trivariate margins as well as utilise the work on limited information test statistics such as the M_2 test statistic by Maydeu-Olivares and Joe (2005), which is also connected to modification indices (Oberski, van Kollenburg, and Vermunt, 2013). Those PPP values can be useful in detecting model misfit in pair or triple of items. Those discrepancies can be investigated in future research within the paper's proposed framework.

2.3.2 Scoring rules and cross validation in SEM

As mentioned in the introduction, it is essential to assess the out-of-sample predictive performance of each model considered in addition to its fit. Although prediction is not necessarily the main aim of factor analysis models, certain ideas from predictive inference can be borrowed here to help us assess model fit and overfit. The focus is on a model's ability to predict new data that was not used for estimating the model parameters. Hence, we divide individuals into two samples: i.) the training sample $\boldsymbol{Y}^{tr}$ used to estimate the model parameters, through the posterior distribution $\pi(\theta|\boldsymbol{Y}^{tr})$, and ii.) the test sample $\boldsymbol{Y}^{te}$ used to check the forecasts of the model estimated above.

More specifically, the predictions for the unseen data come in the form of a distribution $h(\boldsymbol{Y}^{te}|\boldsymbol{Y}^{tr})$ that can be contrasted as a whole against the actual test data $\boldsymbol{Y}^{te}$. In the frequentist case, one option for such a predictive distribution is $f(\boldsymbol{Y}^{te}|\hat{\theta}^{tr})$, where $f(\cdot)$ denotes the likelihood function and $\hat{\theta}^{tr}$ is the MLE obtained from $\boldsymbol{Y}^{tr}$. Under the Bayesian framework, the

standard choice is the posterior predictive distribution

$$f(\boldsymbol{Y}^{te}|\boldsymbol{Y}^{tr}) = \int f(\boldsymbol{Y}^{te}|\theta)\pi(\theta|\boldsymbol{Y}^{tr})d\theta. \tag{2.18}$$

In order to assess the quality of these distributions, scoring rules can be used, (e.g. see Dawid and Musio, 2014; Gneiting and Raftery, 2007) as indices whose small values typically indicate good performance. For example, one common choice for a scoring rule is the log score. For a predictive distribution $h(\boldsymbol{Y}^{te})$ the log score is defined as

$$LS(\boldsymbol{Y}^{te}) = -\log h(\boldsymbol{Y}^{te}). \tag{2.19}$$

The log score is among a class of scoring rules with the desired property of being strictly proper. Strict propriety for a scoring rule ensures that the optimal model among the ones considered will be uniquely identified. More specifically, the score of this optimal model will be strictly lower than the scores of the other models; in the case of it being smaller or equal we get a proper scoring rule rather than strictly proper.

The log score may be seen as a natural extension to the goodness of fit criterion based on the likelihood ratio test statistic for prediction assessment. Consider an SEM model, defined by (2.1) or (2.6) and (2.4), and suppose we want to compare it against the saturated model, e.g. in the case of continuous data the model $\boldsymbol{Y}^{te} \sim N(\alpha, \Sigma)$ for an unconstrained variance-covariance matrix Σ. Denoting with $f^{SEM}(\cdot)$ and $f^{S}(\cdot)$ the density functions of the SEM and saturated models respectively, the difference between the two log scores becomes

$$-\log\left[\frac{f^{SEM}(\boldsymbol{Y}^{te}|\hat{\theta}^{tr})}{f^{S}(\boldsymbol{Y}^{te}|\hat{\alpha}^{tr}, \hat{\Sigma}^{tr})}\right]. \tag{2.20}$$

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

Note that in (2.20), 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 (2.18). Computing some scoring rules, such as the log score require access to the predictive distribution. However, this predictive distribution may be intractable and only samples from it are available, for example via MCMC. When predicting one dimensional data, one can use posterior draws to estimate the distribution using techniques involving kernel densities or the mixture of parameters approach (Krüger et al., 2020; Jordan et al., 2019). But in the SEM context, such techniques cannot be applied since the forecasts for the continuous data case are multivariate. It is possible, however, to use other scoring rules such as the energy score (Gneiting and Raftery, 2007) and the variogram score (Scheuerer and Hamill, 2015). These scores, which are presented in Section 2.3.3, can be computed using posterior draws and do not require access to the exact distribution. For the categorical data case, as we illustrate in Section 2.3.4, the log score can actually be calculated by reformulating the model in terms of the response pattern frequencies.

So far we have assumed a single split between the training and test data, but this may not provide representative results in cases where there are too many peculiar data points in the training or the test data. To limit the effect of such unfortunate splits, cross validation may be used. The procedure can be described as follows:

1. Split the data randomly into K parts.
2. For each of the K parts repeat the following steps:
   (a) Designate the selected group as the test data set and use the other $K-1$ groups together as the training data set.
   (b) Fit the model in the training data set and draw samples from its posterior and its posterior predictive distribution to predict the data in the test set.
   (c) Evaluate the predictions via the chosen scoring rule against the test data.
3. Aggregate the values of the scoring rules across all K groups by summing or averaging.

A nice feature of the above procedure is that all data points appear both in the training and test datasets. Regarding the choice of K, the aim is to ensure a good balance between having

an adequate amount of data points in both training and test samples, so this depends on the sample size. For large enough sample sizes a choice of $K = 3$ often works well. Another option is to use more than one K and average over them as well. For the choice of $K = 3$, the model has to be fitted three times, but in each of these times the sample size is two-thirds the size of the entire sample. The computational time therefore increases with K.

It is interesting to note that the calculation of both PPP values and scoring rules are based on the posterior predictive distribution. Nevertheless, there is an essential difference between the two approaches. PPP values are based on the posterior distribution conditional on the entire dataset and the prediction is made again on the entire dataset. In the scoring rules approach the posterior is conditional only on a subset of the data (training sample) and the prediction is made on the complement of that set (test sample).

2.3.3 Scoring rules for continuous and normally distributed data

In order to assess the predictive performance we need to select a scoring rule. As mentioned in the previous section, the log score is not available in the case of continuous data and the available options are the energy and the variogram scores. To choose between these two, we note that the energy score has been reported to have little sensitivity in detecting misspecifications on correlation matrices (Pinson and Girard, 2012; Scheuerer and Hamill, 2015), which is in line with our empirical findings from our simulation experiments. For this reason, we proceed with the variogram score. To calculate this score for a single data point $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$, we need a set of M samples from the corresponding predictive distribution; let $\tilde{\mathbf{y}}_m = (\tilde{y}_{m1}, \ldots, \tilde{y}_{mp})$ be the m-th sample of a draw and denote all these samples together by $\widetilde{\mathbf{Y}} = \{\tilde{\mathbf{y}}_m\}_{m=1}$. The variogram score for this point is defined in its general form as

$$VS(\mathbf{y}_i, \widetilde{\mathbf{Y}}) = \sum_{j=1}^{p} \sum_{k=1}^{p} w_{j,k} \left(|y_{ij} - y_{ik}|^P - \frac{1}{m} \sum_{m=1}^{M} |\tilde{y}_{mj} - \tilde{y}_{mk}|^P \right)^2 \tag{2.21}$$

where, in (2.21), the j, k are just indices to consider all pairs of each data point $\mathbf{y}_i$ of dimension p, the $w_{j,k}$s are weights and P is the order of the variogram. We follow common practice by

setting all weights to one, and P at its default value of 0.5. In the context of cross-validation, for each split between training ($\boldsymbol{Y}^{tr}$) and test data ($\boldsymbol{Y}^{te}$), the variogram score can be computed by obtaining samples from the posterior based on $\boldsymbol{Y}^{tr}$, and using them to draw samples $\{\tilde{\mathbf{y}}_m\}_{m=1}^{M}$ from the posterior predictive distribution (2.18). The samples can then be inserted in (2.21) together for each of the data points in the test set ($\boldsymbol{Y}^{te}$) in place of $\mathbf{y}_i$. The scores are then added over all points in the test set, to calculate the score corresponding to this train-test split for the model considered. In the case of 3-fold cross-validation, this procedure is repeated for all 3 train-test splits and aggregated by summing or averaging.

2.3.4 Scoring rules for binary and ordinal data

It is possible to compute the log score in the case of binary or ordinal data via the alternative formulation based on frequency patterns, hence we focus on this scoring rule. Note that the posterior predictive density is given by Equations (2.16) and (2.15) where the integral in the latter is with respect to the posterior based on the training data. We can therefore write the log scoring rule for a set of observed frequencies in the test data $\mathbf{O}^{te} = (O_1^{te}, \ldots, O_R^{te})$ based on probabilities $\pi^{tr} = (\pi_1(\theta)^{tr}, \ldots, \pi_R(\theta)^{tr})$, obtained based on the posterior from the training data, as

$$LS(\mathbf{O}^{te}, \pi^{tr}) = -\log f(\mathbf{O}^{te}|\pi^{tr}) = -\log\left[c\prod_{r=1}^{R}\left[\pi_r(\theta)^{tr}\right]^{O_r^{te}}\right] = -\sum_{r=1}^{R} O_r^{te}\log\pi_r(\theta)^{tr} + c, \quad (2.22)$$

where c represents a constant. Note that the log score only differs as a metric to G^2 by a constant, which essentially confirms the argument made earlier in Equation (2.20), about the connection of the likelihood ratio test and the log score, and makes it more specific to categorical data.

2.3.5 Model assessment with fit and predictive performance indices

Our procedure contains two main elements: assessing goodness of fit, as done routinely under current practice, but also assessing out-of-sample predictive performance. For goodness of fit,

the governing well-known procedure is to check if the fit of the hypothesised model, the EZ model in our framework, is no worse than that of the unconstrained model (also known as the saturated model). As described earlier this can be checked by looking at the PPP values of the EZ model. In case of satisfactory PPP value, our recommendation is no different than the standard course of action, to support the hypothesised model, and there is no need to look further.

Now let us consider a situation where the EZ model does not fit the data well, in terms of a PPP value, but the AZ model does. One of the main arguments of this paper is that the researcher should not rush to support the hypothesised model as the satisfactory PPP value may as well be due to the AZ model overfitting the data. Our definition of 'overfit', specifically to the SEM context is the following: If the AZ model is better than its EZ counterpart in terms of goodness of fit but also worse in terms of out-of-sample predictive performance, then it overfits the data. In other words, if the gains in goodness of fit of the AZ model, over the corresponding EZ model, are not based on systematic patterns of the data, then these gains would be of no help when predicting unseen data. Moreover the slightly increased model complexity of the AZ model may have an adverse effect in terms of prediction over the corresponding EZ model. In other words, we are seeking parsimony in addition to goodness of fit. Hence, according to our suggested framework, if the AZ model is worse than the EZ model in terms of the relevant scoring rule, then there is little support in the data for the hypothesised model.

Next, let's consider the case where the AZ model has good PPP value, in contrast with the corresponding EZ model, and also better predictive performance as measured by the relevant scoring rule. Our view in this case is to conduct further checks. There is a possibility that AZ model is just improving upon a poorly specified EZ model but there may exist other EZ or AZ models that predict even better. If the poor fit of the EZ model is due to only some small cross loadings or error correlations, then the AZ model that captures these quantities model should perform really well. But if there are also some other systematic patterns missing from the EZ model, the improvement offered by the AZ model would be limited. Ultimately, the question that we want to answer is whether the predictive performance of each one of these models is good enough. Therefore, it is essential to establish a benchmark when comparing predictive

performances. In the case of goodness of fit assessment this is done by the performance of the unconstrained saturated model. But this may not be a suitable choice for assessing predictive performance (MacCallum et al., 1992). The problem lies in the fact that the saturated model has substantially higher complexity, or else substantially larger number of parameters, than the hypothesised models. Generally speaking, if two models with different numbers of parameters have similar in-sample performance, then the one with the smaller number of parameters will generally perform better out of sample. An alternative option for a benchmark model, exploited in this paper, is the EFA model with the same number of factors as the hypothesised model. This model has generally fewer parameters than the saturated one and is generally expected to perform well in terms of predictive performance as it is allowed to search for systematic patterns in the data without any restrictions, other than having k factors. This is not the case for the EZ and AZ models, where explicit restrictions are given and it is often hoped that they will not be too far from those indicated from the EFA. Hence, in order to regard the predictive performance of the hypothesised model as satisfactory, its scoring rule should be comparable with that of the EFA model chosen as the benchmark.

Caution must be exercised over the choice of the benchmark EFA model, as selecting an over-parameterised EFA model will set the bar too low in terms of predictive performance. Therefore it may be more appropriate, in some cases, to check the parsimony of the EFA model selected as the benchmark. This can be done, for example, by considering EFA models with fewer factors, provided that they fit the data well. In line with such considerations, we note that the presence of the small error correlations induced by the $\mathbf{u}_i$s under the approximate zero framework may offer an advantage to CFA models in terms of prediction as it can be viewed as an additional minor factor. Hence, in order to bring CFA and EFA models onto a level playing field, it may be reasonable to incorporate small error correlations to both of them via the EFA-C model.

Note also that, while the AZ models are more flexible than their EZ counterparts, they can still perform badly in cases of substantial model misspecification. For example, in cases of large enough cross loadings, say more than 0.5, using the Normal$(0, 0.01)$ as prior can still result in poor performance compared to the EFA model. This comparison may thus be exploited to detect misspecified models as we illustrate in Sections 2.4 and 2.5.

We summarise below the recommendations of our proposed framework.

1. If the EZ model has satisfactory fit indices such as PPP values, there is strong support towards the hypothesised model.

2. If both EZ and AZ models have poor PPP values then there is little support of the hypothesised model. Perhaps it may be useful to use more vague priors to explore its weaknesses. It would be expected in this case that the EFA models will have better predictive performance otherwise there maybe 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 overfit, 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 overfit is not the case. To check if 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 towards the hypothesised model.

Model fit assessment is by no means an easy task especially in factor analysis modelling where model misfit can be due to various reasons such as misspecification of the latent variable distribution, item dependencies, skewed data and non-linear predictors. In this paper, for the calculation of PPP values we use a discrepancy function that looks at the overall fit of the model both in the case of continuous and categorical data. It is useful to complement those overall goodness of fit tests with other measures of fit such as residuals and limited information test statistics that check the fit on lower order margins and detect item misfit as explained in Section 3.1. It is because of those complexities that our proposed methodology is trying to shed light to model fit challenges using a different set of tools that look at the model's out of sample prediction performance. This provides new tools within the Bayesian modelling framework in SEM and highlights even further the challenges of fit and problems of PPP values. Furthermore,

the new residual term in the linear predictor defined by the item-individual random effects $\mathbf{u}_i$s plays a key role since it can be used as model diagnostics to detect outliers such as leaked items and cheating behavior in educational testing or secondary response strategies employed by some of the respondents to some of the items.

The approximate zero framework has also been applied in the study of Measurement Invariance in SEM. Measurement Invariance (MI) for multi-group SEM, refers to the constraint of setting parameters to be equal across different groups. Although this is still an active area of research, there have been some indications so far, that the approximate zero framework has limitations in this context. Several simulation studies call into question the effectiveness of the approximate zero framework, at least within the context of measurement invariance. (Pokropek, Schmidt, and Davidov, 2020) studied the effect that different prior choices on the parameters that control cross-group differences have, in the final analysis and model choice. They found that prior misspecifications have little impact on point estimates and more substantial impact on credible interval coverage. Overall, they found that MI testing in the approximate zero framework is meaningfully dependent on the study size, which is not a desirable feature. In the same vain, (Pokropek, Davidov, and Schmidt, 2019) did a large comparative simulation study between approximate MI and alternative approaches, such as exact and partial MI. They found that the approximate MI model is suitable when there are small MI deviations, even when the number of deviations is high. However, it provides biased results when the deviations are moderate or large, even when the number of such deviations are small. The authors concluded that the alternative approaches are preferable in practice. A similar, but smaller scale simulation study by (Van De Schoot, Kluytmans, Tummers, Lugtig, Hox, and Muthén, 2013) reached similar conclusions and highlighted the ineffectiveness of the approximate zero MI approach in cases where the group differences are not small.

2.4 Simulation experiments

2.4.1 Setup

Simulation experiments were conducted to study the performance of the proposed models and demonstrate the assessment framework for continuous and binary data. We focus on two cases of data generated using Equation (2.4), i.e. 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 correlated item-individual random effects.

For both continuous and binary data, we considered $p = 6$ items and $k = 2$ factors. The factor loadings used to generate the data, in each of the three scenarios, are shown in Table 2.1. Although the data were generated under the three scenarios, in all of them the typical

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

Table 2.1: True factor loadings used in the three simulation scenarios.

hypothesised model assumes a simple structure in which the first three items load on the first factor whereas the last three load on the second factor. In other words, for the AZ model,

the first three elements of the first Λ column and the last three of the second Λ column are regarded as the major parameters, whereas the other elements of Λ are cross-loadings. In all three scenarios, the factor correlation was set to 0.2, and the intercepts α were all zero. The sample sizes were set to $n = 1,000$ in the continuous data and $n = 2,000$ in the binary data. For Scenario 2, equation (2.7) was used by setting the matrix $\Omega + \Psi^*$ to have ones in the diagonal, and 6 non-zero off-diagonal elements set to 0.2 with the remaining 9 off-diagonal elements set to zero.

For each scenario, the proposed model assessment framework of Section 2.3 was put into action by computing the PPP values and scoring rules for all the previously mentioned models. After fitting and summarising these models, according to Sections 2.3.3 and 2.3.4, we proceeded according to the recommendations of Section 2.3.

The models and priors were specified as outlined in Section 2.2 and samples from the posterior of each model were obtained using Hamiltonian MCMC programmed using the Stan language. In the case of continuous data, 1,000 iterations were used as the warm-up period and another 2,000 for inference purposes. In the case of binary data, it was 2,000 for warm-up and 2,000 for analysis purposes. The models were run in 4 parallel chains in each case resulting in $4 \times 2,000 = 8,000$ posterior draws. In all cases, we ensured successful convergence of the chains with the help of the automatic metrics implemented in Stan as well as visual inspection of the posterior draws.

In all instances, we applied a 3-fold cross-validation and aggregated the scores by summing. Given that a scoring rule is a comparative index, we reported the difference in scores between each model and the best model. In other words, the best model of each case, or else the one with the smallest score, was given the value of zero.

The next two sections present the results of the simulation experiments for continuous and binary data. The aim of these experiments is to illustrate the performance of the proposed model framework and provide a proof of concept. More detailed simulation experiments will be helpful, as we discuss in the next sections, and are left for future research.

2.4.2 Continuous data

Table 2.2 gives the variogram score (VS) and the PPP values for the three simulation scenarios. We use the variogram score with parameter $P = 0.5$ and weights $w_{ij} = 1$. Starting with Scenario 1, we note that all models fit the data well in terms of the PPP values. In terms of predictive performance, we note that the EZ model performs best, which is not surprising given that the data were generated from it. Note that the EZ model even improves upon the EFA models in terms of predictive performance as it is a more parsimonious model.

	Scenario 1		Scenario 2		Scenario 3	
Model	PPP	VS	PPP	VS	PPP	VS
EZ	0.66	0	0.00	6.93	0.00	17.79
AZ	0.51	4.28	0.31	0	0.53	1.58
EFA	0.62	2.06	0.00	0.23	0.59	0
EFA-C	0.53	1.05	0.38	0.03	0.56	1.45

Table 2.2: Simulation Results for Continuous Data. PPP values and sum of variogram scores of 3-fold cross validation for the relevant models. For each scenario, the best model has 0 variogram score and the differences from it are reported for the other models.

In simulation Scenario 2, both the EZ and EFA models exhibit poor fit according to their PPP values, which is again not surprising given that these models assume zero error correlations. In contrast, the AZ and EFA-C models that allow for small, yet not exactly, zero error correlations both fit well. At this point, the question is whether the improved fit of the AZ model is due to fitting noise or else overfit as defined in *Recommendation 2*. But if AZ was overfitting the data, we would not expect to see an improved performance over the EZ model, as we see here. We can expand the investigation of the AZ model further, wondering whether there is another theory that leads to an AZ model with even better predictive performance. *Recommendation 3* may shed light on this question when we compare the predictive performance against the EFA models. We see that AZ is quite competitive against those models and, in fact, does better, although their variogram scores are quite close. Hence, according to our proposed framework, there is strong support towards the AZ model and, consequently, the hypothesised model. This appears to be a reasonable conclusion in the SEM context given that the poor fit is due

to error correlations that are usually linked with observation error rather than factor loading misspecifications. As before, the use of PPP values alone would not have been enough to reach that conclusion.

Finally, let us consider the simulations for Scenario 3, where the EZ model does not have a good fit, as one would expect, but all the other models have PPP values around 0.5. As before, we are interested in whether the AZ model overfits and what conclusions we can draw on the hypothesised model. To answer such questions, we set the benchmark model to be the EFA model with the higher predictive performance; it is the EFA model in this case, as one would expect since the data were simulated without error correlations. The variogram score of the AZ model is again much better than that of the EZ, as it utilises its approximate zero cross loadings to pick up the two cross loadings of 0.6. But it is not better than the EFA model, thus not ruling out the presence of a different hypothesised theory regarding the loading structure of the six items. Indeed, the theory corresponding to factor loadings according to Scenario 3 described in Table 2.1 provides a better model as the data were simulated from it.

2.4.3 Binary data

In this section, we summarise the results of the three simulation experiments for the case of binary data. Table 2.4.3 gives the PPP values and the log scores. The results are very similar to the continuous case. In the case of Scenario 1, all models demonstrate good fit as indicated by the PPP values. Furthermore, the EZ model is the optimal one in terms of predictive performance (*Recommendation 2*) which is reassuring since data were simulated from the EZ model. In Scenario 2, we see that the EZ model exhibits very poor fit, caused by the additional error correlations in the simulated data, as indicated by the PPP value of 0.02. The rest of the models exhibit a moderately good fit with PPP values above 0.10. Similarly to the continuous case, the AZ model does well in terms of both *Recommendation 2* and *3* being the model with the best predictive performance. Finally, in Scenario 3, in terms of model fit the EZ model also fails, due to the presence of non-zero cross loadings. The other models do well, leaving some questions open in terms of the validity of the hypothesised theory. For this reason,

Recommendation 3 compares the predictive performance of AZ against the best performing EFA model. In this case, the AZ model is not as good as the EFA.

	Scenario 1		Scenario 2		Scenario 3	
Model	PPP	CV-L	PPP	CV-L	PPP	CV-L
EZ	0.52	0	0.02	4.19	0.00	7.31
AZ	0.50	0.68	0.12	0	0.52	1.90
EFA	0.59	1.45	0.13	0.09	0.45	0
EFA-C	0.54	3.27	0.17	0.24	0.50	2.96

Table 2.3: Simulation Results for Binary Data. PPP values and sum of variogram scores of 3-fold cross validation for the relevant models. For each scenario, the best model has log score equal to 0 and the differences from it are reported for the other models.

2.4.4 Parameter recovery for the AZ model in the binary data case.

To investigate the parameter recovery performance of the AZ model in the binary data case, we performed a simulation experiment where 100 different datasets were simulated and the AZ model was fitted on each one of them to obtain samples from its posterior. More specifically the data were drawn from the EZ model, so that we focus on the main parameters of interest, namely the factor loadings and the correlation of the factors, each with sample of size $2,000$. The factor loadings used to simulate the data are the same as in the Scenario 1 of the simulation experiments of Section 2.4 and the correlation between the two factors was 0.2. Finally, the intercept parameters used to simulate the data were all set to 0. We used the parameterisation where the loadings are unrestricted and the factors' variance is fixed to 1 hence their covariance matrix is restricted to be a correlation matrix.

Regarding the prior specification of the AZ model, we assumed that, according to the hypothesised theory, the first factor loads on the first 3 items and the second factor loads on the last 3 items. Hence, the rest of loading parameters were regarded as cross-loadings, and were assigned informative priors around zero. The rest of the priors were assigned as described earlier in the paper.

As informal measures of how well the parameters are recovered, we focused on frequentist

properties of some estimators derived from the posterior samples. The estimators consisted of the 95% credible intervals as interval estimators, obtained from the sample 2.5-th and 97.5-th points extracted from the posterior draws, as well as the posterior mean and median as point estimator. We then examined the coverage probability of the former and the bias of the latter. We note that these summaries (95% credible intervals, posterior mean, and posterior median) may not exhibit the desired frequentist performance even in the case the model fits the data well, as they have not been constructed to do so. Nevertheless, if they happen to perform well, it is definitely reassuring.

We examined the main parameters of interest, such as the loadings Λ and the factor correlation ρ. The results are summarised in Table 2.4 and they contain coverage probabilities and biases of the previously mentioned posterior summaries. As we can see, the coverage probabilities are reasonably close to 0.95 whereas the biases are not substantial, particularly for the posterior median. We therefore conclude, while noting the informal nature of the experiment, than no substantial concerns regarding parameter recovery are raised.

Parameter	True Value	Coverage Rate	Bias of Post. Mean	Bias of Post. Median
$\Lambda_{[1,1]}$	1.0	0.94	0.06	0.03
$\Lambda_{[2,1]}$	0.8	0.96	0.05	0.03
$\Lambda_{[3,1]}$	0.8	0.94	0.05	0.03
$\Lambda_{[4,1]}$	0.0	1.00	0.00	0.00
$\Lambda_{[5,1]}$	0.0	1.00	0.00	0.00
$\Lambda_{[6,1]}$	0.0	1.00	0.00	0.00
$\Lambda_{[1,2]}$	0.0	1.00	0.00	0.00
$\Lambda_{[2,2]}$	0.0	1.00	-0.01	-0.01
$\Lambda_{[3,2]}$	0.0	1.00	0.00	0.00
$\Lambda_{[4,2]}$	1.0	0.99	0.03	0.00
$\Lambda_{[5,2]}$	0.8	0.95	0.06	0.04
$\Lambda_{[6,2]}$	0.8	0.99	0.03	0.02
ρ	0.2	1.00	-0.01	-0.01

Table 2.4: True values, 95% coverage success rate and bias of point estimators out of 100 replications, AZ model for binary data.

2.5 Real-world data examples

In this section, we demonstrate our proposed model assessment framework with two real datasets. The first dataset is a popular psychometric test, usually referred to as the 'Big 5 Personality Test', that decomposes human personality along 5 main traits using 15 items measured on a 7-point likert scale. The second data set is based on the Fagerstrom Test for Nicotine Dependence (FTND) that consists of six binary variables.

2.5.1 Example 1: 'Big 5 Personality Test'

The data were collected as part of the British Household Panel Survey in 2005-06 focusing on female subjects between the ages of 50 and 55; the sample size consists of 589 individuals. The 'Big 5 Personality Test', as it is known, is a 15-item questionnaire on topics of social behaviour and emotional state. Participants answer each item on a scale from $1-7$, 1 being 'strongly disagree' and 7 being 'strongly agree'. Items are treated here as continuous. The test is designed to measure five major, potentially correlated, personality traits. Each trait corresponds to a factor, and each factor is hypothesised to explain exactly 3 out of 15 items.

The data have been analysed in several papers including Muthén and Asparouhov (2012), Stromeyer et al. (2015), and Asparouhov et al. (2015). In these analyses, an interesting finding was that the exact zero (EZ) model did not exhibit good fit based on several standard indices including the PPP values. The approximate zero (AZ) model gave a good fit in terms of the PPP values, but also had many non-zero error correlations. This raised concerns over whether the flexibility of the AZ model is picking up noise, thus resulting in a misleadingly high PPP value. The validity of the 'Big 5' scale on these data remains unclear. In an attempt to shed more light on this question we apply our model assessment framework and summarise the results in Table 2.5.

The picture is very similar to the error correlations scenario in Section 2.4.2, yet much more pronounced. Our analysis confirms the poor fit of the EZ and the EFA with five factors. Both

Model	PPP	CV-V
EZ	0.0	56.43
AZ	0.23	0
EFA	0.00	94.35
EFA-C	0.38	78.47

Table 2.5: 'Big 5' personality test data, BHPS. PPP values and sum of variogram scores of 3-fold cross validation for the relevant models. For each scenario, the best model has 0 variogram score and the differences from it are reported for the other models.

AZ and EFA-C models have reasonably good PPP values. This implies that error correlations contribute to the lack of fit to a large extent. In order to assess the question of overfit and draw conclusions on the validity of the 'Big 5' scale, we calculate the variogram scores for each model. The variogram score of the AZ model clearly dominates all the other models, suggesting that the model is fitting consistent patterns in the data and it clearly outperforms the EFA models. This points to strong support towards the 'Big 5' scale, attributing the fit issues of the EZ model to error correlations that could have been caused by the wording and other issues often present in survey data like the BHPS.

2.5.2 Binary Data: Fagerstrom Test for Nicotine Dependence

In this section we use data on 566 patients available through the National Institute on Drug Abuse (study: IDA-CTN-0051). The Fagerstrom Test for Nicotine Dependence (FTND) (Heatherton et al., 1991) was designed to provide a measure of nicotine dependence related to cigarette smoking. It contains six items that evaluate the quantity of cigarette consumption, the compulsion to use, and dependence. The original scale consists of 4 binary and 2 ordinal items for self-declared smokers:

1. FNFIRST: How soon after you wake up do you smoke your first cigarette? ['3'=Within 5 minutes, '2'=6 - 30 minutes, '1'=31 - 60 minutes, '0'=After 60 minutes]

2. FNGIVEUP: Which cigarette would you hate most to give up? ['1'=The 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, etc.)? [‘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].

The mapping between the FTND scale and a CFA model is not clear, see e.g. Richardson and Ratner (2005) and references therein. Richardson and Ratner (2005) fitted a single factor, a correlated two factor, and a two factor model with one cross loading. These models were also considered in our analysis and are denoted as 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. In addition to these models, we also considered their approximate zero versions, denoted as 1F-C, 2F-AZ, and 2F-AZ-b respectively, 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 2.6.

Examination of the PPP values reveals concerns about the fit of the models 1F and 2F-EZ, so these are ruled out of the discussion. This raises several questions: Is the 2F-EZ-b the best model or do any of the AZ model versions, 2F-AZ or 2F-AZ-b, do better? Is the best of these three good enough? Perhaps more importantly, which measurement scale should be used for the FTND test on the basis of this dataset?

We attempt to shed light on these questions with the use of cross-validated log scores. The best model is the 2F-EZ-b correcting the misspecifications of 2F-EZ with a single additional

Model	PPP	CV-L
1F	0.01	15.98
1F-C	0.32	6.63
2F-EZ	0.04	10.45
2F-AZ	0.40	6.23
2F-EZ-b	0.41	0.00
2F-AZ-b	0.44	2.01
2F-EFA	0.44	2.66
2F-EFA-C	0.58	2.38

Table 2.6: PPP values and and sum of log scores of 3-fold cross validation for the relevant models. The models with '-b' refer to the measurement model with the first item loading to both factors. The best model had log score equal to 0 and the differences from it are reported for the other models.

parameter. The fact that the log score of 2F-EZ-b is smaller than that of the EFA models provides support towards the scale with two correlated factors where the item 'FNFIRST' loads on both of them.

2.6 Discussion

In this paper, we generalise the Bayesian SEM framework, introduced in Muthén and Asparouhov (2012), along two directions. First, by expanding the model to allow for other data distributions than the Normal; e.g. logistic often used in IRT models. Second, in terms of model exploration and assessment, by developing a suitable framework that goes beyond goodness of fit and allows us to address questions that naturally arise from the application of Bayesian SEM. This framework incorporates scoring rules combined with cross-validation to the existing fit indices.

As illustrated on simulated data and real-world examples, the use of the scoring rules can prove quite useful in SEM analysis. Nevertheless, as with any index, it would be helpful to explore it further and get a better understanding of the range of values indicating a good model in different settings. This range may depend on the sample size, the number of factors and parameters, the type of the data, the choice of the scoring rules, the number of folds or the

form of cross-validation in general, the choice of the benchmark model etc. Another important component, present in any form of Bayesian analysis, is the prior specification. The behaviour of the scoring rules under different priors, e.g. the spike and slab priors, as in Lu et al. (2016), rather than the ridge-type priors, is also an interesting question.

The approximate zero framework has also been applied in the study of Measurement Invariance in SEM. Measurement Invariance (MI) for multi-group SEM, refers to the constraint of setting parameters to be equal across different groups. Although this is still an active area of research, there have been some indications so far, that the approximate zero framework has limitations in this context. Previous research has revealed that the use the approximate zero framework can be the problematic, especially when the size of the mis-specifications is moderate or high.

The calculations can be implemented using MCMC through standard user-friendly software like Stan and can be combined with existing packages for SEM. This opens up the possibility of using fast approximate methods such as Variational Bayes (Kucukelbir et al., 2017a) that are automated and readily available. This can be particularly useful in categorical data application where the use of MCMC and the presence of high-dimensional latent variables can result in computation times that are larger than the users' expectations. Moreover, Variational Bayes can be used to improve the efficiency of MCMC samplers.

Further extensions of the generalised family of models can also be explored; for example, non-Normal errors $\mathbf{e}_i$s or random effects $\mathbf{u}_i$s. Inspection of the latter may also provide diagnostic information for detecting outliers and removing items to purify the constructs. It would also be interesting to explore the connections with Bayes factors, as they tend to provide parsimonious models that typically do well in terms of cross-validation. Calculating Bayes factors is not always straightforward and they are also more sensitive to the choice of priors. However, such issues can be alleviated by suitable choice of priors, as done in this paper.

Finally, it is important to note that the developed model assessment framework and the CV index 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 (Yang and Dunson, 2010; Song et al., 2013). 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.

Chapter 3

Sequential Bayesian Inference for Factor Analysis

3.1 Introduction

Factor analysis is a statistical technique that uses a small number of latent factors to model the behaviour of a potentially larger number of observed variables. It can be used to model directional (regression coefficients) and non-directional (correlations) relationships amongst latent variables (structural model) which are identified by observed variables (measurement model). Factor analysis is part of Structural Equation Modelling (SEM) and Confirmatory Factor Analysis (CFA), where the focus is on verifying scientific hypotheses. Alternatively, Exploratory Factor Analysis (EFA) uses factor analysis as a dimension reduction technique or as a tool to uncover patterns in multivariate data. In both cases, Bayesian approaches have been developed and offer several benefits such as providing a natural framework for parsimonious model choice (Lopes and West, 2004; Frühwirth-Schnatter and Lopes, 2018; Conti et al., 2014), performing well in small sample sizes (Depaoli and Clifton, 2015) and assessing model fit (Muthén and Asparouhov, 2012).

Sequential Bayesian modelling has received attention in the setting of hypothesis testing where it has been proven superior to the traditional null hypothesis (NHST) paradigm. Specifically,

Sequential Bayes Factors (SBF) do not suffer from bias associated to the stopping rule, the practice of stopping the processing of new data only when conclusive evidence is reached. Contrary to NHST theory, which requires a prespecified sampling plan, Bayes Factors allow for flexible sampling design and unlimited testing (Pramanik et al., 2021; Schnuerch and Erdfelder, 2020). One drawback of Bayes Factors is that they are not always easy to compute except from some simple cases of mean difference analysis (Schönbrodt et al., 2017). Several schemes have been proposed but they are only approximate and typically they are also non trivial to compute; see Vitoratou, Ntzoufras, and Moustaki (2016) and the references therein for some examples in the context of factor analysis. The sequential scheme developed in this paper offers another alternative with the additional benefit that there is no need to refit the model from scratch when new data become available. The goal of a sequential scheme is to recursively explore the sequence of posterior distributions

$$\pi_0(\theta) = p(\theta), \quad \pi_i(\theta) = p(\theta|\mathbf{y}_{1:i}), \quad i = 1, \ldots, n, \tag{3.1}$$

where θ denotes all the unknown parameters in our model. The sequential inference paradigm approximates recursively these posterior distributions and the model evidence $p(\mathbf{y}_{1:i})$. For models where the likelihood $f(y|\theta)$ is available, the sequential scheme of Iterative Batch Importance Sampling (IBIS) (Chopin, 2002) and its more general framework (Del Moral et al., 2006) provide the standard option. Factor models based on continuous and normally distributed data fall in this category and in this paper we tailor the IBIS algorithm for this case. For models including latent variables, such as factor models based on binary data, interest lies in the augmented joint posterior of parameters and latent variables

$$\pi_0(\theta, z_0) = p(\theta)p(z), \quad \pi_i(\theta, \mathbf{z}_{1:i}) = p(\theta, \mathbf{z}_{1:i}|\mathbf{y}_{1:i}), \quad i = 1, \ldots, n, \tag{3.2}$$

In such cases one option is provided by the SMC^2 scheme of Chopin et al. (2012) which focuses on the case where the latent variables satisfy the Markov property, e.g. Hidden Markov Models. In this paper we aim to construct an alternative scheme that takes advantage of the independence, rather than Markov dependence, between the latent variables that is typically

assumed in factor analysis.

Despite the fact that the developed computational scheme of the paper is sequential, it offers several benefits even in non-sequential contexts. First, it can provide the posterior joint distributions of the parameters and the model evidence in one go. Second, it allows the computation of scoring rules under the prequential framework; see for example (Dawid and Musio, 2014). Third, it can provide a more robust alternative computational scheme than MCMC schemes that can be helpful when the target distribution has problematic landscape, e.g. being multi-modal.

In summary, a sequential framework for factor analysis can be desirable for three main reasons. Firstly, it is needed to develop methodology that offers sequential posterior distributions for the cases where sequential inference is needed. Typically because the data stream is available one data point at a time and the analysis must be done in an online fashion. Second, because we may wish to compute the model evidence which allows us to perform model choice based on formal Bayesian quantities. This connects to a contribution of Chapter 2 which develops a model choice paradigm based on predictive performance and scoring rules. An interesting further research question is to compare the two approaches for model choice. Third, an efficient sequential approach provides an alternative inference algorithm to standard MCMC. Sequential algorithms can be useful even when used in a non-sequential way, because they can be more robust computationally than standard MCMC.

We begin by laying out the framework for continuous data in Section 3.2. We then proceed to explore the more challenging case of categorical data, given the presence of latent variables, in Section 3.3. In Section 3.4 we demonstrate the framework with simulation experiments and a real data example. Finally Section 3.5 concludes with some relevant discussion.

3.2 Sequential Monte Carlo for Factor Analysis Based on Continuous Data

3.2.1 Model and Priors

We use a unified Bayesian framework that encompasses models for categorical and continuous observations. In this Section we focus on continuous data. Suppose there are p observed variables (items) denoted by $\mathbf{y} = (y_1, \ldots, y_p)$ and that their associations are explained by k continuous latent variables (factors) denoted by $\mathbf{z} = (z_1, \ldots, z_k)$. The classical linear factor analysis model (*measurement model*) is:

$$\mathbf{y}_i = \alpha + \Lambda \mathbf{z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{3.3}$$

where α is a $p \times 1$ vector of intercepts, Λ is the $p \times k$ matrix of factor loadings and n is the sample size. The vector of latent variables $\mathbf{z}_i$ has usually a Normal distribution, $\mathbf{z}_i \sim N_k(0, \Phi)$. The $\boldsymbol{\epsilon}_i$s are error terms assumed to be independent from each other and from the $\mathbf{z}_i$s. If they are assumed to be Normally distributed as $\boldsymbol{\epsilon} \sim N(0_p, \Psi)$, $\Psi = \text{diag}(\psi_1^2, \ldots, \psi_p^2)$, the latent variables $\mathbf{z}$ can be integrated out using standard properties of the Normal distribution, to obtain

$$\mathbf{y}_i \sim N(\alpha, \Lambda\Phi\Lambda^T + \Psi). \tag{3.4}$$

The model defined in (3.3) and (3.4) applies to EFA, CFA or more generally SEM depending on how the parameters Λ and Φ are formulated. EFA is the practice of factor analysis for the purpose of reducing the dimensionality of the observed outcomes. This is typically done by fitting a model where all the elements of Λ are all free parameters, subject to some identifiability restrictions, while enforcing $\Phi = I_k$. In CFA and more broadly SEM, the focus is on establishing whether a hypothesised social, psychological, or other scientific theory, that determines which items load on each factor, is compatible with the data. Researchers express the theory by restricting certain elements of Λ to zero (e.g. the so-called cross-loadings) to

assign items to factors, and assessing the model's goodness of fit. The covariance matrix Φ is either unstructured (CFA) or defined by a parametric model that relates the latent variables with each other and optionally with the observed covariates (SEM). For both approaches it is common to set Ψ to be a diagonal matrix, an assumption known as conditional independence of the variables given the factors.

Fitting the model requires resolving certain indeterminacies that arise. Since the scale of the latent factors is not identifiable there are generally two parametrisations to choose from: either the leading loadings in Λ for each factor are fixed to a constant, usually 1, or Φ is constrained to be a correlation matrix. In the second case, further care is needed as the sign of the factor loadings is not identifiable so the parameter space needs to be constrained to the positive or the negative side. Alternatively the parameter space could be unconstrained and instead post-sampling processing can be applied whereby the posterior samples of the loading columns corresponding to each factor should be multiplied by -1 if the relevant leading loading is negative, otherwise they are left as they are. This formulation may be viewed as a special case of the parameter expansion suggested in Ghosh and Dunson (2009) for EFA. In the case of EFA further challenges have to be addressed due to the fact that the likelihood is specified in terms of $\Lambda\Lambda^T$ while the parameter of interest Λ is free. Contrary to CFA, in the EFA setting there are no modelling constraints on the loading matrix, hence the likelihood is invariant under rotations of Λ. Enforcing Λ to be lower triangular (see e.g. Geweke and Zhou, 1996b) is one way to remove the rotational indeterminacy, but it introduces order dependence amongst the observed variables. The choice of the first k variables, which is an important modelling decision (Carvalho et al., 2008b), thus becomes inadvertently impactful. Alternative schemes have been proposed by Conti et al. (2014); Frühwirth-Schnatter and Lopes (2018); Bhattacharya and Dunson (2011), which have the additional benefit of helping to identify the number of factors in a single MCMC run.

Regarding priors we start with the factor correlation matrix Φ which under the full covariance matrix parametrisation receives a prior with a low amount of information, e.g. the Inverse Wishart with the identity as the scale matrix and $p + 4$ degrees of freedom or lower. Alternatively, if we use a correlation matrix Φ, the LKJ prior is assigned, introduced in Lewandowski

et al. (2009), with a similar amount of low information, e.g. LKJ(2). The free loadings in Λ are assigned zero-centred Normal priors $N(0, \sigma^2)$. The prior variance is frequently set to be a large constant, which however can lead to issues related to Lindley's paradox (Lindley, 1957). An alternative choice that protects against such problems is the unit information priors (Kass and Wasserman, 1996), according to which variance is set to a small value that correspond to the amount of information from a single observation point. In this work, our items are in similar scales hence we fix the prior variance to 1 in line with recommendations by Lopes and West (2004) and Ghosh and Dunson (2009). Regarding the diagonal matrix Ψ, we assign independent Inverse Gamma priors on each ψ_j^2

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

where S_y is the empirical covariance matrix and c_0 is a constant that the researcher can choose in order to limit the probability of running into Heywood issues as per recommendations in Frühwirth-Schnatter and Lopes (2018) and Conti et al. (2014). Finally, large variance Normal priors are assigned on the α parameters. In every analysis that follows we use the following wide prior Normal, $\alpha \sim N(0, 10^2)$.

3.2.2 Sequential Algorithm

To target the recursive posteriors $\pi(\theta|\mathbf{y}_{1:i})$ we adopt the Iterated Batch Importance Sampling algorithm (IBIS), introduced by Chopin (2002). At a high level, IBIS works by propagating forward in time a set of parameter particles, each weighted by the likelihood function evaluated at its parameter values. For the standard IBIS approach we need to evaluate the likelihood function $f(y|\theta)$ which is available for the continuous case via formulation (3.4). The ability to integrate out the latent variables, combined with the efficiency of the IBIS algorithm, achieves the goal we set in Section 3.1, i.e. to get an efficient process to draw samples from the sequence of posterior distributions $p(\theta|\mathbf{y}_{1:i}),\ \ i \geq 1$ and compute the model evidence. In what follows we describe the steps of the algorithm, while the full process is presented in Algorithm 1. We begin by drawing N_θ samples of the parameter vector $\theta = (\alpha, \Lambda, \Phi, \Psi)$ from the prior distribution,

Algorithm 1 IBIS

Sample θ^m, for $m = 1, \ldots, N_\theta$ from $\pi(\theta)$ and set $\omega^m = 1$. All operations are assumed to be repeated for all $m \in 1 : N_\theta$.
Then at time $i = 1, \ldots, n$, do:

1: Compute the incremental weights and their weighted average

$$u_i(\theta^m) = f(\mathbf{y}_i|\mathbf{y}_{i:i-1}, \theta^m) = f(\mathbf{y}_i|\theta^m), \qquad L_i = \frac{1}{\sum_{m=1}^{N_\theta} \omega^m} \times \sum_{m=1}^{N_\theta} \omega^m u_i(\theta^m),$$

2: Update the importance weights

$$\omega^m = \omega^m u_i(\theta^m)$$

3: **if** ESS$(\omega) < \gamma$ **then**
4: **procedure** RESAMPLE(θ, ω)
5: **return** θ
6: **procedure** JITTER$(\theta^m, \mathbf{y}_{1:i})$ using an MCMC algorithm
7: **return** $\tilde{\theta}^m$
8: $(\theta^m, \omega^m) = (\tilde{\theta}^m, 1)$

called θ particles and denoted with $\{\theta_m\}_{m=1}^{N_\theta}$. We proceed in increments of time by considering one data point $\mathbf{y}_i$ at each time i. This is also known as data tempering and is adopted in this paper so that we can assess the out of sample predictive performance of the model in question.

At each time i, or else data point, we compute the weights specified by the likelihood function $\omega_i^m = f(\mathbf{y}_{1:i}|\mathbf{y}_{1:i-1}\theta^m)$. Note that the latter simplifies to $f(\mathbf{y}_{1:i}|\theta^m)$ for the factor models defined in the previous subsection. Doing so, the weighted draws of the θ particles, $\{\theta^m, \omega_i^m\}_{m=1}^{N_\theta}$ at time i can be used to evaluate summaries of the posterior $\pi(\theta|\mathbf{y}_{1:i})$. More specifically, expectations with respect to that posterior, $E[g(\theta)|\mathbf{y}_{1:i}]$, can be computed using the estimator

$$\frac{\sum_m [\omega_m g(\theta^m)]}{\sum_m \omega_m} \to E[g(\theta)|\mathbf{y}_{1:i}]. \tag{3.5}$$

(Chopin, 2004) shows consistency and asymptotic normality of this estimator as $N_\theta \to \infty$ for all appropriately integrable $g(\cdot)$. The same holds for expectations with respect to the posterior predictive distributions such as $f(\mathbf{y}_{i+1}|\mathbf{y}_{1:i})$. Since

$$f(\mathbf{y}_{i+1}|\mathbf{y}_{1:i}) \propto f(\mathbf{y}_{i+1}|\mathbf{y}_{1:i}, \theta)\pi(\theta|\mathbf{y}_{1:i}),$$

the weighted θ particles, from $\pi(\theta|\mathbf{y}_{1:i})$, can be transformed into weighted y_{i+1} particles from

$f(y_{i+1}|\mathbf{y}_{1:i})$ by simply drawing y_{i+1}^m from $f(y_{i+1}|\mathbf{y}_{1:i}, \theta^m)$ which, as mentioned earlier, equals to $f(y_{i+1}|\theta^m)$ in our case. Moreover, a very useful by-product of the IBIS algorithm is the ability to compute the model evidence $f(\mathbf{y}_{1:i})$, to calculate Bayes factors. Computing the following quantity in step 1 in Algorithm 1 yields a consistent and asymptotically normal estimator of $f(\mathbf{y}_i|\mathbf{y}_{1:i-1})$

$$\frac{1}{\sum_{m=1}^{N_\theta} \omega^m} \sum_{m=1}^{N_\theta} \omega^m u_i(\theta^m) \to f(\mathbf{y}_i|\mathbf{y}_{i:i-1}). \tag{3.6}$$

In other words the output of the IBIS output allows the calculation of all the summaries often obtained from the MCMC outputs, such as the posterior mean, mode, or median, 95% credible intervals, samples from the predictive distribution, but for all the posteriors $\pi(\theta|\mathbf{y}_{1:i})$, $i = 1, \ldots, n$. Moreover it provides estimates of the model evidence for all i.

Note that if we were to only propagate the particles according to steps 1 and 2 of the (IBIS) Algorithm 1, the weights of the particles will eventually deteriorate with very few or even one of them dominating the others, which will lead to inaccurate estimates of the posterior summaries. One index that measures the quality of the weighted θ particles is the effective sample size (ESS)

$$\text{ESS}(\omega) = \frac{\left(\sum_{m=1}^{N_\theta} \omega^m\right)^2}{\sum_{m=1}^{N_\theta} \left(\omega^m\right)^2}. \tag{3.7}$$

The protocol of the IBIS algorithm requires to monitor a degeneracy criterion, which is typically to check if the ESS is less than a prespecified threshold γ, the violation of which triggers a two-step procedure to improve the quality of the θ particles. The first step of this procedure is to resample the θ particles with replacement, e.g. via the multinomial distribution with the normalised weights as probabilities. At that point we reset all weights to 1 but we end up having multiple copies of the θ particles with high weights, whereas some θ particles with low weights are removed. The purpose of this step is to drop θ particles of low weights and focus on the ones with high weight. This can be particularly helpful in the presence of local modes, since the θ particles that can potentially get trapped there will eventually be removed if the density at those modes is low. The second step of this procedure, called jittering, is to apply a MCMC algorithm with initial value at each θ^m particle to sample from the posterior given data up to that point. The MCMC algorithm is run for a few iterations and the last value of the

MCMC chain, denoted by $\tilde{\theta}^m$ becomes the new value θ^m. The purpose of jittering is to avoid having exact multiple copies in the set of θ particles and the use of MCMC ensure that the desirable asymptotic properties of the IBIS output are not violated; see (Chopin, 2002, 2004; Del Moral et al., 2006) for details on the relevant theory.

Hence, in order to fully define the IBIS algorithm, it necessary to provide a MCMC algorithm to sample from the posteriors $\pi(\theta|\mathbf{y}_{1:i})$ for all i. Note that the standard Gibbs sampler of (Geweke and Zhou, 1996b) is not immediately suitable for this purpose as it returns samples from $\pi(\theta, \mathbf{z}|\mathbf{y}_{1:i})$. We therefore proceed with Hamiltonian MCMC targeting the posterior based on the likelihood in (3.4) where the latent factors have been marginalised out. The Hamiltonian MCMC algorithm can be applied using the standard publicly available platform Stan (Carpenter et al., 2017). The code used for this paper, which is provided in the accompanying repository[1], combines the IBIS algorithm with the use of PyStan, the Python interface of the Stan language. The fact that the MCMC and IBIS target $\pi(\theta|y)$, as opposed to $\pi(\theta, \mathbf{z}|\mathbf{y}_{1:i})$, does not imply that it is no longer possible to explore the posterior of the latent factors. Note that under the model of (3.3) and for all $i \leq j \leq n$

$$\pi(\mathbf{z}_i|\mathbf{y}_{1:j}) \propto \pi(\mathbf{z}_i|\theta, \mathbf{y}_{1:j})\pi(\theta|\mathbf{y}_{1:j}) = \pi(\mathbf{z}_i|\theta, \mathbf{y}_i)\pi(\theta|\mathbf{y}_{1:j}),$$

which for the case of $\Phi = I_k$ is $(\mathbf{z}_i|\theta, \mathbf{y}_i) \sim N\big((I_k + \Lambda^T\Psi^{-1}\Lambda)^{-1}\Lambda^T\Psi^{-1}\mathbf{y}_i\ ,\ (I_k + \Lambda^T\Psi^{-1}\Lambda)^{-1}\big)$, see e.g. Geweke and Zhou (1996b); Lopes and West (2004). A similar expression is available for the general case. Hence the IBIS output can be used to transform the θ particles into $\mathbf{z}_i$ particles by simply drawing from the full conditional above for each θ^m given a data point $\mathbf{y}_i$.

From a computational point of view, the most expensive step of the IBIS algorithm is the jittering step that requires to run an MCMC routine for a few iterations per θ particle. This is roughly equivalent to running several MCMC algorithms based on the smaller dataset $\mathbf{y}_{1:i}$ for some i. Nevertheless, note that jittering is more likely to occur when, in the transition between $\pi(\theta|\mathbf{y}_{1:i-1})$ and $\pi(\theta|\mathbf{y}_{1:i})$, these two posteriors are substantially different. As a consequence, most of jittering steps tend to take place for small is, where the learning curve is steeper, and

[1] https://github.com/bayesways/smc2

become less frequent as i increases. This suggests that the computational cost of the IBIS algorithm is typically larger than running a single MCMC algorithm on the full data $y_{1:n}$ but usually not by much. The difference can often be eliminated by using parallel computing; more specifically running the MCMC chains of the jittering step in parallel for each θ particle. For more coding details see Appendix B.1 and the accompanying repository.

3.3 Sequential Monte Carlo methods for Binary data

3.3.1 Model, Priors and MCMC Scheme

Binary and ordinal type data can be accommodated by extending the model framework used for continuous data and viewing the categorical responses as 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 in the general form then becomes:

$$\mathbf{y}_i^* = \alpha + \Lambda \mathbf{z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{3.8}$$

For binary data, the connection between the observed binary variable y_j and the underlying variable y_j^* is $y_j = \mathcal{I}(y_j^* > 0)$. Specific choices for the distribution of the error term $\boldsymbol{\epsilon}$ lead to the following well known models:

$$\boldsymbol{\epsilon} \sim \begin{cases} N(0_p, \Psi), \quad \Psi = \text{diag}(\psi_1^2, \ldots, \psi_p^2), & \text{if } \mathbf{y}_i \text{ is continuous} \\ N(0_p, \Psi), \quad \Psi = I_p, & \text{if } \mathbf{y}_i \text{ is binary and the probit model is adopted} \\ \prod_{j=1}^{J} \text{Logistic}(0, \pi^2/3), & \text{if } \mathbf{y}_i \text{ is binary and the logit model is adopted,} \end{cases} \tag{3.9}$$

where 0_p is a p-dimensional vector of zeros and I_p denotes the identity matrix of dimension p. The marginal distribution of the underlying variable becomes:

$$\mathbf{y}_i^* \sim N(\alpha, \Lambda \Phi \Lambda^T + \Psi). \tag{3.10}$$

The expression above is equivalent to the following

$$\begin{cases} \mathbf{y}_i \sim \prod_{j=1}^{p} \text{Bernoulli}\big(\pi_{ij}(\eta_{ij})\big) \\ \pi_{ij}(\eta_{ij}) = \sigma(\eta_{ij}) \;\text{ or }\; \pi_{ij}(\eta_{ij}) = \Phi(\eta_{ij}), \;\; \eta_{ij} = [\boldsymbol{\eta}_i]_j \\ \boldsymbol{\eta}_i \coloneqq \alpha + \Lambda \mathbf{z}_i, \\ \mathbf{z}_i \sim N(0, \Phi) \end{cases} \tag{3.11}$$

where $\sigma(\cdot)$ denotes the sigmoid function and leads to the logit model, whereas $\Phi(\cdot)$ denotes the cumulative density function of the standard Normal distribution and leads to the probit model.

Priors are set similarly to the continuous data case described in Section 3.2. Parameters α and Φ are treated identically. However, the unit information prior for the loading parameters Λ could be different. In the case of the 2PL IRT model this translates to a $N(0, 4)$ prior (Vitoratou et al., 2014), which is the one we adopt for most models as well. Finally, in the binary data case there is no longer a matrix Ψ to consider under formulation (3.11).

In order to sample from the posterior given all the available data, the Gibbs sampler is only available for the case of the probit link (Chib and Greenberg, 1998b), whereas a Metropolis withing Gibbs algorithm has also been used, see Vitoratou et al. (2016) and the references therein. Aiming for an efficient and general option, we again resort to Stan, which can in principle be used for all such models. Stan was also used in Vamvourellis et al. (2021) where it performed reasonably well. As described in the previous Section, an efficient MCMC scheme is an essential ingredient for sequential schemes, the topic we turn to next.

3.3.2 Sequential Scheme

The IBIS approach taken in the continuous data is not directly relevant for the case of binary data because we cannot integrate the latent factor to obtain a likelihood of the form $f(y|\theta)$ for categorical data, except for the case of the probit link. Instead, we can easily evaluate the augmented likelihood $f(y \mid z, \theta)$ according to (3.9) where the underlying variable is given by (3.11) depending on the model chosen. There are two potential routes in order to construct a

sequential Monte Carlo scheme that includes both parameters θ and latent variables $\mathbf{z}_{1:i}$. The first is to consider an IBIS algorithm on the higher dimensional parameter vector, that includes both parameter and latent variables, and explore the properties of factor analysis models to mitigate potential issues caused by the increased dimension. An example application of the IBIS algorithm in high dimensions is provided by (Kantas, Beskos, and Jasra, 2014). The second route is to pursue the development of a scheme in the spirit of the SMC^2 of (Chopin et al., 2012). SMC^2 focuses mostly on the case of Markov dependent latent variables and combines the IBIS algorithm with the pseudo marginal framework of Andrieu and Roberts (2009) and, more specifically, the particle MCMC algorithm (Andrieu, Doucet, and Holenstein, 2010). In this paper we proceed along the lines of the former route, by constructing an efficient IBIS algorithm on the augmented parameter vector $\Theta = (\theta, \mathbf{z}_{1:n})$, enhanced with importance sampling targeting the posterior of the latent factor $\mathbf{z}_{1:n}$ based on the Laplace or Variational Bayes approximations (Blei, Kucukelbir, and McAuliffe, 2017). Note that the developed approach requires a single z particle for each θ particle, as opposed to N_z particles in the SMC^2 framework.

Getting to the specifics of the IBIS algorithm, we take Algorithm 1, formulated on Θ rather than θ, as starting point and present our modifications that lead to the Algorithm 2, which we recommend in this paper for the case of binary data. The initialisation in the case of Algorithm 1, requires to draw samples from the priors of θ and $\mathbf{z}_i$ for $i = 1, \ldots, n$. Note however, that at time i of the IBIS the latent factors $\mathbf{z}_j$ for $i < j < n$ do not contribute at all to the algorithm; neither in the incremental weights nor in the likelihood for the MCMC in the jittering step. They can therefore be omitted and drawn retrospectively when the algorithm reaches $i = j$.

One of the key features regarding the efficiency of the IBIS algorithm is the number of times the degeneracy criterion is triggered. Under the data tempering schedule the ESS is being reduced at time i due to the differences between the posteriors based on $\mathbf{y}_{1:i-1}$ and $\mathbf{y}_{1:i}$. When it comes to θ such differences typically become smaller as i increases. This is not the case, however, for latent factors $\mathbf{z}_i$s. As noted earlier, for factor analysis models, $\pi(\mathbf{z}_i|y_{1:n}, \theta) = \pi(\mathbf{z}_i|\mathbf{y}_i, \theta)$, which implies that given θ the learning regarding $\mathbf{z}_i$ takes place only at time i. Hence, if the posteriors $\pi(\mathbf{z}_i|\mathbf{y}_i, \theta)$s tend to be substantially different than the priors, the degeneracy criterion of Algorithm 1 may end up being triggered at all times, thus leading to a very inefficient

computational scheme. The problem can potentially be addressed by replacing data tempering with another schedule, but this will no longer enable desirable features such as sequential testing and evaluation of scoring rules. We therefore proceed by retaining the data tempering schedule and resort to importance sampling to address this issues. More specifically, rather than drawing each $\mathbf{z}_i$ from its prior, which leads to incremental weights $f(\mathbf{y}_i|\mathbf{y}_{1:i-1}, \theta, \mathbf{z}_i) = f(\mathbf{y}_i|\theta, \mathbf{z}_i)$, we draw them from a proposal distribution $q(.)$ and compute the weights according to

$$u_i(\mathbf{z}_{1:i}^m, \theta^m) = \frac{f(\mathbf{y}_i|\theta^m, \mathbf{z}_{1:i}^m)\pi(z)}{q(\mathbf{z}_i^m|\theta^m, \mathbf{y}_{1:i})} \tag{3.12}$$

We seek a proposal $q(\cdot)$ that resembles the posterior $\pi(\mathbf{z}_i|\theta^m, \mathbf{y}_{1:i}) = \pi(\mathbf{z}_i|\theta^m, \mathbf{y}_i)$. The density of this posterior is not available in closed form but it is typically low dimensional and its likelihood function consists of a single data point. Hence, it is not hard neither computationally expensive to obtain approximations of it, such as the Laplace method which is easy to program in the presence of the relevant derivatives; see Appendix B.2. Another option is provide by Variational Bayes (VB) (Kingma and Welling, 2013; Kucukelbir et al., 2017b) to derive a distribution that approximates the posterior $\pi(\mathbf{z}_i|\mathbf{y}_i, \theta)$. The VB method has the advantage of not requiring the model specific expressions of the derivatives and can be automated.

We now summarise and present the full process in Algorithm 2. The main differences of this IBIS algorithm when contrasted with the Algorithm 1 is the augmentation of the particles with latent variables, thus having $\{\Theta^m\}_{m=1}^{N_\Theta}$ particles, the usage of the augmented likelihood in (3.3) instead of the marginal in (3.4) and the incremental addition of the latent factor particles.

3.4 Applications

We perform sequential parameter inference and model choice in simulated and real datasets considering cases of continuous and categorical data. We start with simulations of continuous data in Section 3.4.1 to show that the final posterior distribution are correctly recovered. We then extend those results to the case of binary data in Section 3.4.2 where we also compare the suggested methodologies from Section 3.3. Section 3.4.3 focuses on Bayes Factors for sequential

Algorithm 2 IBIS-LVM

Sample θ^m from $\pi(\theta)$ and set $\omega^m = 1$ for $m \in 1 : N_\Theta$. All operations are assumed to be repeated for all $m \in 1 : N_\Theta$.
Then at time $i = 1, \ldots, n$ do:

1: Sample $\mathbf{z}_i^m \sim q(\mathbf{z}_i|\mathbf{y}_i, \theta^m)$
2: Append $\mathbf{z}_i^m$ to $\mathbf{z}_{1:i-1}^m$ to maintain the matrix of latent variables (if $t = 1$ initialise matrix $\mathbf{z}_{1:1}^m = \mathbf{z}_1^m$)
3: Compute the incremental weights and their weighted average

$$u_i(\theta^m, \mathbf{z}_i^m) = \frac{f(\mathbf{y}_i|\theta^m, \mathbf{z}_i^m)\pi(\mathbf{z}_i^m|\theta^m)}{q(\mathbf{z}_i^m|\mathbf{y}_i, \theta^m)}, \quad L_i = \frac{1}{\sum_{m=1}^{N_\theta} \omega^m} \times \sum_{m=1}^{N_\theta} \omega^m u_i(\theta^m, \mathbf{z}_i^m),$$

4: Update the importance weights

$$\omega^m = \omega^m u_i(\theta^m, \mathbf{z}_i^m)$$

5: **if** ESS$(\omega) < \gamma$ **then**
6: **procedure** RESAMPLE$(\theta^m, \mathbf{z}_{1:i}^m, \omega)$
7: **return** θ^m, $\mathbf{z}_{1:i}^m$
8: **procedure** JITTER$(\theta^m, \mathbf{z}_{1:i}^m, \mathbf{y}_{1:i})$ using an MCMC algorithm
9: **return** $\tilde{\theta}^m, \tilde{\mathbf{z}}_{1:i}^m$
10: $(\theta^m, \mathbf{z}_{1:i}^m, \omega^m) = (\tilde{\theta}^m, \tilde{\mathbf{z}}_{1:i}^m, 1)$

model choice in different scenarios including choosing the number of factors. Finally, Section 3.4.4 contains a real data example from the British Household Panel Survey.

3.4.1 Continuous Data Simulations

We generated data $\mathbf{y}$ from the model of equation (3.3) with two factors according to the following process, denoted as 'Continuous Scenario 1': the sample size n is 200, and α is a vector of zeros, the factor scores $\mathbf{z}_i$s were generated from the $N(0, \Phi)$, whereas the error terms were distributed according to the $N(0, \Psi)$ with the associated parameters Λ, Ψ and Φ being as

follows

$$\begin{aligned}
\Psi &= \text{diag}(0.35, 0.58, 0.58, 0.35, 0.58, 0.58) \\
\Phi &= \begin{pmatrix} 0.65 & 0.13 \\ 0.13 & 0.65 \end{pmatrix} \\
\Lambda &= \begin{pmatrix} 1.0 & 0 \\ 0.8 & 0 \\ 0.8 & 0 \\ 0 & 1 \\ 0 & 0.8 \\ 0 & 0.8 \end{pmatrix}
\end{aligned} \tag{3.13}$$

A CFA model was fit to the simulated data where the loading matrix structure mirrors the loading matrix used in the data generation process, in other words the locations of zeros were assumed to be known but the remaining loading were considered as unknown parameters. Moreover, the residual covariance matrix was restricted to be diagonal.

We performed inference using the IBIS algorithm as presented in Section 3.2 and recovered posterior samples for all parameters at each step $i = 1, \ldots, n$. In Figures 3.1, 3.2, 3.3, and 3.4, we plot the 95% credible intervals after processing the last data point ($i = 200$) and verify that they include the correct values (marked with red dots). Note that for identifiability reasons the leading loading parameters were fixed to 1, hence the free loading parameters to be estimated in this model are the 4 elements in the positions containing the values 0.8 in Λ matrix in (3.13). We also fit the same model using a batch MCMC algorithm to the same exact data set and compared the posterior draws. More specifically, we compared the posterior draws of the sequential IBIS methodology with the posterior draws from an MCMC batch algorithm that was fit to the full dataset all at once. We confirmed that the posterior draws are essentially identical, which demonstrates the correctness of the sequential methodology. In Figures 3.5, 3.6, 3.7, and 3.8, we present the density plots of the loadings parameters overlaying the density charts for both algorithms. We see that the charts coincide almost completely, which shows

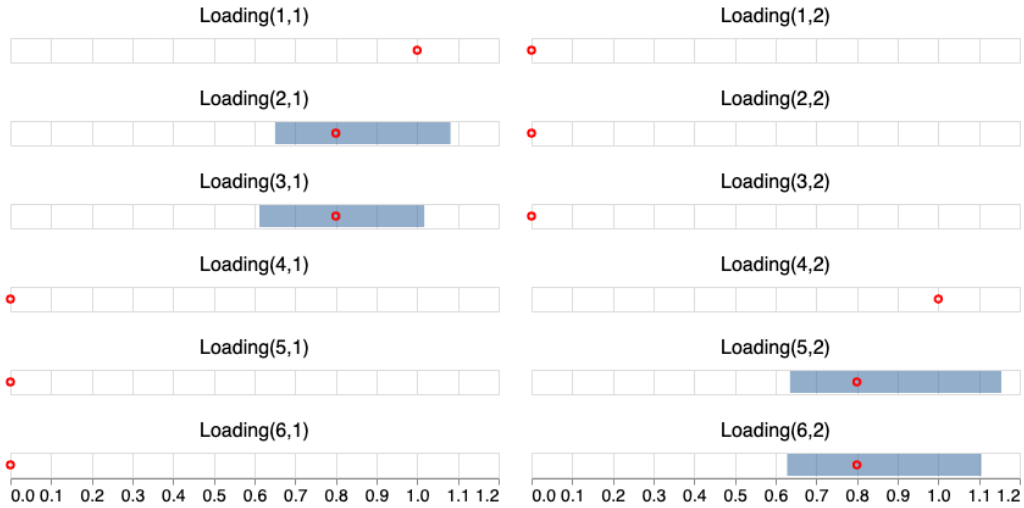

Figure 3.1: Real data values marked in red dots overlaid with the 95% credible intervals for Λ, the loading matrix parameters in Continuous Scenario 1.

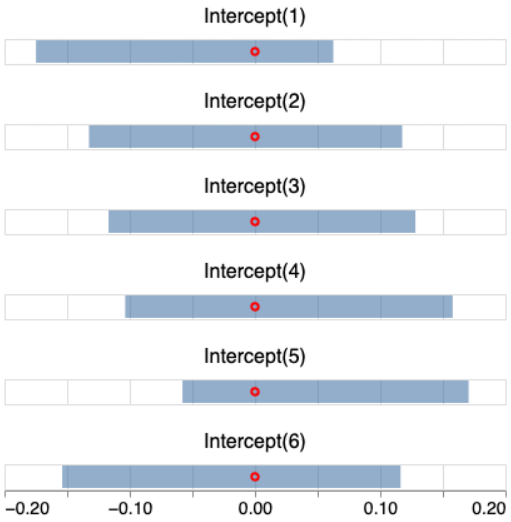

Figure 3.2: Real data values marked in red dots overlaid with the 95% credible intervals for α, the intercept parameters in Continuous Scenario 1.

that at the end of the process the sequential algorithm produces the same posterior density plots as the standard MCMC batch algorithm.

3.4.2 Parameter Estimation for Binary Data

We now proceed to demonstrate the sequential inference algorithm for a case of simulated binary data. We run three inferential processes on the same data set, each time using one of the three methodologies presented in Section 3.3.2, sampling the latent variables from their prior,

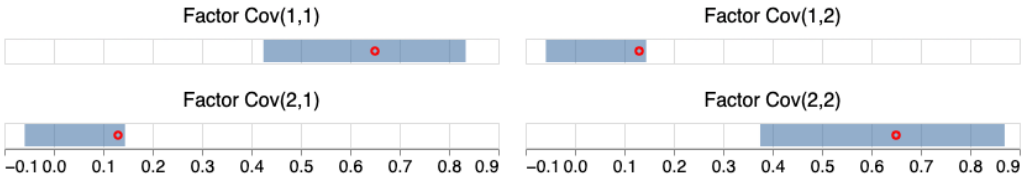

Figure 3.3: Real data values marked in red dots overlaid with the 95% credible intervals for Φ, the factor covariance parameters in Continuous Scenario 1.

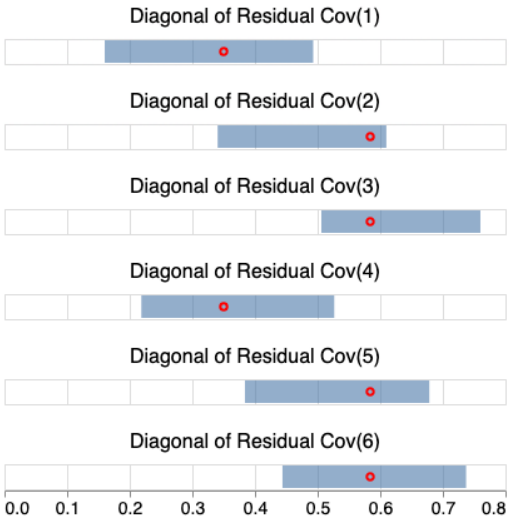

Figure 3.4: Real data values marked in red dots overlaid with the 95% credible intervals for diag(Θ) the diagonal elements of the residual covariance parameters in Continuous Scenario 1.

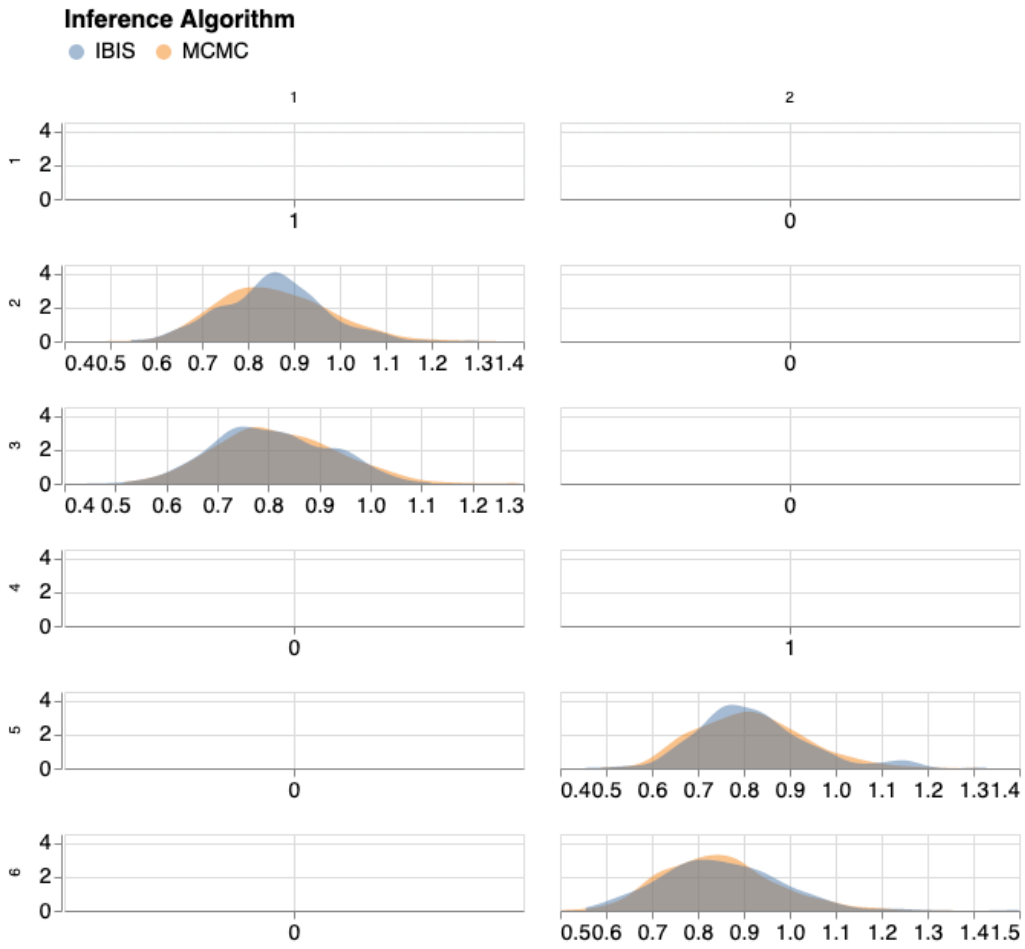

Figure 3.5: Posterior Draws for the Loading matrix Λ in Continuous Scenario 1, for both IBIS and batch MCMC methodologies.

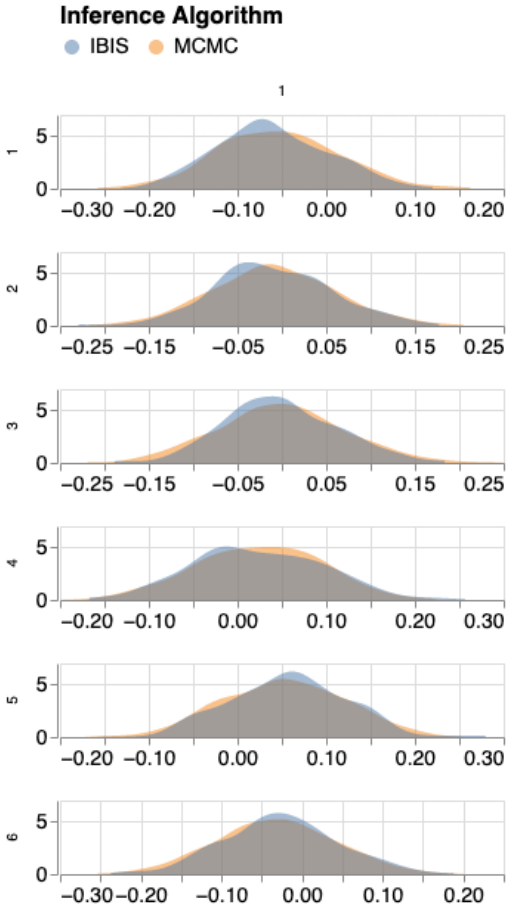

Figure 3.6: Posterior Draws for the Intercept parameters α in Continuous Scenario 1, for both IBIS and batch MCMC methodologies.

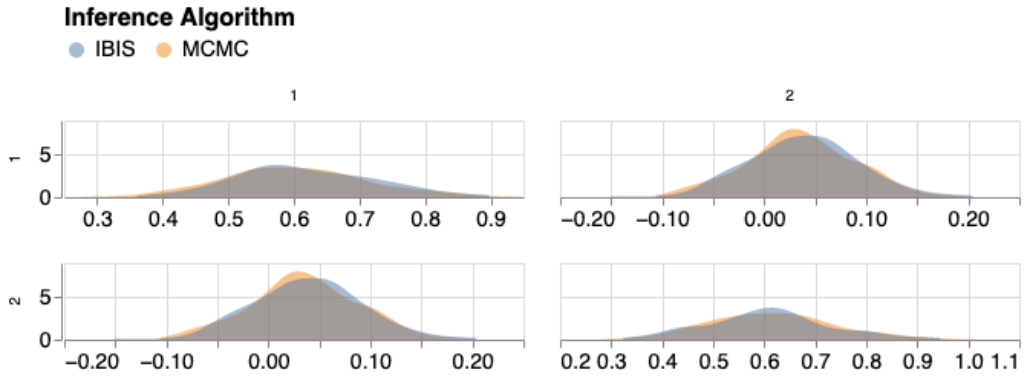

Figure 3.7: Posterior Draws for the Factor Covariance matrix Φ in Continuous Scenario 1, for both IBIS and batch MCMC methodologies.

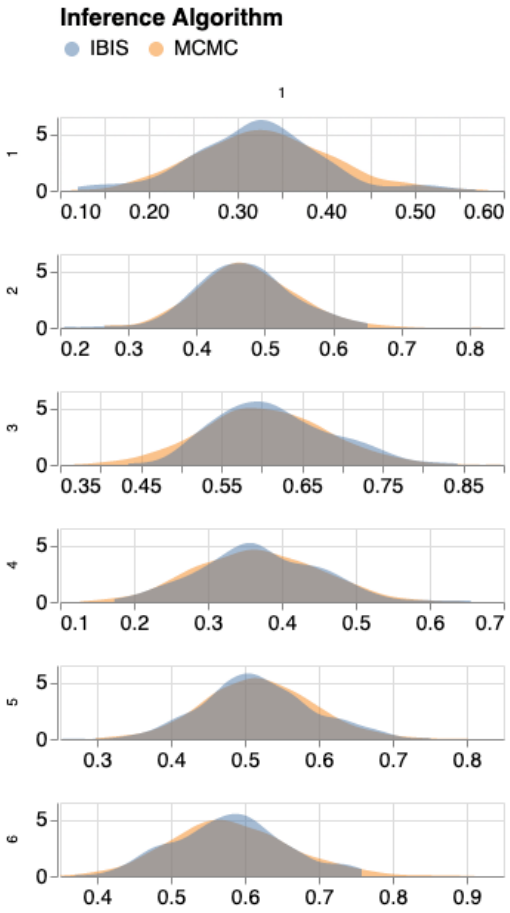

Figure 3.8: Posterior Draws for the diagonal of the covariance matrix of the residual Errors diag(Θ) in Continuous Scenario 1, for both IBIS and batch MCMC methodologies.

the Laplace approximation of their posterior, and the Automatic Variational Bayes (ADVI) approximation (Kucukelbir et al., 2017b) of their posterior. The data set was generated from a single factor model based on formulation (3.11) adopting the logit link and using the following parameter values: $n = 100$, $\alpha = (-0.53, 0.35, -1.4, -1.4, -0.96, -2.33)$, the loading matrix $\Lambda = (1, 1, 1, 1, 1, 1)$, and the factor scores were generated as $z \sim N(0, 1)$. We call this data generation process 'Binary Scenario 1'.

The simulations demonstrate that the naive approach of simulating from the prior, denoted by 'PRIOR' is inefficient. We contrast PRIOR to the two alternative methods we propose in this paper, namely the Laplace approximation method (LAPLACE) and the Variational Bayes method (VB). The PRIOR method is relatively easy to implement because the latent variables are drawn from the prior. The drawback of this method is that because the latent particles come from the prior, very few of them achieve high likelihood values. As a result, we need to refresh the particles very often by running a full MCMC chain for each particle, which is typically the most costly step in a sequential algorithm. In the 100 data point data set we used, the ESS criterion was triggered 57 times under the PRIOR scheme. The problem is particularly acute because the resampling and jittering rate does not necessarily drop as the data index i increases, contrary to the other methods we examined. More specifically, the degeneracy criterion was triggered 30 times in the first 50 data points and 27 in the last 50 data points; in other words the rate fell by only 10% in the second half. The LAPLACE methodology requires extra work to derive the derivatives but is far more efficient in terms of resampling rate. More specifically, the ESS criterion was triggered only 27 times in the 100 data points. More importantly, the resampling and jittering rate in the second half of the data set was (9/50) whereas in the first half it was (18/50), thus resulting in 50% reduction. Regarding the additional run time cost of computing the Laplace approximation, we found it to be minor in our simulations. Informally, we note that in the experiments we ran the run time of LAPLACE method was only 10-20% higher compared to the PRIOR method, in the order of added minutes. A potential drawback of this scheme is that the Laplace approximation requires the analytical derivatives, which may not be simple to derive. A more automatic approach is offered by the VB method which does not require the user to manually derive the derivative

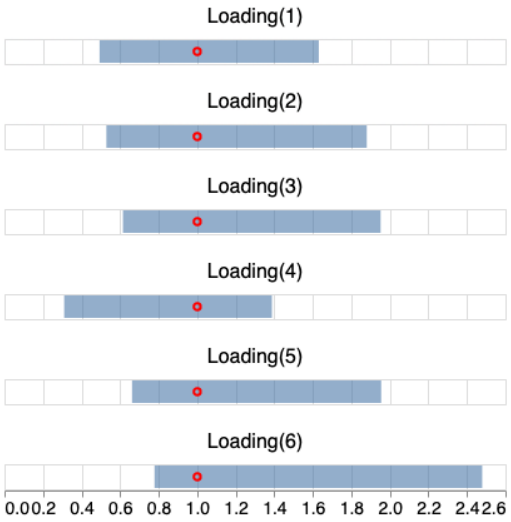

Figure 3.9: Real data values marked in red dots overlaid with the 95% credible intervals for the loading matrix parameters Λ in the Binary Scenario 1 data simulation.

formulas needed for the Laplace approximation. The efficiency of VB in our simulations was exactly the same as that of LAPLACE, with the ESS criterion being triggered at 18/50 times in the first half and 9/50 times in the second half of the data points. One possible drawback of the VBA is the extra run time needed to compile and run the VBA step for each draw. While the exact run time is highly dependent on the software and hardware used, to reduce run time we advise using a language that can run compiled models in order to amortise the compile time of the model across the rest of the resampling steps. In Figures 3.9 and 3.10 we plot the 95% credible intervals for the loading parameters along with the real data values and verify that the credible intervals cover the true values used to generate the data. Additionally, just as we did with the continuous data simulations, we fit both the batch MCMC and the sequential algorithm to the same dataset and compare the posterior draws. We verify that the draws of the sequential framework at the last step coincide with those from the batch MCMC algorithm fit to the entire data set. We present the comparison in Figures 3.11 and 3.12. For simplicity we show the output of the LAPLACE method only since the rest of the methods produce similar results.

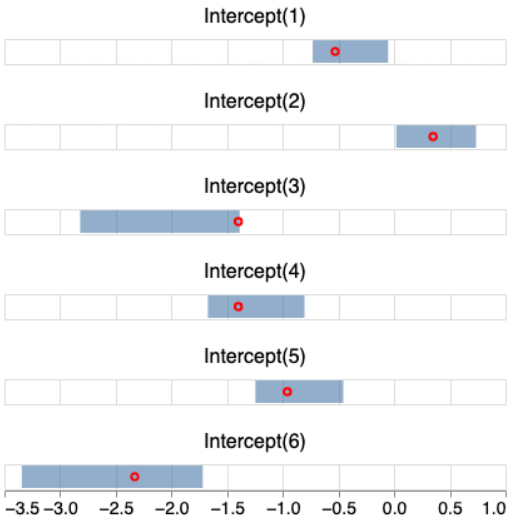

Figure 3.10: Real data values marked in red dots overlaid with the 95% credible intervals for the intercept parameters α in the Binary Scenario 1 data simulation.

3.4.3 Sequential Model Choice

3.4.3.1 Overview

In this Section, and also the next one, we demonstrate the use of Sequential Bayes Factors in order to facilitate model choice. We include special considerations for choosing the right number of factors. After we have briefly introduced the concept, we lay out the range of candidate models. We then proceed to run simulations using synthetic continuous data, which allows us to check our approach since we know the best model in advance.

The fully Bayesian approach to model choice is based on computing the model evidence and the probability of the observed data under the model of choice. The evidence of model $\mathcal{M}$, also called marginal likelihood, is defined as follows

$$\pi(y|\mathcal{M}) = \int f(y|\theta, \mathcal{M})\pi(\theta, \mathcal{M})d\theta \tag{3.14}$$

where $f(y|\theta, \mathcal{M})$ is the likelihood function under the model $\mathcal{M}$, and $\pi(\theta, \mathcal{M})$ is the prior distribution under model $\mathcal{M}$. When the model is implied, we omit it from the formulas for ease of notation. In the Bayesian framework, to compare two models $\mathcal{M}_1$ and $\mathcal{M}_2$ is to compare

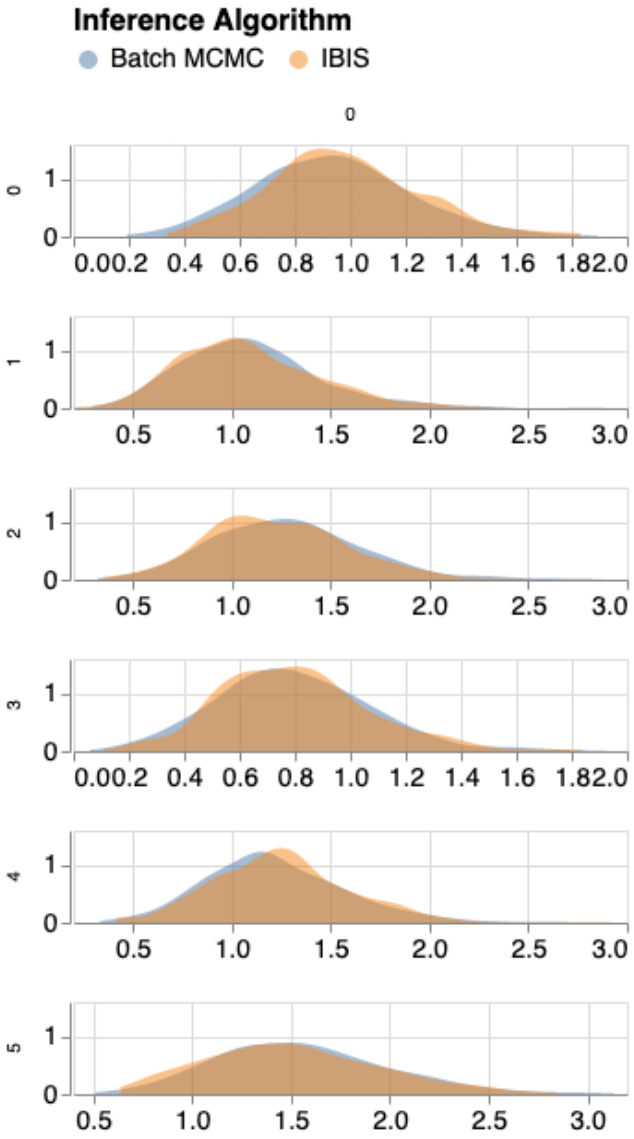

Figure 3.11: Posterior Draws for the Loading matrix Λ in for EZ model in the binary data simulation, for both IBIS and batch MCMC methodologies.

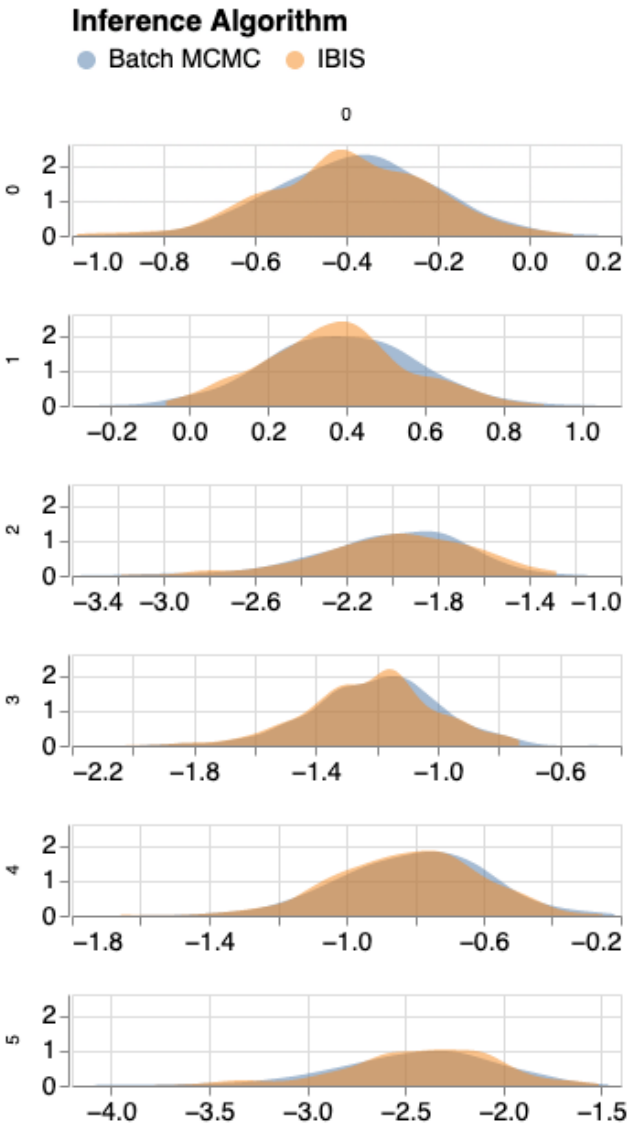

Figure 3.12: Posterior Draws for the intercept parameters α of EZ model in the binary data simulation, for both IBIS and batch MCMC methodologies.

their evidence. The ratio of their evidence is the Bayes Factor which we will abbreviate as

$$BF(\mathcal{M}_1/\mathcal{M}_2) := \frac{\pi(y|\mathcal{M}_1)}{\pi(y|\mathcal{M}_2)}$$

The Bayes Factor of two models provides a numerical comparison as for which model is most supported by the data. For example if $\pi(y|\mathcal{M}_1)/\pi(y|\mathcal{M}_2) = 2$ that means that the probability of y under model $\mathcal{M}_1$ is twice bigger than under model $\mathcal{M}_2$. Jeffreys provided guidelines for qualitative interpretations of Bayes Factors (Kass and Raftery, 1995), according to which, a BF of 4 and above is substantial evidence that $\mathcal{M}_1$ is the optimal model amongst the two under the Bayesian framework.

We focus on the task of model choice within a range of factor models that span the spectrum from CFA to EFA which we describe in detail now. As described in Section 3.2, researchers routinely express a factor hypothesis by setting certain loading parameters to zero, while freeing the rest. Such 'exact zero' models, however, are very sensitive to model misspecifications and typically demonstrate poor fit in the presence of even small cross loadings, which are often found in real world data sets. For this reason researchers have questioned the suitability of such models, denoted by EZ henceforth, in research hypothesis testing (Stromeyer et al., 2015; Hoijtink and van de Schoot, 2018a). A proposed solution to this issue is to replace exact zero parameters with approximate zero (AZ) parameters through the use of prior distributions that are highly concentrated around zero. Such a model can be seen as the middle road between confirmatory and exploratory analysis. Specifically, in the classical confirmatory EZ models the loading matrix parameters are either free or zero. Freeing the zero parameters completely would result in an EFA model, so AZ offers a middle path where we free them while strongly constraining them to be near zero using appropriate prior distributions. The loading structures of all three models, exact zeros (EZ), approximate zeros (AZ), and exploratory factor model (EFA), are presented in Table 3.2. Note that the the first loading of each factor in EZ and AZ are set to 1 for identifiability reasons, as the scale of the latent variable is non-identifiable. Similarly, exact zero models impose a diagonal structure in the residual covariance matrix, whereas the AZ model allows off-diagonal elements to be free but constrained near zero. For more information

we refer the interested reader to Muthén and Asparouhov (2012) and Vamvourellis et al. (2021).

In the discussion above we have assumed that the true number of factors is known, since we generated the data ourselves using two factors. Hence, all three models so far (EZ, AZ and EFA) have been assumed to have two factors. In the absence of specific knowledge however, choosing the right number of factors remains an important challenge of applied factor analysis. To address this, we propose fitting EFA models with different number of factors and comparing their Bayes Factors. The most supported EFA model amongst them will likely be the one with the appropriate number of factors. That of course assumes a case of perfect model specification, as in Scenario 1. If there are cross loadings or correlated residual errors, such as in Scenario 2, it is possible that an EFA model with an extra factor would perform better because such an extra factor allows the model to accommodate the misspecifications. We demonstrate these cases by including the 1 factor EFA (EFA1) and the three factor EFA (EFA3) along with the two factor EFA (EFA2) and the rest of the models.

3.4.3.2 Simulation Experiments

Our starting point is the case study presented in Section 3.4.1 using data generated according to Continuous Scenario 1. We fit the exact zero model (EZ) which we deliberately chose to match exactly the structure of the model that we used to generate the data. Hence we expect that EZ is supported by this data more than any other model, and indeed we find that it is. In the next scenario, we examine how model performance metrics change in the presence of small model misspecifications, a process we denote by 'Continuous Scenario 2'. In particular, we generated data from the same model as before, except the loading structure introduces small cross loadings in three positions, as shown in Table 3.1.

Now we demonstrate the use of Bayes Factors for finding the optimal model amongst the ones considered under the Bayesian framework in the two scenarios discussed above. In Scenario 1 the EZ model matches exactly the data generation process so it is expected to be the best model. In Scenario 2, however, the loading structure imposed by the EZ does not exactly match the data generation process anymore. We expect that a more flexible factor model, such as the

Scenario 1		Scenario 2	
$\Lambda_{:1}$	$\Lambda_{:2}$	$\Lambda_{:1}$	$\Lambda_{:2}$
1	0	1	0
.8	0	.8	.3
.8	0	.8	0
0	1	0	1
0	.8	.3	.8
0	.8	.3	.8

Table 3.1: True factor loadings used in continuous data simulation.

EZ		AZ		EFA	
$\Lambda_{:1}$	$\Lambda_{:2}$	$\Lambda_{:1}$	$\Lambda_{:2}$	$\Lambda_{:1}$	$\Lambda_{:2}$
1	0	1	~ 0	x	x
x	0	x	~ 0	x	x
x	0	x	~ 0	x	x
0	1	~ 0	1	x	x
0	x	~ 0	x	x	x
0	x	~ 0	x	x	x

Table 3.2: Loading structure of all three models, x represents a free parameter, ~ 0 represents a free parameter with a prior distribution concentrated around 0.

EFA, would perform better. Such flexibility comes at the cost of higher risk of overfitting, which occurs when a model fits noise rather than systematic patterns in the data. Furthermore, EFA models generally have slightly more free parameters than CFA models, and thus may result in lower estimation accuracy compared to EZ model. Hence, it is not clear a-priori whether the EFA model is better or worse than the EZ model in the presence of cross loadings in the data. The middle grown candidate model is the AZ, which preserves some of the structure of the EZ model but with some added flexibility. In the presence of small cross loadings we expect that the AZ model will be better than the EZ, but it is not clear how it would fair against the EFA.

Below we summarise the results of Scenario 1 in Tables 3.3 and 3.4 as it stands at the end of the inferential process (at point $i = 200$). Table 3.3 shows the log model evidence for each model from which we can deduce that EZ is the optimal model amongst those considered, since it has the highest value. In table 3.4 we provide all the log Bayes Factors for each pair of models, which we will denote by LBF. For example the first column shows the log Bayes Factors $LBF(\text{EZ}/M) \coloneqq \log BF(\text{EZ}/M) = \log(\pi(y|\text{EZ})/\pi(y|M))$ for $M \in \{\text{AZ, EFA1, EFA2, EFA3}\}$. We can verify that all ratios are above 0 which is a different way of saying that EZ is supported by the data the most. The second best model is AZ, as in the second column we can see it has positive ratios with the all other models but EZ.

The table also shows us how we could have concluded that the right number of factors is two, in the absence of prior knowledge. We are looking for the EFA model that is most supported by the data amongst the three options. From Table 3.3 we can see that EFA2 has the highest value of the three EFA models. Alternatively, we can verify this from Table 3.4; $LBF(\text{EFA1}/\text{EFA2}) = -58.2 < 0$ which means that EFA2 beats EFA1; and also $LBF(\text{EFA2}/\text{EFA3}) = 6.6 > 0$ which means that EFA2 also beats EFA3.

Moving on, we now turn to the results in Scenario 2, where the data present cross-loadings. The results for Scenario 2 are summarised in Tables 3.5 and 3.6 where we can verify that the EZ model is no longer the optimal of the candidate models. In fact, the most supported model is now the AZ model since it has the highest marginal likelihood value in Table 3.5. This is a confirmation that the Bayes Factors are picking up on the fact that the exact zero model is

Name	Log(Model Evidence)
EZ	$\mathbf{-1330.98}$
AZ	-1331.27
EFA1	-1391.82
EFA2	-1333.67
EFA3	-1340.21

Table 3.3: Log Marginal Likelihood or Log Model Evidence for candidate models in simulation Scenario 1, at the final point $i = 200$. We highlight model EZ with the highest value.

	EZ	AZ	EFA1	EFA2	EFA3
EZ					
AZ	0.3				
EFA1	60.8	60.6			
EFA2	2.7	2.4	-58.2		
EFA3	9.2	8.9	-51.6	6.6	

Table 3.4: Log Bayes Factor for candidate models in simulation Scenario 1, at the final point $i = 200$. The table values represent the log ratio of the model on the top row divided by the model on the column. For example the $LBF(\text{EZ}/\text{AZ}) = 0.3$.

no longer the best match for the data generation process, given the presence of the small cross loadings. It is interesting to gain a qualitative insight into how much does the data support each of the three models EZ, AZ and EFA2. If we exponentiate the values presented on Table 3.6 we can read all four comparisons involving AZ from the table as follows:

$$BF(\text{AZ})/BF(\text{EZ}) = \frac{1}{\exp(LBF(\text{EZ})/\text{AZ}))} = \frac{1}{\exp(-9.8)} = 1\text{e}5$$

$$BF(\text{AZ}/\text{EFA2}) = \exp(2.6) = 13.4$$

We can verify that the hypothesised loading structure represented by EZ is too restrictive as AZ is clearly more supported. Additionally, because the cross loadings used to generate the data are relatively small, the loading structure is for the most part correct and for this reason the AZ model is strongly preferred also over the exploratory model with the correct number of factors EFA2.

Furthermore, we can note, that even in the presence of cross loadings we can deduce the right number of factors by observing which EFA performs best in terms of Bayes Factors. As in the previous scenario, EFA2 beats both EFA1 and EFA3, hence the evidence is pointing decisively towards EFA2 and two factors. Interestingly, EFA3 is also supported to a degree as the Bayes Factor $BF(\text{EFA2/EFA3}) = \exp(1.2) = 3.3$ is relatively low at the value of 3. That is probably happening because the additional factor is able to accommodate some of the cross loadings.

Name	Log(Model Evidence)
EZ	−1330.69
AZ	**−1290.85**
EFA1	−1347.83
EFA2	−1293.43
EFA3	−1294.61

Table 3.5: Log Marginal Likelihood or Log Model Evidence for candidate models in simulation Scenario 2, at the final point $i = 200$. We highlight model AZ with the highest value.

	EZ	AZ	EFA1	EFA2	EFA3
EZ					
AZ	−9.8				
EFA1	47.1	57.0			
EFA2	−7.3	2.6	−54.4		
EFA3	−6.1	3.8	−53.2	1.2	

Table 3.6: Log Bayes Factor for candidate models in simulation Scenario 2, at the final point $i = 200$. The table values represent the log ratio of the model on the top row divided by the model on the column. For example the $LBF(\text{EZ/AZ}) = -9.8$.

So far we have focused on model evaluation after having fit the model to the full batch of the data. However, the sequential framework we propose allows us to dynamically evaluate the models at each data point if we wanted. In practice it is not uncommon for sequential algorithms to be slow to reach a good set of particles if initiated at the first data point. One solution is to use the Adaptive Tempering algorithm which is designed to mitigate this issue exactly, see for example (Kantas et al., 2014; Schäfer and Chopin, 2013). Another solution is to initiate the sequential process particles using the output of a batch MCMC run using the first few data points. In this work, for demonstration purposes, we chose to initialise using

the first 30 data points for all our runs. In general, further work is required to choose the size of the batch used for the initialisation step. After the initialisation step, the particles achieve good values (values of high likelihood) and remain stable from then on. The inference output suffices to carry out the comparisons between the models we described above at each new data point we process.

We first address the question of how to choose the right number of factors to choose in Scenario 1. Recall that we initiate the sequential paradigm after having initialised the particles by running a batch MCMC on the first 30 points. We present the chart of log Bayes Factors between the EFA models in Figure 3.13 where the horizontal axis shows the index of the data point and the vertical axis shows the Log Bayes Factor. We can confidently claim that the right number of factors is two based on the fact that EFA2 outperforms easily both EFA1 and EFA3. We present the Sequential Log Bayes Factors amongst the rest of the models for Scenario 1 in Figure 3.14. We choose to present all the factors with EZ in the numerator and the other 2 models in the denominator. The chart highlights two interesting points. First, it shows that the AZ and EZ models present very comparable level of support throughout the dataset as their log factor ratio stays near zero throughout. Second, we see that the EZ model starts to outperform the best EFA model, EFA2, after about 50 points when their log ratio turns positive. However, it does not surpass the notional mark of 1.4 until it hits point 180. We note that $BF(\text{EZ/EFA2}) = \exp(1.4) = 4$ and recall that a Bayes Factor of 4 is considered substantial support in favour of the better model.

In Figure 3.15 we present the Sequential Log Bayes Factors for the EFA models in Scenario 2 in order to choose the right number of factors. We see that EFA2 outperforms EFA1 from the beginning but $LBF(\text{EFA2/EFA3})$ remains above 0 but below 2.5 for the rest of the dataset. This would be a sign that the right number of factors is probably 2, but there are additional model misspecifications that might call for an additional factor. More generally, when the data support an EFA model with K factors only slightly more an EFA with $K+1$ factors, it could be an indication that the true number of factors is K but there are model misspecifcations that justify using the EFA model with $K+1$ factors as well. We then present the comparisons of the rest of the models in Figure 3.16 where we choose to present all the ratios with AZ in

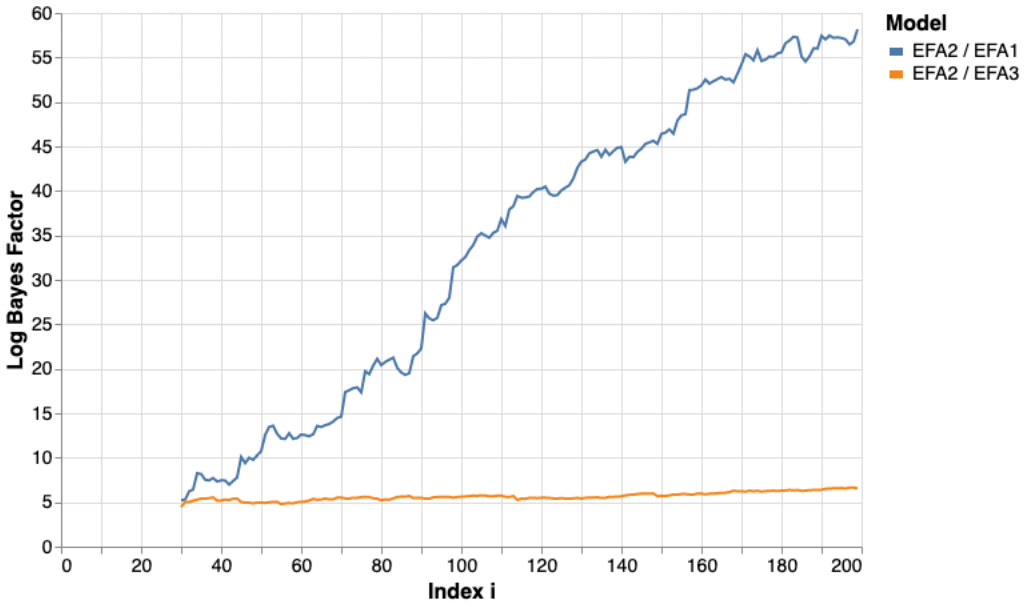

Figure 3.13: Bayes Factors for EFA models of 1, 2 and 3 factors respectively in Scenario 1. We see that EFA2 outperforms both of the other models, which is an indication that the right number of factors is 2.

the numerator. The EZ model remains competitive for approximately the first 100 points, and after that the log ratio grows decisively above 5. The ratios with EFA2 favours AZ in the range of $2-3$ throughout which confirms the interpretation we discussed in relation to Table 3.6.

3.4.4 Application: Big 5 Personality Test - British Household Panel Survey

In this Section we apply the proposed framework to a real word dataset to highlight the benefits of sequential model selection in the case where the true data generation process is unknown. The data set we work with comes from the British Household Panel Survey in 2005-06, which concentrated on female subjects between the ages of 50 and 55; the sample size consists of 676 individuals. The 'Big 5 Personality Test', as it is known, is a 15-item questionnaire on topics of social behaviour and emotional state. Each item receives an answer from each participant on a scale from $1-7$, 1 being 'strongly disagree' and 7 being 'strongly agree'. The test is meant to measure five major, potentially correlated, personality traits. Each factor corresponds to one trait, and is hypothesised to explain exactly 3 out of 15 items. Typically, as is the case in our

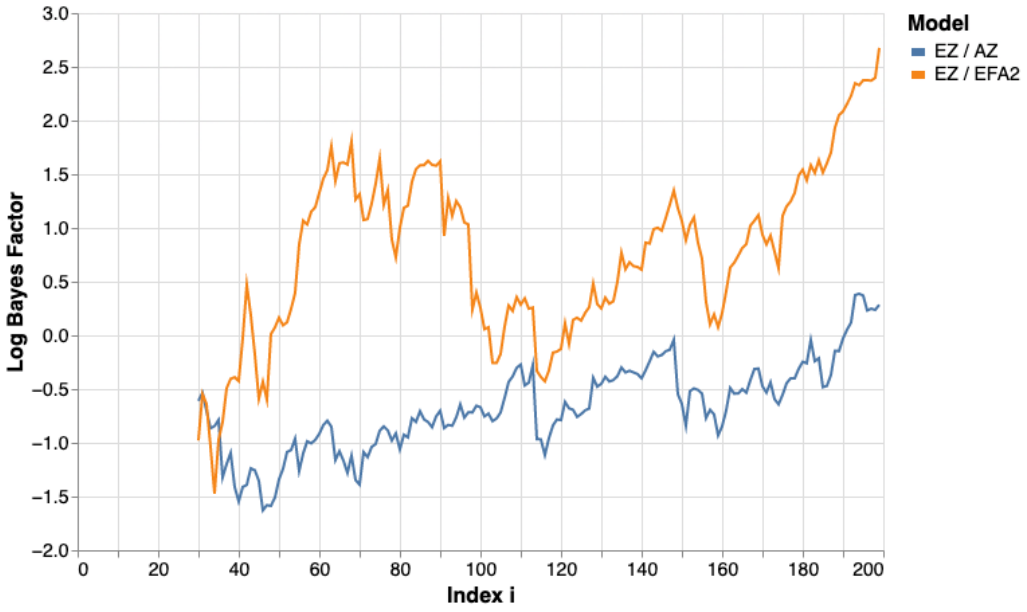

Figure 3.14: Bayes Factor values of 3 candidate models, in Scenario 1 where the data was generated from a structure with zero cross loadings. We expect EZ to be the model of choice since it matches the structure of the data generation model the best of all the candidate models.

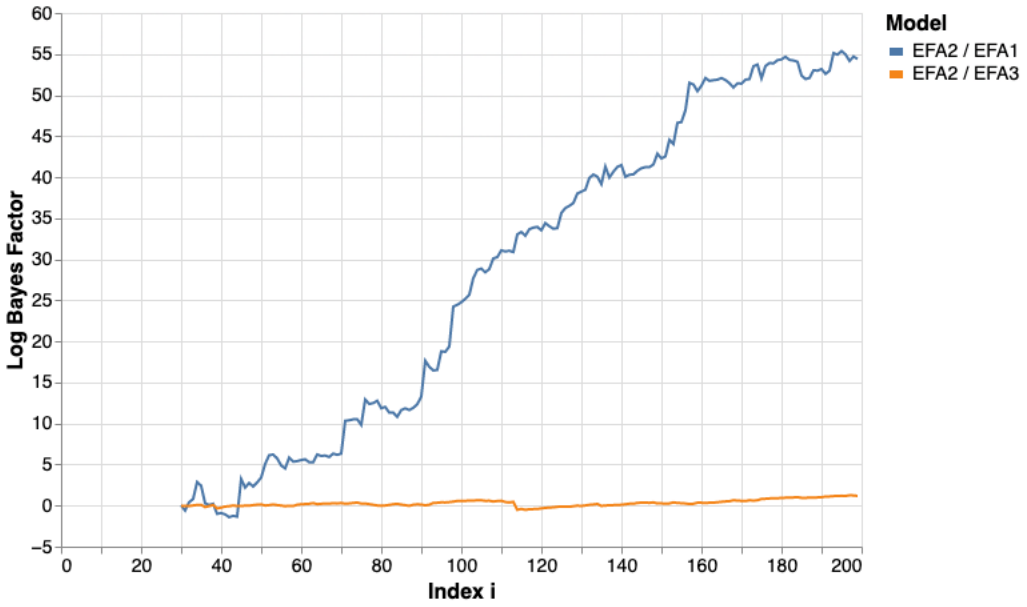

Figure 3.15: Bayes Factors for EFA models of 1, 2 and 3 factors respectively in Scenario 2. We see that EFA2 outperforms both of the other models, which is an indication that the right number of factors is 2. However, because of model misspecification the EFA model with an additional factor, EFA3, remains competitive.

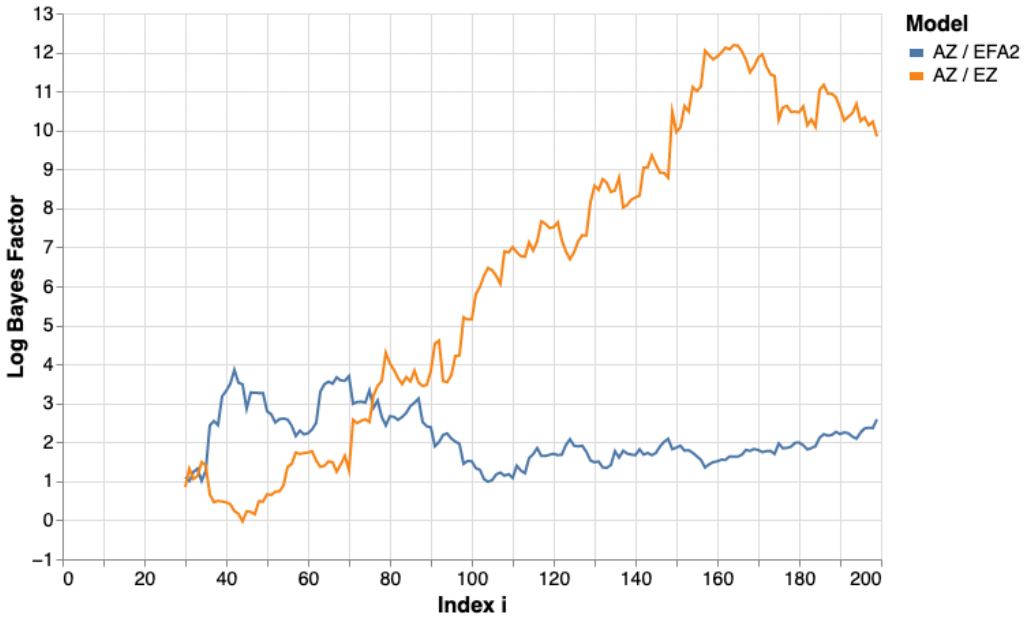

Figure 3.16: Bayes Factor values of the 4 candidate models, in Scenario 2 where the data was generated from a structure that included cross loadings. We expect AZ to be the model of choice since it matches the structure of the data generation model the best of all the candidate models.

analysis, the questions are ordered in a way such that the first three questions correspond to the first factor, the next three questions correspond to the second factor; the pattern continues so that the last three questions correspond to the fifth factor. For our analysis we standardised the data by removing the mean and set the standard deviation to 1 so that we receive standardised loading values.

Prior research has suggested that the dataset demonstrates potentially small cross loadings, as well as correlated residual errors, possibly as a result of negative wordings of some of the items in the questionnaire. As a result, the exact zero model has been found to have poor fit, while the equivalent AZ model fits the data much better (Muthén and Asparouhov, 2012; Vamvourellis et al., 2021). We are also interested in examining whether more flexible models are more supported by the data, to benchmark against the confirmatory models with the hypothesised structure. Since the questionnaire was constructed to measure 5 personality traits, we will include the 5 factor EFA model which will perform better than the rest if the hypothesised factor structure is not correct. Finally, we include the most flexible model possible, the saturated model denoted by 'SAT', that imposes no factor structure on the dataset.

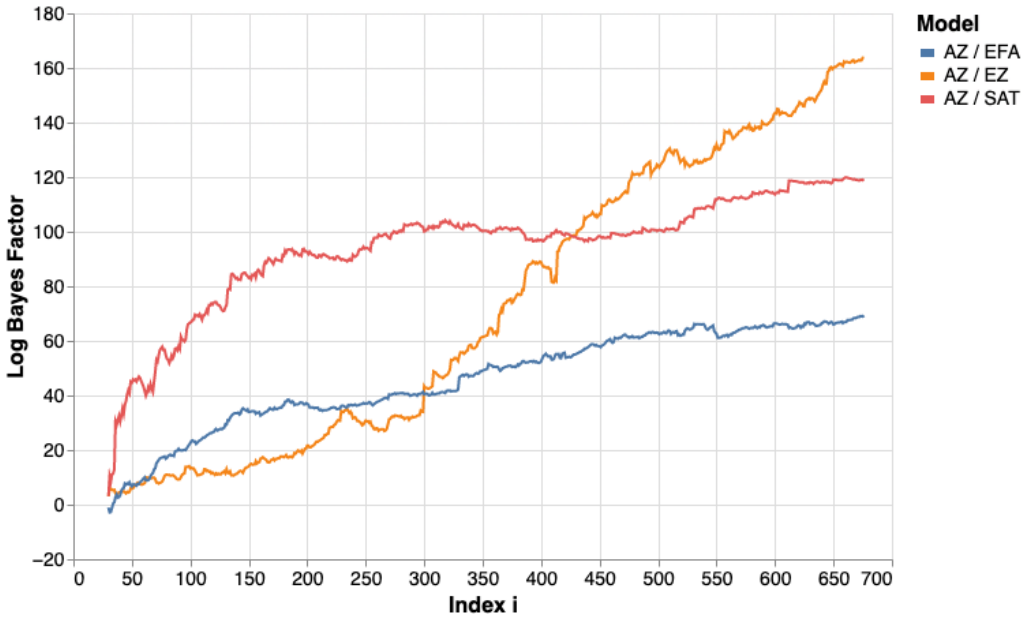

Figure 3.17: Bayes Factor values of the 4 candidate models. Based on prior research we expect AZ to be the model of choice.

In the sequential model comparison chart, Figure 3.17, it becomes clear early on that the hypothesised factor structure is correct. Even though the cross loadings and residual errors cause the EZ model to not perform optimally in terms of Bayes Factors, the overall loading structure is supported by the data. We can conclude this because the most supported model is the AZ model, which easily outperforms EFA and SAT, both of which are presumed to be more flexible. Their flexibility would result in higher model evidence if indeed the factor structure hypothesised in AZ was wrong. The fact that the most flexible models fail to perform better is an indication that the hypothesised structure of AZ is supported. The SBF framework can facilitate dynamic adjust to surveys, such as the one analysed here. As we can see,it becomes immediately clear that the AZ is better than the EZ, say after the first 80 participants. If the analysis was done in real time, the researchers could have stopped the study early and potentially revised the structure or the wording of the questionnaire.

Having selected the right model we can also draw posterior samples at each point and learn the loadings of the factors for each item. Here we present the posterior mean estimated loadings after the last data point has been processed in Table 3.7. We can see that some cross loadings, such as 1-st factor 8-th item are estimated to be non-zero, despite the strong priors towards zero. These items are prime candidates for sources of model misspecification. Overall, all such

cross loadings are estimated to be smaller than 0.1 in absolute value while the main loading structure takes values around 1. A posterior density plot of the draws, Figure 3.18 can reveal some further characteristics of the areas of misfit. For example, some cross loading densities, such as the the 4-th factor on the 14-th item, present substantial skewness to the right, which means that there is potentially a stronger source of cross loading than the cross loading of the 1-st factor on the 8-th item which is a more symmetric density plot.

$\Lambda_{:1}$	$\Lambda_{:2}$	$\Lambda_{:3}$	$\Lambda_{:4}$	$\Lambda_{:5}$
1.	0.	-0.1	-0.	0.
1.2	-0.	0.1	0.	0.
1.5	0.	0.	0.	-0.
0.	1.	-0.	-0.	0.
0.	0.9	0.1	0.	-0.
0.	1.3	-0.	-0.	-0.
-0.	0.	1.	-0.	-0.
0.1	0.	1.4	-0.	-0.
-0.	-0.	1.3	0.	0.
-0.	0.	0.	1.	0.
0.	-0.	-0.1	1.1	0.
-0.	-0.	0.	1.	-0.
-0.	0.1	0.	-0.1	1.
-0.	-0.	0.	0.1	1.3
0.	-0.	-0.	-0.	1.

Table 3.7: Posterior mean estimates of loading values for the Big 5 dataset at the final point of inference. We present the values at 1 decimal point.

3.5 Discussion

In this paper we propose an efficient sequential scheme for factor analysis with continuous latent variables. The modelling case of continuous Normally distributed data can be handled using the existing framework of IBIS (Chopin, 2002). The crucial fact is that it is possible to write down the distribution of the data, marginalising out the latent variables. However, when that marginal distribution is not available, as is the case in IRT models, the standard IBIS algorithm does not apply. We develop an efficient scheme based on IBIS that can handle categorical and non-normally distributed continuous data, using the Laplace and Variational

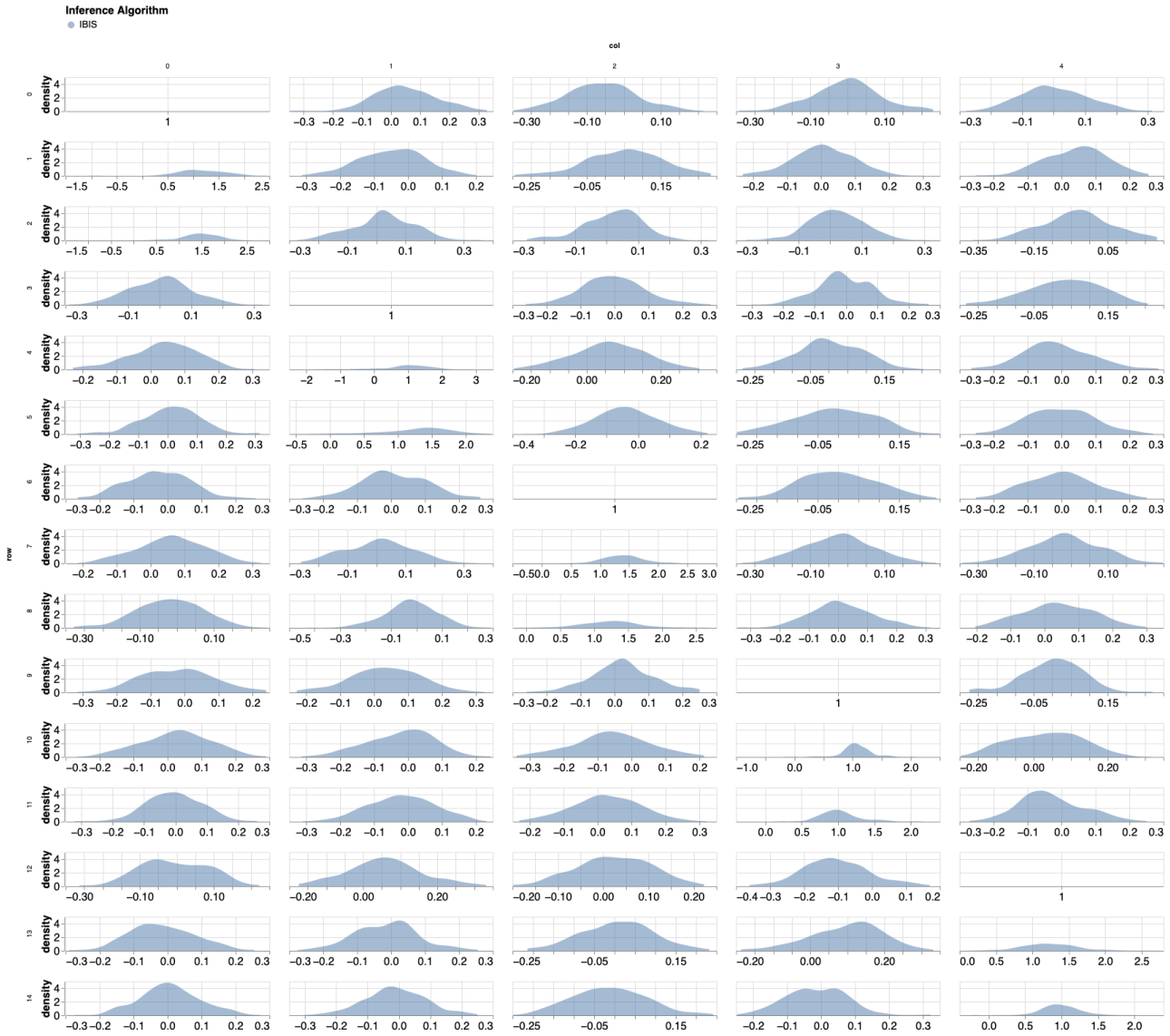

Figure 3.18: Posterior density plots for the loadings parameter of the AZ model fit to the Big5 dataset.

Bayes approximations. Other SMC^2 based approaches, while wholly valid, they tend to be inefficient in the context of factor analysis. The reason is that they are constructed to work with latent variables that present temporal correlations, whereas factor modelling usually involves independent latent variables. Our scheme takes advantage of the independence structure to offer an efficient alternative. Merging the fields of Sequential Monte Carlo methods and Factor Analysis could be a fertile area of research ahead. Additionally, future practitioners stand to benefit from more research into the implementation details of these frameworks. Moreover, we note that the sequential inferential framework proposed in this work provides a robust computational alternative to MCMC schemes for factor analysis even in the non-sequential setting. An interesting future research direction is a more formal performance comparison between the two approaches.

Finally, another area of future research is that of sequential model choice. The sequential framework proposed here facilitates model choice in two ways. While in this work we focus on Bayes Factors, scoring rules is a powerful alternative tool. Scoring rules assess a model via predictive performance, even when when prediction is not amongst the research objectives. The philosophical basis of this approach suggests that a model fits the data well when it is able to predict new data that was not used for estimating the model parameters. Scoring rules evaluated on out-of-sample data are available as a direct output of the data tempering strategy adopted in this work, under the prequential framework (Dawid and Musio, 2014). A formal comparison of Sequential Bayes Factors and such scoring rules is an interesting field of further research.

Chapter 4

Bayesian Benefit Risk Analysis

4.1 Introduction

Regulatory authorities called to authorise new drugs, often face a difficult challenge on how to choose an action in the presence of multiple competing objectives. Rising demand for individualised treatments, increasing number of stakeholders and higher level of complexity in the kind of questions asked today, all make the challenge even harder. Multiple high profile drugs have been withdrawn over the past 20 years (Guo, Pandey, Doyle, Bian, Lis, and Raisch, 2010), some permanently while others have been re-marketed under revised labelling and guidance (Wallach, Wang, Zhang, Cheng, Nardini, Lin, Bracken, Desai, Krumholz, and Ross, 2020). Given the complexity and the gravity of this challenge, it is important to provide stakeholders with structured quantitative approaches for benefit-risk assessment of medicines. Such approaches, as the one developed in this paper, typically rest on using observational data to estimate suitable models and the parameters thereof, while at the same time balancing competing objectives.

One high profile drug with a controversial past is rosiglitazone. Rosiglitazone is a drug for the treatment for Type II diabetes, that was marketed under the commercial name Avandia. It gained market authorisation in the United States in 1999 and in the European Union in 2000. New data subsequently emerged about possible cardiovascular risks associated with rosiglitazone, confirmed by a meta-analysis in 2007 (Nissen and Wolski, 2007), which resulted

in a European suspension of the Marketing Authorisation in 2010. This suspension included its use as a fixed dose combination with metformin or glimepiride for Type II diabetes, which had been approved in 2003 for metformin and 2006 for glimepiride. The drug remained available in the United States, but only under a restricted-access program put in place in 2010. In 2011 the US regulators reverted the recommendation and suspended the drug. In 2013 the drug regained approval, following a study that found rosiglitazone to be as safe as other diabetes drugs. Today the drug is withdrawn from most countries in the world, but remains available in the United States. For a complete review of regulatory history of rosiglitazone see (Wallach et al., 2020).

Benefit risk analysis (BRA) is an umbrella term that encompasses any structured approach for weighing the benefits (typically the efficacy) against the risks (typically dangerous or unwanted side effects) of drug treatments. The goal is to assist stakeholders decide if a drug works, and whether the potential side effects are acceptable. Since the decision depends on personal preferences and risk tolerances, the proposed frameworks include various ways of embedding formal or informal utility functions. Mt-Isa et al. (2014) conducted a review of the benefit risk literature categorising the approaches into quantitative frameworks, metrics for benefit-risk assessment, estimation techniques and utility survey techniques, and qualitative frameworks. Out of these four groups, the first two are formal quantitative approaches based on statistical theory, whereas the latter two are more informal. Common single metrics are the benefit risk ratio (BRR) (Carlisle et al., 2016) and net clinical benefit (NCB) (Shakespeare et al., 2001) whereas more recently researchers have suggested extension metrics such as RV-NNT and MCE (Holden, 2003). Single metrics like these, while providing useful summaries, in practice cannot holistically evaluate a drug as they focus instead on specific aspects of the decision making process. A practical limitation is that they require the benefit and risks to be already expressed in common scales. A theoretical limitation is that complex decision making often cannot be easily summarised by single numbers especially if the uncertainty is not taken into account. Ratio based metrics in particular, have received criticism for being unstable or misleading often hiding the uncertainty embedded into the final decision that rests on the actual value (Lynd and O'brien, 2004; Shaffer and Watterberg, 2006; Sutton et al., 2005).

Multi-criteria decision analysis (MCDA) is one of the most comprehensive quantitative frameworks for benefit risk assessment and it is widely used today (Keeney et al., 1993; Mussen et al., 2009; Dodgson et al., 2009). It is a quantitative framework that allows users to weight outcomes by their utilities and can support any integrated single metric score. It has been developed over the last 50 years (Jong and Stone, 1976) and has subsequently been used for assessing the benefit-risk balance of drugs (Glasziou and Irwig, 1995), or other interventions (Ponce, Bartell, Wong, LaFlamme, Carrington, Lee, Patrick, Faustman, and Bolger, 2000). In particular it has been recommended by Mussen et al. (2007) and Garrison Jr et al. (2007) as a necessary tool in the regulatory setting (Mühlbacher, Juhnke, Beyer, and Garner, 2016). MCDA also lends itself to sensitivity analysis around crucial variables, a methodology commonly referred to as SMAA (Tervonen and Figueira, 2008; Tervonen et al., 2011; Lahdelma et al., 1998) for which there is also a software package (Tervonen, 2014).

There has been some recent work on Bayesian modelling for MCDA analysis. Waddingham, Mt-Isa, Nixon, and Ashby (2016) conducted an analysis of multiple data sets, formed by both aggregated and patient level data, to incorporate uncertainty in the MCDA score. They propose a Bayesian probabilistic model to propagate the uncertainty caused by sampling variability to the final outcome of an MCDA study. The methodology is demonstrated using a model applied to 9 binomial variables and one count variable. The model assumes that all variables are independent, and does not include assessing the model fit. Li, Luo, Yuan, and Mt-Isa (2019) proposed a two factor latent trait model to account for correlation amongst outcome variables, accommodating a combination of continuous and binary outcomes of interest. Furthermore, their approach introduced the use of latent factors that represent the benefit and risks.

This recent line of research successfully demonstrates the benefits of Bayesian modelling for drug evaluations. Waddingham et al. (2016) used Bayesian modelling to incorporate parameter estimation uncertainty into the final MCDA score, and demonstrated the importance of taking that uncertainty into account for drug evaluation. Costa et al. (2017); Li et al. (2019) showed how to account for the correlation between continuous types of data, typically the treatment efficacy outcomes, and discrete types of data, typically adverse events. Naturally, in order to account for any correlation, such models are fit on individual level data, rather than summary

data, which is aligned with the recent efforts to create personalised treatment according to individual preferences. These modelling advances are important steps towards comprehensive drug evaluations. Since the final decision depends on parameter estimates from a model, it is crucial to be able to check if the proposed model is supported by the data before basing any decisions on the results of inference. At the same time assessing the predictive performance is essential in order to choose amongst competing models.

In this paper, we aim to enhance the available methodology in a few new directions. We build a framework that can accommodate multivariate sets of data that consists of categorical and continuous variables, henceforth denoted as mixed type. Importantly, we do not only model the marginal distributions of the observations, but also their dependencies. The first contribution expands the pool of models by considering Bayesian structural equation modelling (Muthén and Asparouhov, 2012; Vamvourellis et al., 2021), tailored to the data features of this paper. Given the increased number of models it becomes essential to develop schemes for model choice and assessment. In addition to account for the data dependencies, it is highly important to seek a parsimonious model in order to achieve good predictive performance and avoid overfitting. The second contribution of this paper offers a way to search between the available models by extending the methodology of Vamvourellis et al. (2021) to mixed type data. Thirdly, we introduce a sequential clinical study design paradigm that offers multiple benefits over standard batch clinical trials: (i) it allows us to recursively update model estimates with each new subject receiving the treatment; (ii) it permits stopping the exposure as soon as the research requirements are satisfied reducing unnecessary further exposure to undesirable treatments; (iii) it allows us to assign treatment groups dynamically based on research objectives, for example if the effects of one treatment are estimated with high confidence within a few data points while for another they remain uncertain, we can shift subjects away from the first group and into the second as needed. We note that while early stopping can be problematic in the frequentist hypothesis testing framework, due to the multiple comparisons involved, it is generally not a problem under the Bayesian framework (Schnuerch and Erdfelder, 2020; Dienes, 2016; Schönbrodt et al., 2017; Pramanik et al., 2021). From a computational point of view, this is achieved by developing an efficient Sequential Monte Carlo (SMC) scheme to accommodate

the models considered in this paper and facilitate assessment and choice thereof. Last but certainly not least, sequential algorithms, such as the one we propose in this paper, offer meaningful benefits over standard Markov Chain Monte Carlo (MCMC) approaches even when used as a non-sequential batch learning algorithm. Among other benefits, SMC algorithms produce estimates of the model evidence and Bayes factors, as well as provide an alternative inferential framework that is computationally more robust than batch data MCMC methods.

The paper is structured as follows. In Section 2, we lay out the comprehensive modelling framework for benefit risk analysis of different drug treatments, including modelling the data generation process and the MCDA approach. We review various model alternatives based on latent variable modelling, including exploratory factor analysis as well as structural equation models, in order to hone into specific aspects of the data generation process that is of interest. Section 3 introduces the framework for assessing different models and choosing the most suitable for the data at hand. In Section 4, we introduce the sequential inference paradigm. In Section 5, we demonstrate the benefits of the proposed methodology by applying it to a real world clinical case study for diabetes treatments. We conclude with a discussion of limitations and future work. All code associated with this paper is available at the accompanying github repository named 'bayesways/mcda'[1].

4.2 Factor Models for Multi-criteria Decision Analysis

4.2.1 Multi-criteria Decision Analysis Score and Data

We define MCDA amongst R treatments, based on data collected on P criteria for a set of N subjects in the following way. The criteria of a clinical trial typically consist of the benefits of the treatment and the adverse events that are experienced by the subjects, denoted by y_{ijr} for $i = 1 \dots N$, $j = 1 \dots p$, and $r = 1 \dots R$. Commonly, the benefits are efficacy measurements represented by continuous variables, whereas adverse events refer to measurable experiences

[1] https://github.com/bayesways/mcda

of negative side effects, commonly expressed as the frequency of occurrences for each event considered. The MCDA approach adopted in this paper requires that the expected value of each outcome, with respect to the population of N individuals, $E(y_{ijr})$ is accompanied by a partial value function $U_{jr} := u_j\{E(y_{ijr})\} \to [0,1]$, that maps from a pre-specified range in the observations space onto an integrated benefit-risk scale from 0 to 1. These functions, typically common across subjects and treatment groups, map the range of outcomes to a subjective continuous measure of utility or value, with 0 representing the worst case scenario, and 1 representing the best case scenario. In this paper, we work with simple linear mappings, however other types of mappings can be used as well in our approach. The MCDA score also requires multiplying each U_{jr} with a weight w_j that reflects the relative importance or preference of a full swing from worst to best of the j-th criterion relative to the rest. Such weights are elicited by expert clinicians tasked with evaluating a drug treatment, or by individual patients who want to compare the possible treatments available to them according to their personal benefit risk preferences. To keep notation clean and simple we assume these weights are common across all subjects in this work. However, it is easy to accommodate individual-specific weights $\{w_{ij}\}_{i,j}$ as long as $\sum_j w_{ij} = 1$ for all i. Without loss of generality we proceed with common weights and define population based MCDA score for the r-th treatment as

$$M_r := \sum_j w_j U_{jr} = \sum_j w_j \cdot u_{jr}\{E(y_{ijr})\}. \tag{4.1}$$

4.2.2 Factor Analysis for Mixed Type Data

As the MCDA score of (4.1) for each treatment depends on the unknown values of the population means, its calculation requires estimating them from data based on an appropriate model. The data consist of p observed variables denoted by $\mathbf{y} = (y_1, \ldots, y_p)$ that can be of mixed type and potentially dependent. We consider several models in this papers aiming for a model that captures the dependencies between the data while at the same time is parsimonious and achieves good predictive performance. The models are defined in the following subsections.

4.2.2.1 Latent variable framework

One approach that can accommodate continuous or categorical variables is a factor analysis model. According to this approach, to model a set of observations for p items $\mathbf{y} = (y_1, \ldots, y_p)$ we introduce k continuous latent factors, denoted by $\mathbf{z} = (z_1, \ldots, z_k)$. The associations between the observed items can be explained through the latent factors and their loading structure Λ as follows

$$\mathbf{y}_i^* = \alpha + \Lambda \mathbf{z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{4.2}$$

where $\mathbf{y}^* = (y_1^*, \ldots, y_p^*)$ are auxiliary variables that help us express a general framework. When the j-th item y_j is a continuous variable, we assume that the underlying variable is directly observed $y_j = y_j^*$. In the context of clinical trials, the amount of glucose detected in the blood stream could be such an item measured on the continuous scale. When the j-th item is a categorical variable, we can model it as a manifestation of the underlying latent variable y_j^* as follows. For a binary item we model it as $y_j = \mathcal{I}(y_j^* > 0)$, where $\mathcal{I}(\cdot)$ is the indicator function. For an ordinal item with m_j categories, $y_j = a$ if $\tau_{a-1}^{(j)} < y_j^* \leq \tau_a^{(j)}, a = 1, \ldots, m_j$ where $\tau_0^{(j)} = \infty, \tau_1^{(j)} < \tau_2^{(j)} < \ldots < t_{m_j-1}^{(j)}, \tau_{m_j}^{(j)} = +\infty$. Examples of such categorical variables in the context of clinical trials are typically related to adverse events. For instance experiencing allergic reactions could be a binary item, whereas the experiencing pain on scale of 1 to 5 would be an ordinal variable.

We can then interpret equation (4.2) in sufficient generality that encompasses more than one type of data, as is often the case in clinical trials. The parameters of equation (4.2) are understood as follows: α is a $p \times 1$ vector of intercepts and expressed the population mean for each item, Λ is the $p \times k$ matrix of factor loadings and n is the sample size. The factors $\mathbf{z}_i$ usually are assumed to be normally distributed, $\mathbf{z}_i \sim N_k(0, \Phi)$ though other distributions could be used. The covariance matrix Φ can either be assumed unstructured or defined by a parametric model that relates latent variables with each other and observed covariates. The $\boldsymbol{\epsilon}_i$s are error terms assumed to be independent from each other and from the $\mathbf{z}_i$s.

Specific choices for the distribution of the error term $\boldsymbol{\epsilon}$ lead to the following well known models:

$$\boldsymbol{\epsilon} \sim \begin{cases} N(0_p, \Psi), \quad \Psi = \text{diag}(\psi_1^2, \ldots, \psi_p^2), & \text{if } \mathbf{y}_i \text{ is continuous} \\ N(0_p, \Psi), \quad \Psi = I_p, & \text{if } \mathbf{y}_i \text{ is binary and the probit model is adopted} \\ \prod_{j=1}^{J} \text{Logistic}(0, \pi^2/3), & \text{if } \mathbf{y}_i \text{ is binary and the logit model is adopted,} \end{cases} \tag{4.3}$$

where 0_p is a p-dimensional vector of zeros and I_p denotes the identity matrix of dimension p.

When we know in advance which factors contribute to what items, the model is considered to be confirmatory factor analysis (CFA). Such knowledge is typically expressed by constraining certain elements of the loading matrix Λ to zero. For example, a confirmatory model used in this paper sets several element of Λ to zero in a way so that the first factor loads only onto the outcomes that measure benefits, and the second factor loads only onto the outcomes that measure risks. When the loading structure is unknown the loading matrix is Λ unconstrained and the model is referred to as exploratory factor analysis (EFA). EFA has more free parameters hence it is often necessary to place restrictions to ensure identifiability, such as restricting Λ to be upper triangular or setting $\Phi = I_k$. Both approaches may have different numbers of factors with a limiting case being the saturated model, under which

$$\mathbf{y}_i^* \sim N(\alpha, \Sigma), \tag{4.4}$$

where Σ is a full covariance matrix, except for the constraint of having ones in its diagonal entries corresponding to binary variables for identifiability reasons.

4.2.2.2 Bayesian structural equation modelling

The Bayesian formulation of factor models, introduced by Muthén and Asparouhov (2012) and extended by Vamvourellis et al. (2021), provides a alternative model that may be viewed as being between exploratory and confirmatory factor analysis or between separate effects and pooled models. This is achieved by replacing any exact zero restrictions of parameters with approximate zero ones, using informative priors that place a large amount of probability mass

around zero; these models are henceforth denoted as approximate zero. The general equation framework is expressed with the help of item-individual random effects The model in (4.2) is generalised with the addition of an item-individual specific random effect u_{ij} as follows

$$\mathbf{y}_i^* = \alpha + \Lambda \mathbf{z}_i + \mathbf{u}_i + \mathbf{e}_i, \tag{4.5}$$

where $\boldsymbol{\epsilon}_i$ in (4.2) is split into $\mathbf{u}_i$, a p-dimensional vector of random effects with a non-diagonal covariance matrix Ω, and $\mathbf{e}_i$, an error term with a diagonal covariance matrix Ψ^*. The $\mathbf{u}_i$ terms aim to capture associations amongst the items beyond those explained by the vector of latent variables $\mathbf{z}_i$. Those associations can be due to question wording, method effect, etc. In the case of continuous data it is possible to marginalise over the latent variables to get the following expression

$$\mathbf{y}_i^* \sim N(\alpha, \Lambda\Phi\Lambda^T + \Omega + \Psi^*), \quad i = 1, \ldots, n, \tag{4.6}$$

where $\mathbf{u}_i \sim N(0, \Omega)$, $\mathbf{z}_i \sim N(0, \Phi)$ and $\mathbf{e}_i \sim N(0, \Psi^*)$.

Overall, it becomes clear that several models can be used to carry out MCDA thus providing several options. On the other hand, this introduces the challenge of choosing an appropriate model amongst the competing models. Our suggested approach is presented in the next section where the aim is to strike a good balance between goodness of fit and out-of-sample predictive performance in the MCDA context of correlated mixed-type data.

4.2.2.3 Multiple group models

The models presented so far refer to a single treatment group. When data from more than one group are available, one can proceed with a separate model for each group or a pooled model aiming for more parsimony. For example in a classical clinical trial there are two arms, control and treatment. More generally a clinical trial can contain R arms where subjects are allocated to one of the R groups. We can denote such data by $\{\mathbf{y}_i^{(r)}\}_{i=1}^{n_r}$ where the r-th group contains n_r subjects. In this case, rather than modelling each group completely separately, it may be beneficial to pool together certain parameters of the model, such as the covariance structures

(Ψ, Ω, Ψ^*). At the same time the intercept parameters α and the loadings Λ can be group specific. This way expression (4.6) becomes

$$\mathbf{y}_i^{*(r)} \sim N(\alpha_r, \Lambda_r \Phi \Lambda_r^T + \Omega + \Psi^*), \quad i = 1, \ldots, n, \tag{4.7}$$

where now $r = 1, \ldots, R$ is the index of the clinical arm. Note that there are now R intercept parameters $\{\alpha_r\}_{r=1}^R$ and loading matrices $\{\Lambda_r\}_{r=1}^R$ to be estimated. However because we pooled the rest of the parameters there is still only one of each (Ψ, Ω, Ψ^*) to be estimated which is common across all R groups.

4.2.2.4 Priors

We use priors according to the recommendations in (Vamvourellis et al., 2021) and references within. The cross loadings in the Λ matrix are assigned Normal distributions with zero mean and a variance of 0.01 as in Muthén and Asparouhov (2012). For the rest of the loadings we use $\Lambda_{ij} \sim N(0, 1)$ when y_j is continuous and $\Lambda_{ij} \sim N(0, 4)$ when y_j binary. The prior given to the idiosyncratic variance, for the continuous variables, is

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

where S_y is the sample covariance matrix of the continuous observations y and c_0 is a constant that we choose so as to avoid Heywood issues, and bound the samples away from 0. The idiosyncratic variance for the binary variables are fixed to 1, since they are not identifiable. We set Φ as a correlation matrix and remove the constraints imposed on the corresponding elements of Λ (equal to one) for identifying the scale of the latent variables. Also rather than constraining these Λ entries to be positive, a restriction required to ensure identifiability, we assign Normal priors to all of them and apply post processing on the MCMC output based on their sign. The LKJ prior, introduced in Lewandowski et al. (2009), was used for the correlation matrix Φ. For the Ω matrix, we assign the Inverse Wishart distribution with identity scale matrix and $p + 6$ degrees of freedom to reflect prior beliefs of near zero residual covariances (Muthén and

Asparouhov, 2012). The scale can be thought of as controlling the magnitude of model flexibility the researcher is willing to allow for capturing the effects of external factors on measurement. Hence, it is important when setting this prior to ensure that the $\mathbf{u}_i$s are of lower magnitude than the $\mathbf{e}_i$s. Finally, we set large variance normal priors for the α parameters, $\alpha \sim N(0, 10^2)$.

4.3 Model Assessment

In this section we propose a model assessment framework suitable for potentially correlated mixed type data and provide guidance on how to use it in a real world case study. In short, our approach looks at goodness of fit and predictive performance indices separately for continuous and categorical data. More specifically, we recommend complementing PPP values by monitoring the out-of-sample performance of the marginal distributions of each set of variables, continuous and binary. In other words, we model all the variables jointly but check the model fit separately for continuous and binary. In mixed data, to our knowledge it is not possible to come up with a single metric for the PPP values or a single scoring rule for both the binary/ordinal and continuous data. Hence we adopt the approach of looking at separate indices of these parts. Examining separately the fit for the categorical and the continuous items is in line with previous work by Moustaki (1996).

While PPP values are commonly used as the default metric for Bayesian factor modelling, they are not directly suitable for the approximate zero framework (Muthén and Asparouhov, 2012). The reason being that they cannot detect overfitting, the case where the gains in goodness of fit are based on finding circumstantial patterns of the data that do not replicate in new unseen data. To guard against overfitting Vamvourellis et al. (2021) proposed evaluating out-of-sample scoring rules via cross-validation. We need to evaluate the performance of the predictive distributions, which are typically not available in closed form, against the observed data in the test set. This is done using proper scoring rules that can be evaluated with posterior sample draws. While the specific scoring rules used differ depending on the type of data under consideration, the basic idea is to avoid overly flexible models that often fail to disentangle

noise from systematic patterns in the data.

4.3.1 PPP Values

Our proposed approach considers measures of goodness of fit (to assess if a model fits the data sufficiently well) and of relative fit (to compare competing models). Since the tests are specific to the type of data we need to apply them to the marginal distributions of the target variables, separately for the continuous and binary parts of the outcomes. The goodness of fit is regularly used to assess if the model is supported by the data. In the Bayesian setting, the common metric used is posterior predictive p-values (PPP) which is an absolute measure of in-sample performance. A model is considered to not fit the data if it achieves a PPP value that is close to zero, though it does not have the same distribution properties as traditional p-values. What PPP values cannot do is to compare models beyond the question of adequate fit. For that, Vamvourellis et al. (2021) suggested using out-of-sample predictive performance measured by scoring rules and cross-validation. Since factor models have fewer variables than the unconstrained saturated versions, if a factor model has adequate fit, it is expected to perform better out-of-sample due to parsimony. Generally speaking, the higher the number of parameters the better a model will perform in-sample, but that is not necessarily true out-of-sample where performance grows as a function of parameters until it starts declining. The focus of out-of-sample check is on a model's ability to predict new data that was not used for estimating the model parameters. In practice, it is recommended to use cross-validation instead of randomly sampling a single test set, to limit the effect of an unfortunate split. These cross-validated scoring rules are a useful compliment to goodness of fit measures such as the Posterior Predictive P-values (PPP) which are commonly used, although researchers have questioned their suitability for Bayesian factor models (Stromeyer et al., 2015; Hoijtink and van de Schoot, 2018b). One of the challenges here is that PPP performance can be sensitive to the choice of prior distributions (MacCallum et al., 2012; Van Erp et al., 2018; Liang, 2020). A more robust approach is to additionally monitor the predictive performance of the models, even when prediction is not strictly the main aim. The idea is to avoid models that achieve a

high PPP value but fail to learn patterns that generalise to unseen data. An alternative and more formal Bayesian approach to model choice uses the marginal likelihood conditional on the model which we do not explore further in this work.

The object of interest for model assessment is the predictive distribution. In the Bayesian setting the object that is the posterior predictive distribution

$$f(\boldsymbol{Y}^{te}|\boldsymbol{Y}^{tr}) = \int f(\boldsymbol{Y}^{te}|\theta)\pi(\theta|\boldsymbol{Y}^{tr})d\theta. \tag{4.8}$$

where $\mathbf{y}^{tr}$ denotes the data used for learning the model parameters and $\mathbf{y}'$ can be $\mathbf{y}^{tr}$ (in-sample prediction) or new data that was not used for inference (out-of-sample prediction). In the frequentist case, one option for such a predictive distribution is $f(\boldsymbol{Y}^{te}|\hat{\theta}^{tr})$, where $f(\cdot)$ denotes the likelihood function and $\hat{\theta}^{tr}$ is the maximum likelihood estimate obtained from $\boldsymbol{Y}^{tr}$.

We first describe the general procedure of calculating the PPP value and then get into the specific choices of functions. Given a suitable MCMC algorithm, the PPP value is then calculated 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 $\widetilde{\mathbf{Y}}$ having the same size as $\boldsymbol{Y}$ using the current value $\boldsymbol{\theta}_m$ in Equation (4.2) or (4.6) (depending which model is under consideration).

   (c) Calculate $D(\widetilde{\mathbf{Y}}, \boldsymbol{\theta}_m)$ and $d_m = \mathcal{I}\big[D(\boldsymbol{Y}, \boldsymbol{\theta}_m) < D(\widetilde{\mathbf{Y}}, \boldsymbol{\theta}_m)\big]$, where $\mathcal{I}\big[\cdot\big]$ is an indicator function.

2. Return PPP$= \frac{1}{M}\sum_{m=1}^{M} d_m$.

The PPP value is defined based on a discrepancy function $D(\boldsymbol{Y}, \boldsymbol{\theta})$. A standard choice of $D(\cdot)$ for continuous data is the likelihood ratio test (LRT) statistic between the restricted and the unrestricted model (see e.g. Scheines, Hoijtink, and Boomsma, 1999b). For binary data, a

common choice of discrepancy measure, see for example Sinharay (2005c), is the G^2 statistic given by

$$D(\boldsymbol{Y}, \boldsymbol{\theta}) = \sum_{r=1}^{R} O_r \log\left(\frac{O_r}{n\pi_r(\boldsymbol{\theta})}\right). \tag{4.9}$$

where the model is reformulated in terms of response patterns and their observed frequencies in a sample size of n. For binary items there are $R = 2^p$ possible response patterns, denoted by $\mathbf{y}_r$, with corresponding observed frequencies denoted by O_r where $r = 1, \ldots, R$. The probability of occurrence of a response pattern, based on a logistic model with a parameter vector $\boldsymbol{\theta}$, can be calculated from the following equation using the fact that the p variables are conditionally independent given $\mathbf{z}$ and $\mathbf{u}$

$$\pi_r(\theta) = \int \prod_{j=1}^{p} \text{Bernoulli}\left\{[\mathbf{y}_r]_j | \sigma([\boldsymbol{\eta}]_j)\right\} f(\mathbf{z}) f(\mathbf{u}) d\mathbf{z} d\mathbf{u}, \tag{4.10}$$

where $\sigma(\eta_{rj})$ is defined in Section 2.2.2, and $\mathbf{z}$ and $\mathbf{u}$ are the latent components in the model considered. The integral in (4.10) can be computed using Monte Carlo. For more details we refer the interested reader to (Vamvourellis et al., 2021).

4.3.2 Scoring Rules

While PPP is computed using the entire dataset $\boldsymbol{Y}$, the out-of-sample scoring rules are computed on a held out test dataset $\boldsymbol{Y}^{te}$ which is completely separate from the dataset $\boldsymbol{Y}^{tr}$ used to learn the parameters. Once we split the data in training and testing sets, we perform inference to learn the model parameters using the training set, and compare the posterior predictive distribution against the held out testing set. The comparison is done using scoring rules.

In the case of continuous data we use the variogram score which is defined as:

$$VS(\mathbf{y}_i, \widetilde{\mathbf{Y}}) = \sum_{j=1}^{p} \sum_{k=1}^{p} w_{j,k} \left(|y_{ij} - y_{ik}|^P - \frac{1}{m} \sum_{m=1}^{M} |\tilde{y}_{mj} - \tilde{y}_{mk}|^P \right)^2 \tag{4.11}$$

where, in (4.11), the j, k are just indices to consider all pairs of each data point $\mathbf{y}_i$ of dimension p, the $w_{j,k}$s are weights and P is the order of the variogram. We follow common practice by

setting all weights to one, and P at its default value of 0.5. Hence, for each cross validation split between training ($\boldsymbol{Y}^{tr}$) and test data ($\boldsymbol{Y}^{te}$), the variogram score can be computed by obtaining samples from the posterior based on $\boldsymbol{Y}^{tr}$, and using them to draw samples from the posterior-predictive distribution for $\boldsymbol{Y}^{te}$. The latter can then be inserted in (4.11), together with the test data, to calculate the score corresponding to this train-test split for the model considered.

In the case of binary data we use the the log scoring rule for a set of observed frequencies in the test data $\mathbf{O}^{te} = (O_1^{te}, \ldots, O_R^{te})$ based on probabilities $\pi^{tr} = (\pi_1(\boldsymbol{\theta})^{tr}, \ldots, \pi_R(\boldsymbol{\theta})^{tr})$, obtained based on the posterior from the training data, as

$$LS(\mathbf{O}^{te}, \pi^{tr}) = -\log f(\mathbf{O}^{te}|\pi^{tr}) = -\log\left[c\prod_{r=1}^{R}\left[\pi_r(\boldsymbol{\theta})^{tr}\right]^{O_r^{te}}\right] = -\sum_{r=1}^{R} O_r^{te}\log\pi_r(\boldsymbol{\theta})^{tr} + c, \quad (4.12)$$

where c represents a constant. For more details we refer the interested reader to (Vamvourellis et al., 2021).

4.4 Sequential Monte Carlo

Having made our model choice, as explained in section 4.3, we proceed to construct an efficient sequential inference scheme that provides the sequence of parameter posterior distributions. In this section we present our developed sequential inference paradigm that allows us to recursively update model and parameter estimates each time new data become available. This framework can be used towards a sequential clinical trial design which can offer improvements over standard designs on moral and financial grounds. It permits stopping the treatment as soon as the research requirements are satisfied, reducing unnecessary further exposure to undesirable treatments. More broadly, it allows us to allocate dynamically subjects to treatment groups based on parameter inference. Without a sequential inference algorithm, achieving such goals can only be done by re-running the MCMC algorithms for all sequential batches.

4.4.1 IBIS Algorithm for Benefit Risk Analysis

We use the Iterated Batch Importance Sampling algorithm (IBIS) with the Laplace approximation proposal for the latent variables. At a high level, IBIS works by propagating forward in time a set of parameter samples, also known as particles, each weighted by the likelihood function evaluated at its parameter values. For the standard IBIS approach we need to have access to the likelihood function $f(y|\theta)$, as in formula (4.6). However, such forms are not available for the logit model, or any mixed type data models that include the logit. Instead we can easily evaluate the augmented likelihood $f(y|\theta, z)$. In order to use the augmented likelihood we need to augment the set of particles in the IBIS algorithm to include the latent variables as well. For each data point $\mathbf{y}_i$ we draw latent variable particles $\{\mathbf{z}_i^m\}_{m=1}^{N_\theta}$ from a proposal distribution $q(.)$ and compute the weights according to $u^m = f(\mathbf{y}_i|\theta^m, \mathbf{z}_i^m)$. The simplest and most standard choice for the proposal distribution is the prior $q(z) = \pi(z)$. However, the efficiency of the scheme can be improved if the proposal is the posterior $q(z) = \pi(z|y, \theta^m)$. Even though the exact posterior is not available we can use the Laplace or the Variational Bayes (Blei et al., 2017) approximations. In this paper we proceed with the former which we denote as $p^L(\mathbf{z}_i|\mathbf{y}_i, \theta^m)$.

For ease of illustration we develop the sequential algorithm below for one of the models we consider in this work. We choose the exact zero model, which performs quite well in the case study we present in the following section 4.5. However, the methodology we propose extends to the rest of the models presented in this paper.Without loss of generality let us assume that the first c variables are continuous and the next $p-c$ are binary. The model is defined by equations (4.2) or (4.3) in the case of exact zero models or (4.5) and (4.6) in the case of approximate zero models. In the case of the saturated model the definition comes from (4.4).

The particles of the parameter vector θ contain $(\alpha, \Lambda, \Psi, \Phi)$ and the latent variables are denoted by $\mathbf{z}_i$. Note than in the case of EZ models, the $\boldsymbol{\eta}_i$s is a deterministic function of z, α and Λ, here $\eta(z, \theta) = \alpha + \Lambda z$; under the AZ formulation they would have also been separate random variables. All the operations involving the particle index m must be understood as operations performed for all $m \in 1 : N_\theta$, where N_θ is the total number of θ-particles. The incremental

Algorithm 3 IBIS-Laplace for Benefit Risk Analysis

Sample θ^m from $p(\theta)$ and store and set $\omega^m = 1$. All operations are assumed to be repeated for all $m \in 1 : N_\theta$.

Then for point $i = 1 : N$ do:

1: Sample $\mathbf{z}_i^m \sim p^L(\mathbf{z}_i|\mathbf{y}_i, \theta^m)$ using the Laplace approximation

2: Compute and store $\boldsymbol{\eta}_i^m$ along with θ^m and $\mathbf{z}_i^m$ as $\eta(\mathbf{z}_i, \theta) = \alpha + \mathbf{z}_i\Lambda'$ so that every time we refer to $(\theta^m, \mathbf{z}_i^m)$ we also have access to the associated $\boldsymbol{\eta}_i^m$

3: We also compute and store the MCDA population score associated with particle θ^m for each treatment r as follows $s_r^m = \sum_j w_j \cdot u_j(\alpha_{jr})$, using formula (4.1)

4: Update $\mathbf{z}_{1:i}^m = \mathbf{z}_i^m$

5: Compute the incremental weights and their weighted average

$$u_i(\theta^m, \mathbf{z}_i^m) = \frac{f(\mathbf{y}_i|\mathbf{y}_{1:i-1}, \theta^m, \mathbf{z}_i^m)\pi(\mathbf{z}_i^m|\theta^m)}{p^L(\mathbf{z}_i^m|\mathbf{y}_i, \theta^m)}, \quad L_i = \frac{1}{\sum_{m=1}^{N_\theta} \omega^m} \times \sum_{m=1}^{N_\theta} \omega^m u_i(\theta^m, \mathbf{z}_i^m),$$

6: Update the importance weights

$$\omega^m = \omega^m u_i(\theta^m, \mathbf{z}_i^m)$$

7: **if** ESS$(\omega) < \gamma$ **then**

8: **procedure** RESAMPLE$(\theta, \mathbf{z}_{1:i}, \omega)$

9: **return** θ, $\mathbf{z}_i$

10: **procedure** JITTER$(\theta^m, \mathbf{z}_{1:i}^m, \mathbf{y}_{1:i})$ using an HMC algorithm with

11: **return** $\tilde{\theta}^m, \tilde{\mathbf{z}}_{1:i}^m$

12: $(\theta^m, \mathbf{z}_{1:i}^m, \omega^m) = (\tilde{\theta}^m, \tilde{\mathbf{z}}_{1:i}^m, 1)$

weight in the Algorithm 3 can be computed as follows

$$u_i(\theta^m, \mathbf{z}_i^m) = \frac{f(\mathbf{y}_i|\mathbf{y}_{1:i-1}, \theta^m, \mathbf{z}_i^m)\pi(\mathbf{z}_i^m|\theta^m)}{p^L(\mathbf{z}_i^m|\mathbf{y}_i, \theta^m)} = \frac{f(\mathbf{y}_i|\theta^m, \mathbf{z}_i^m)\pi(\mathbf{z}_i^m|\theta^m)}{p^L(\mathbf{z}_i^m|\mathbf{y}_i, \theta^m)} \tag{4.13}$$

where

$$f(\mathbf{y}_i|\alpha^m, \Lambda^m, \Psi^m, \mathbf{z}_i^m) = \prod_{j=1}^{c} N\left\{\mathbf{y}_{i,1:c}|(\alpha^m + \mathbf{z}_i^m\Lambda^m)_{ij}, \psi_j^2\right\} \cdot \prod_{j=p-c+1}^{p} \text{Bernoulli}\left\{y_{ij}|\ \sigma\left[(\alpha^m + \mathbf{z}_i^m\Lambda^m)_{ij}\right]\right\} \tag{4.14}$$

$$\pi(\mathbf{z}_i^m|\Phi^m) = N(\mathbf{z}_i|0, \Phi^m) \quad \text{(prior for } \mathbf{z}_i\text{)} \tag{4.15}$$

$$p^L(\mathbf{z}_i^m|\mathbf{y}_i, \theta^m) = N(\mathbf{z}_i|\mu^L, \Sigma^L) \quad \text{(Laplace approximation)} \tag{4.16}$$

with $N(x|A, B)$ denoting the probability density function of a random variable x based on the $N(A, B)$ and, similarly, Bernoulli$(y_{ij}|P)$ denotes the probability mass function of the Bernoulli(P) distribution. The last formula is given by the Laplace approximation of the posterior $\pi(\mathbf{z}_i^m|\mathbf{y}_i, \theta^m)$. Further simplifications may be possible depending on the model. Appendix B.2 contains details for EZ models with separate independent factors for the continuous and binary variables.

The jittering procedure uses the MCMC developed for the batch MCMC described in section 4.2. We now derive the Laplace approximation for this model for demonstration purposes. Alternatively, one can find numerical approximations to the Laplace method parameters.

4.5 Rosiglitazone Case Study

We now demonstrate our proposed framework by applying it to the study of rosiglitazone treatment for type 2 diabetes. We first give an overview of the clinical trial setup and the data that was collected as a result of it, and lay out the MCDA framework and the parameter choices within. In section 4.5.2 we describe in detail the range of models we consider and demonstrate

the proposed model choice methodology. In section 4.5.3 we compute the final MCDA scores based on the estimated parameters of the chosen model. Finally, in section 4.5.4, we present our proposed sequential framework. Specifically, since we do not have a dedicated sequential dataset, we create one synthetically to highlight the benefits of the sequential clinical trial design.

4.5.1 Data and MCDA Setup

We analyse data from a clinical trial of three different treatments that were administered over a 12 week period and the difference in health outcomes was compared before and after the treatment. The data consists of 449 diabetic subjects that received one of three treatments under examination: metformin (MET), rosiglitazone (RSG) and a combination of the two (AVM), marketed under the commercial name 'Avandia'. The sample sizes for each group were as follows, 146 subjects were given MET, 153 were given RSG and 150 were given AVM.

We employed the MCDA methodology to arrive at a comprehensive score for each treatment reflecting their corresponding benefit risk profiles. The targeted efficacy outcome was a reduction of the haemoglobin and glucose levels detected in the blood stream of each subject at the end of the treatment, compared to the level observed in the pre-treatment screening. Measurements for both haemoglobin and glucose difference from the baseline were taken on a continuous scale. The treatments under consideration are known to have relatively mild side effects, most common of which are nausea and vomiting. For our case study we consulted with researchers who studied the drugs before and landed on 4 adverse events of note: diarrhoea, nausea, vomiting, and dyspepsia, encoded as binary outcomes that were equal to 0 unless a patient experienced the event at least once during the trial. Overall, our data consisted of 8 columns, one for the anonymised subject id, one for treatment received, two for the efficacy continuous outcomes and four for the binary adverse events.

In this study, we adopted all the MCDA parameters, such as the partial value functions and the specified ranges of the variables in questions in consultation with researchers who have

conducted similar studies of rosiglitazone in the past using MCDA. The ranges and weights for each outcome are presented in Table 4.1.

Name	Type	Outcome Range	MCDA Weight
haemoglobin	continuous	[-6, 3]	0.592
glucose	continuous	[-15, 7.5]	0.118
prob(diarrhoea)	binary	[0.10, 0.35]	0.089
prob(nausea)	binary	[0.10, 0.25]	0.178
prob(vomiting)	binary	[0.10, 0.20]	0.018
prob(dyspepsia)	binary	[0.10, 0.25]	0.005

Table 4.1: Outcomes of interest and MCDA parameters.

Note that the continuous variables, the first two items, indicate the difference in measured haemoglobin and glucose levels respectively, at the end of the study versus at the beginning. The binary adverse events, last 4 items, measure the probability of experiencing the even in question at least once during the course of the treatment.

4.5.2 Model Choice

Given the data and the MCDA scores, it is important to design an appropriate model in order to extract the relevant population parameters of interest. We consider a range of different models and utilise the proposed model assessment framework to choose between them. Focus is given on CFA models with two factors and structure such that the first factor loads onto the efficacy variables and the second loads onto the risk variables as shown in Table 4.2.

The models considered are the full saturated model, with full covariance matrix (SAT), and the independence model where the covariance matrix S is fixed to be diagonal (IND). Then we considered four factor models where one factor loads to the first two items, being the efficacy variables which are continuous variable, and the second factor loads to the rest 4 items, the adverse events which are binary variables. The first model is the EZ model with independent factors (EZ1), and the second is the same model with correlated factors (EZ2). The approximate zero model includes a model with correlated errors for the binary data, without cross loadings (AZ1), and with cross loadings (AZ2). For all the above models, we fit also the

$\Lambda_{:1}$	$\Lambda_{:2}$
1	0
x	0
0	x
0	x
0	x
0	x

Table 4.2: Hypothesised factor loading structure: the first factor z_1 loads onto the efficacy variables, the first two items, and the second factor z_2 loads onto the risk variables, last four items.

version (denoted by '-p') where the covariance structured is assumed to be common amongst the three groups. Among these option, the model that fits the data the best has to be the one the highest predictive performance for out-of-sample data. The results are summarised in Table 4.3. Before proceeding one should also check that the model selected achieves a reasonable PPP value, typically values greater than 0.1 are not problematic and values around 0.5 indicate excellent fit. The PPP values for the top for the top four models are shown in Table 4.4.

The collective results, presented in Table 4.3, lead to the following conclusions. First, pooling the covariance parameters results in higher predictive performance for the binary data and lower performance for the continuous data. We can verify that by comparing models in pairs while recalling that lower scoring rules indicate a better predictive performance. For example we see that SAT underperforms SAT-p in both continuous and binary scoring rules by about 1 and 2 units respectively, EZ1 underperforms EZ1-p in both continuous and binary scoring rules by about 5 and 9 units respectively, and finally IND underperforms IND-p in both continuous and binary scoring rules by about 6 and 1 unit respectively. When the covariance estimates of each of the three groups are reasonably close, as seems to be the case here, pooling benefits predictive accuracy; whereas in the opposite case the pooled model will underperform. Second, the results suggest that the saturated model overfits the data leading to lower out-of-sample performance than the more parsimonious models. To verify that observe that the best performing saturated model amongst all the versions (SAT, SAT-p, IND, IND-p) is SAT-p, and is still underperfoming the worst performing factor model, EZ1, by about 1 unit in the continuous part and about 6 units in the binary part. Third, we suspect that the benefits and risks are not substantially

correlated, since EZ1-p fits better than EZ2-p. Recall that EZ1 is the same model as EZ2 except that the factors are assumed independent. Furthermore, we see that there is non trivial correlation within the benefit items, and within the risk items respectively, since the IND scores worse than SAT while at the same time IND-p scores worse than SAT-p. Finally we see the power of model parsimony as the most parsimonious of the models, EZ1-p, turns out to be the best in terms of predictive performance of all models. Before moving on with the model of choice, EZ1-p, we also confirm that it achieves a satisfactory fit. In Table 4.4 we compare the PPP values for all the best models so far, and confirm that all of them, including the top pick, EZ1-p, demonstrate excellent fit with PPP values near 0.5 for both continuous and binary.

In any case, the importance of model choice and assessment becomes clear. While all the models presented hear are plausible, there could be substantial differences between them as far as predictive performance is concerned. Since we will be basing our final analysis on the estimated parameters of the model, choosing a model with good predictive performance is crucial to protect the validity of the final results.

Model	Continuous-VS	Binary-LS
SAT	349.93	573.10
SAT-p	348.75	570.91
EZ1	347.88	564.36
EZ1-p	342.03	555.36
EZ2-p	345.06	558.53
AZ1-p	346.41	557.56
AZ2-p	343.44	558.51
IND	388.44	576.37
IND-p	382.90	575.07

Table 4.3: Summary of out of sample predictive performance for all candidate models using scoring rules. Sum of variogram and log scores of 3-fold cross validation for continuous and binary data respectively.

In Table 4.5 we also present the parameter estimates for EZ1-p, the chosen model, on the basis of goodness of fit and out-of-sample predictive performance.

Model	Continuous-PPP	Binary-PPP
EZ1-p	0.49	0.36
EZ2-p	0.49	0.37
AZ1-p	0.42	0.35
AZ2-p	0.44	0.39

Table 4.4: Model Fit Metrics for the best performing models.

Parameter	95% Coverage	Post. Mean	Post. Median
$\Lambda_{[2,1]}$	$[1.45, 2.15]$	1.78	1.78
$\Lambda_{[3,2]}$	$[0.05, 0.98]$	0.46	0.45
$\Lambda_{[4,2]}$	$[-1.21, 0.40]$	-0.39	-0.37
$\Lambda_{[5,2]}$	$[-1.43, 3.46]$	2.15	2.23
$\Lambda_{[6,2]}$	$[-1.67, 3.71]$	2.36	2.44
α_1^{AVM}	$[-2.49, -2.10]$	-2.30	-2.30
α_2^{AVM}	$[-4.48, -3.62]$	-4.05	-4.05
α_3^{AVM}	$[-2.42, -1.40]$	-1.87	-1.86
α_4^{AVM}	$[-2.94, -1.70]$	-2.27	-2.25
α_5^{AVM}	$[-4.08, -1.88]$	-2.83	-2.78
α_6^{AVM}	$[-7.29, -3.60]$	-5.19	-5.10
α_1^{MET}	$[-2.03, -1.64]$	-1.83	-1.83
α_2^{MET}	$[-3.41, -2.49]$	-2.95	-2.95
α_3^{MET}	$[-1.85, -0.97]$	-1.39	-1.39
α_4^{MET}	$[-3.31, -1.96]$	-2.58	-2.56
α_5^{MET}	$[-4.91, -2.38]$	-3.45	-3.38
α_6^{MET}	$[-7.23, -3.63]$	-5.17	-5.07
α_1^{RSG}	$[-1.79, -1.41]$	-1.60	-1.60
α_2^{RSG}	$[-3.25, -2.38]$	-2.81	-2.81
α_3^{RSG}	$[-3.42, -2.08]$	-2.70	-2.68
α_4^{RSG}	$[-3.15, -1.87]$	-2.44	-2.42
α_5^{RSG}	$[-5.80, -2.87]$	-4.11	-4.04
α_6^{RSG}	$[-9.13, -4.53]$	-6.55	-6.45

Table 4.5: True values, 95% coverage success rate and bias of point estimators out of 100 replications.

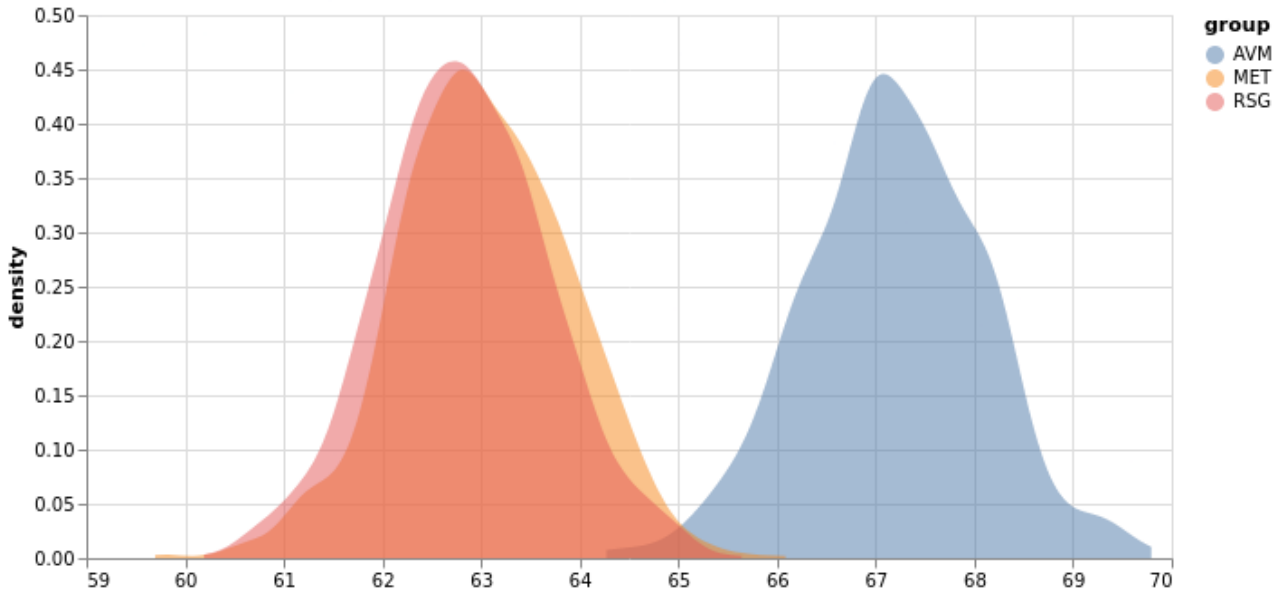

Figure 4.1: MCDA Scores at the end of the sequential run. Avandia (AVM) scores substantially higher than the other two treatments, Rosiglitazone (RSG) and Metformin (MET).

4.5.3 MCDA Scores

Our goal is to compute the final benefit risk score, including the parameter uncertainty that is propagated from the Bayesian model. Having access to samples from the posterior distributions of the parameters allow us to compute any metric of interest, including the MCDA socres. For each sample from the posterior α_r^m for treatment r we compute the implied sample s_r^m from the posterior of the MCDA score as follows $s_r^m = \sum_j w_j \cdot u_j(\alpha_{jr})$. We run the batch MCMC algorithm to draw 1000 samples from the posterior distribution of α and the implied posterior distribution of the scores s_r for $r = 1, 2, 3$ representing the three treatment groups MET, RSG and AVM respectively. The scores posterior density distributions are show in Figure 4.1. As we can see AVM has higher mean scores compared to both RSG and MET. We computed the posterior probability that $P(s_{\text{AVM}} > s_{\text{MET}}) = 0.99$ and that $P(s_{\text{AVM}} > s_{\text{RSG}}) = 0.99$.

4.5.4 Sequential Analysis

In our data the subjects were pre-allocated according to a random algorithm to one of the three treatment arms ahead of the trial. For illustration purposes we now analyse the same

data as if the allocation to the treatment arms was done sequentially. Specifically, we shuffled all subjects and then inter-weaved them cycling through the groups as shown in Table 4.6. We will demonstrate what could have been done differently if we conducted a sequential design trial, rather than a traditional pre-specified one.

New Index	Subject ID	Group
1	324	AVM
2	422	MET
3	124	RSG
4	121	AVM
5	224	MET
6	231	RSG
7	...	...

Table 4.6: Schematic sequential schedule of synthetic re-ordering of the original clinical trial data.

To demonstrate the use of the sequential paradigm, we run our analysis inputting one subject at a time. Since we didn't have access to the original order that the subjects were treated in, we randomised the order before interweaving the treatment groups so that we cycle through all three groups at a similar rate as we analyse one data point at a time. We computed the MCDA score of the parameters at each data point looking for potential early stopping points for any of the treatments. In particular we monitored the probability $P(s_{\text{AVM}} > s_{\text{MET}})$ at each data point and found that it converged to 0.99 within the first 198 patients. Of those patients, the effective sample size exposed to either AVM or MET is two thirds as per the schedule (Table 4.6). That means that we were able to conclude that AVM is better than MET using information from only 66 of 150 patients, which represents about 43% of the subject sample size. Similarly, we monitored $P(s_{\text{AVM}} > s_{\text{RSG}})$ which converged to 0.99 at within the first 300, as shown in Figure 4.2. This translates to reaching a conclusion using 66% of the sample size. Using the sequential paradigm we would have concluded the exposure at the 300-th subject, removing the need to expose subjects to the less effective treatments. This way the trials would have reached the same conclusion faster and exposing fewer subjects. Alternatively, we could have allocated more subjects to RSG and MET in order to facilitate that comparison. As we can see in Figure 4.3 we can also see a case of an inconclusive comparison between RSG and MET. Under a dynamic

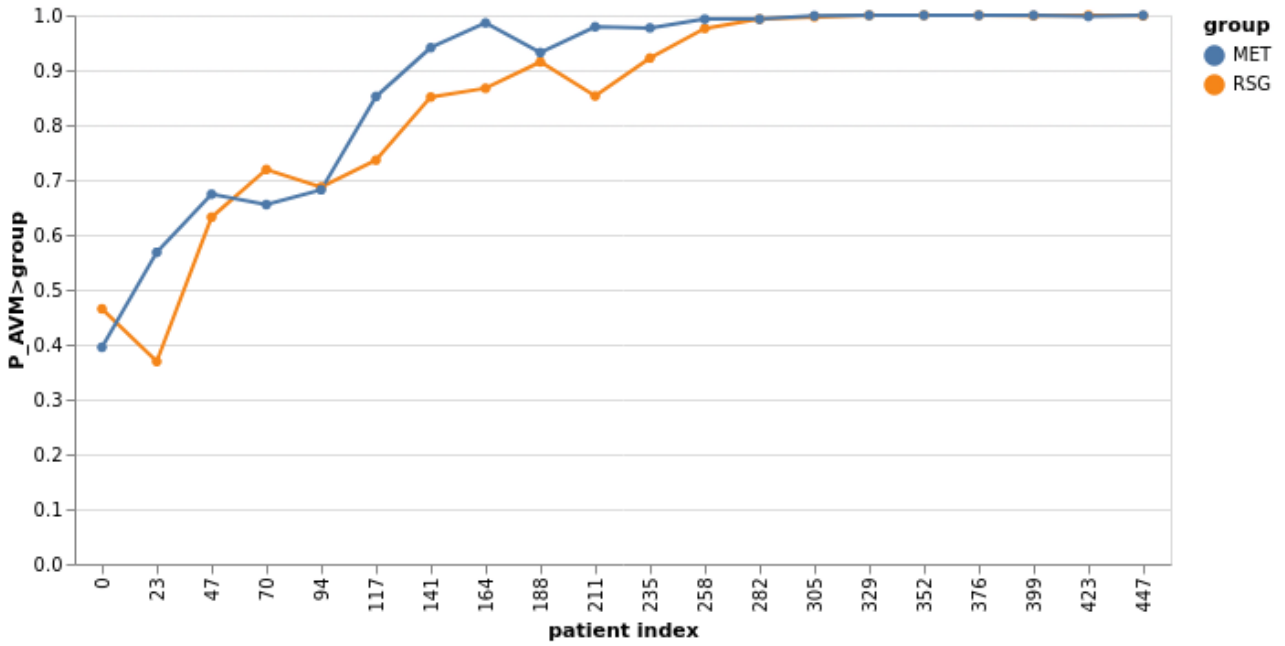

Figure 4.2: Sequentially updated probabilities of $P(s_{\text{AVM}} > s_{\text{MET}})$ and $P(s_{\text{AVM}} > s_{\text{RSG}})$. We see tha the probabilities converge to 1 within the first 300 patients. A dynamic trial could have concluded early or could have assigned the remaining patients to AVM given that it is considered a better treatment based on MCDA scores.

clinical trial it would be possible to monitor these probabilities and allocate patients away from treatment arms that results have converged early, such as the AVM-RSG and AVM-MET pair comparisons, and into groups that require more data to become conclusive, such as RSG-MET depicted in Figure 4.3.

We are also able to make a sequential chart of distribution of the MCDA population scores for each treatment. Figure 4.4 shows the posterior mean (lines) and the 95% central quantiles (shade bands) of the posterior distributions for the final MCDA population scores for each treatment, at each point in the trial. This way we can get a high level view of the evolution of the clinical trial results, and the progression of the uncertainty quantification as more data is gathered. The lines indicate that AVM (in blue) is shown to have a higher predicted score early on, within the first 100 data points. However, the fact that the uncertainty bands around the posterior mean lines of the three treatments at that point remain fairly indistinguishable, indicate that there is considerable uncertainty about which treatment is better at that point. We can also see that the AVM 95% band separates from the the other two bands after about the first 300 subjects while the other two treatments remain very close throughout. At that

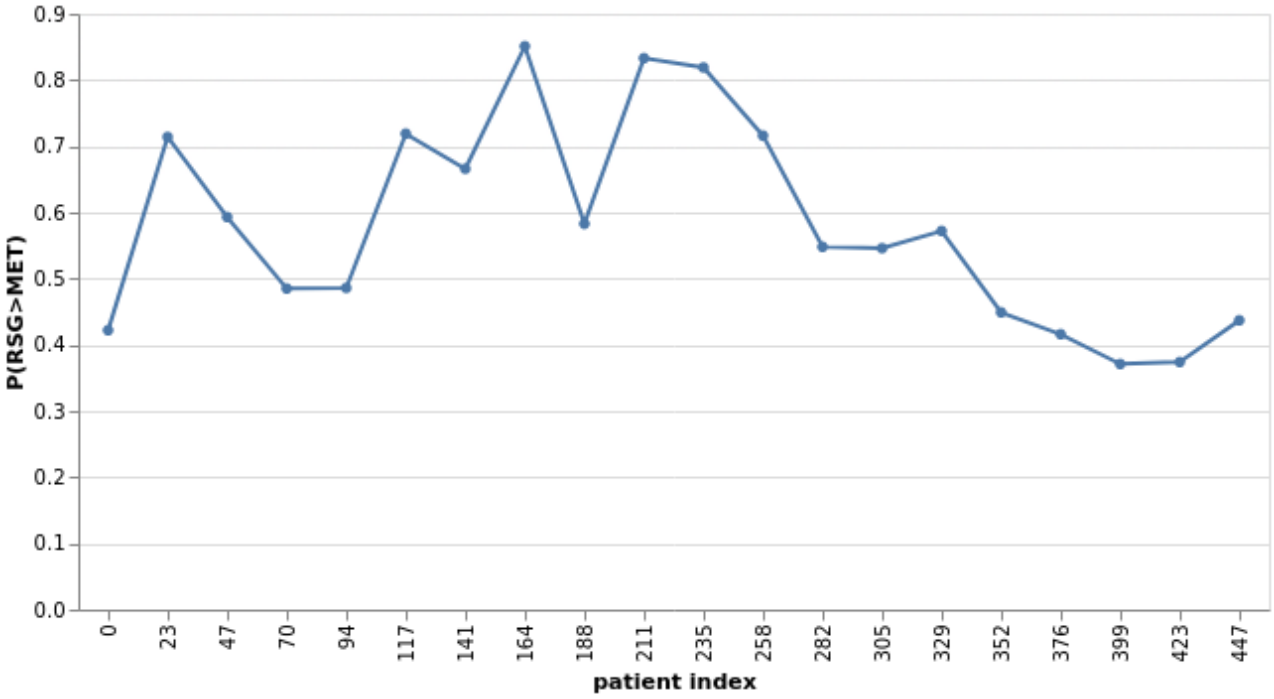

Figure 4.3: Sequentially updated probabilities of $P(s_{\text{RSG}} > s_{\text{MET}})$. We see that the comparison is inconclusive as the probability fluctuates around 0.5. Under a dynamic clinical trial it would be possible to monitor these probabilities and allocate patients away from treatment arms that results have converged early, such as the AVM-RSG and AVM-MET pair comparisons, and into groups that require more data to become conclusive, such as RSG-MET presented here.

point we are almost certain that the AVM score is higher than that of RSG and MET.

4.6 Discussion

In this paper we introduce a comprehensive benefit risk framework for the assessment of clinical treatments. To holistically assess competing treatments it is important to use a framework that encompasses all aspects of the treatment. MCDA is a general framework that can be used in association with any statistical model and any types of data. It also allows for stakeholders, clinicians, policy makers, individual patients, or others, to express their preferences through the use of weights. MCDA rests on accurate estimation of the statistical parameters of interest, which are embedded in the data generation process.

For this purpose we introduce a range of models that accommodate mixed type data and a methodology to assess the model fit. The more modelling options the more important it is

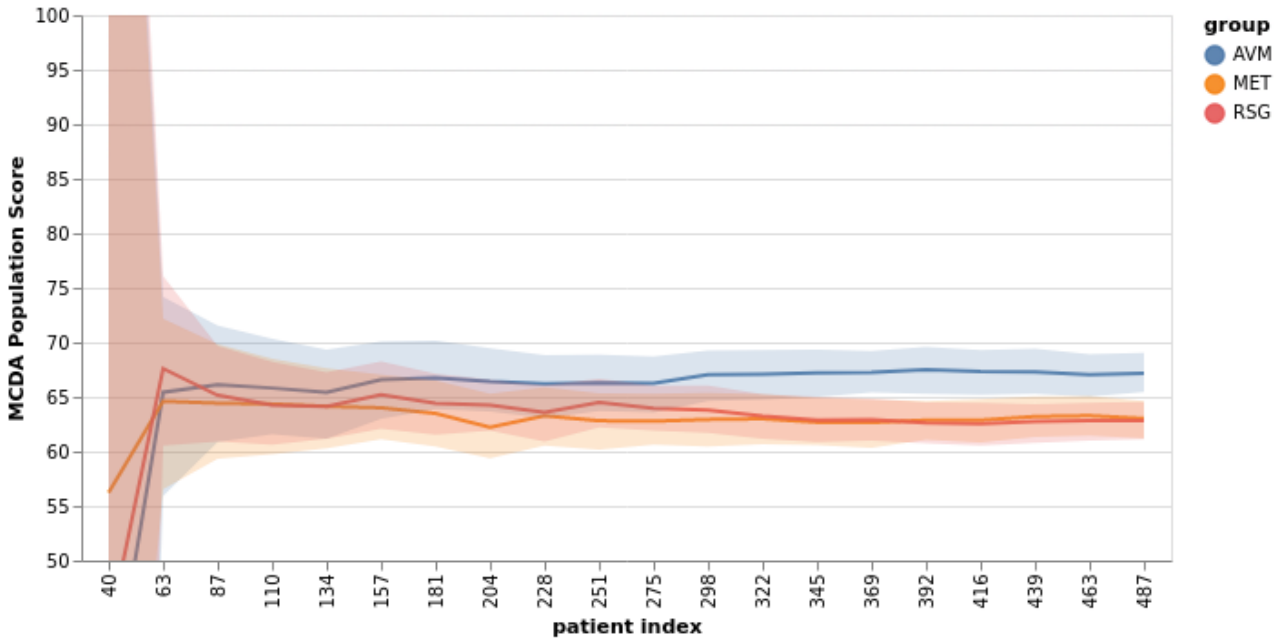

Figure 4.4: Sequentially updated posterior mean (lines) and the 95% central quantiles (shade bands) of the posterior distributions for the final MCDA population scores s_r for each treatment $r \in \{\text{AVM, MET, RSG}\}$, at each point in the trial.

to choose the model carefully. On the one hand, more flexible models are desired in ordered to capture more features of the data, such as the dependence amongst the items. On the other hand, the more flexible the model the more susceptible to overfitting. It is vital to check carefully the model fit before using the parameter estimates further. In particular, a model that overfits can produce misleading estimates which in turn can lead to false conclusions regarding the treatment. Robust model assessment frameworks are needed, especially when there are multiple competing models to choose from. The introduced model assessment framework does not examines cross dependencies between binary and continuous data when it comes to goodness of fit and predictive performance as this is not a straightforward task. Finding metrics to do so is an interesting open problem which is left for further research. In this paper we proceeded requiring good performance on both marginal parts of the data in the absence of a better alternative in line with Moustaki (1996).

In this paper we also propose a sequential clinical trial framework that can be a meaningful improvement in the process of assessing clinical treatments. This framework uses normality assumptions to derive the formulas for the sequential scheme. A promising direction for future research would be to drop these assumptions. Another assumption of the framework is the

linearity of the model, it would be interesting to extend it for the non-linear models.

Regarding model fit, an alternative assessment method is that of comparing the model evidence quantities of the candidate models. These quantities are a by product of the sequential framework we propose, without which it would be non-trivial to compute. To pick between the two approaches we reflect on the ultimate goal of our research: ranking the treatments from best to worse. In that regard we need to pick the model that matches reality the best and whose parameters map as well as possible to the desired variables of clinical interest. In other words we are not looking for the right model, but the most useful one. Of the two methods at hand, predictive performance is the more practical and for this reason, we go with predictive performance over model evidence as the framework of choice.

Appendices

Appendix A

Generalised Bayesian Structural Equation Models

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 performed a sensitivity analysis to examine the effect of the data-dependent priors on the final result. In order to amplify the prior effect we used a relatively small sample size, by simulating 200 data points from a standard two-factor model, according to simulation Scenario 1 in Section 2.4. We fit the EZ model with the data-dependent prior of Frühwirth-Schnatter and Lopes (2018); Conti et al. (2014) that protects against Heywood cases for the idiosyncratic variances

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

with $c_0 = 2.5$. Moreover, the following data-independent priors were also used: InvGamma$(0.1, 0.1)$, Half-Cauchy(5), and Uniform$(0, 10)$. The posterior samples from all four priors were used to produce kernel density plots for the posterior of the free Λ elements that are depicted in Figure A.1. As we can see, the posterior density plots are almost identical for all these priors. Similar results were also obtained for the remaining parameters. We therefore conclude that the data-dependent prior does not impact the final results, while helping guard against Heywood cases.

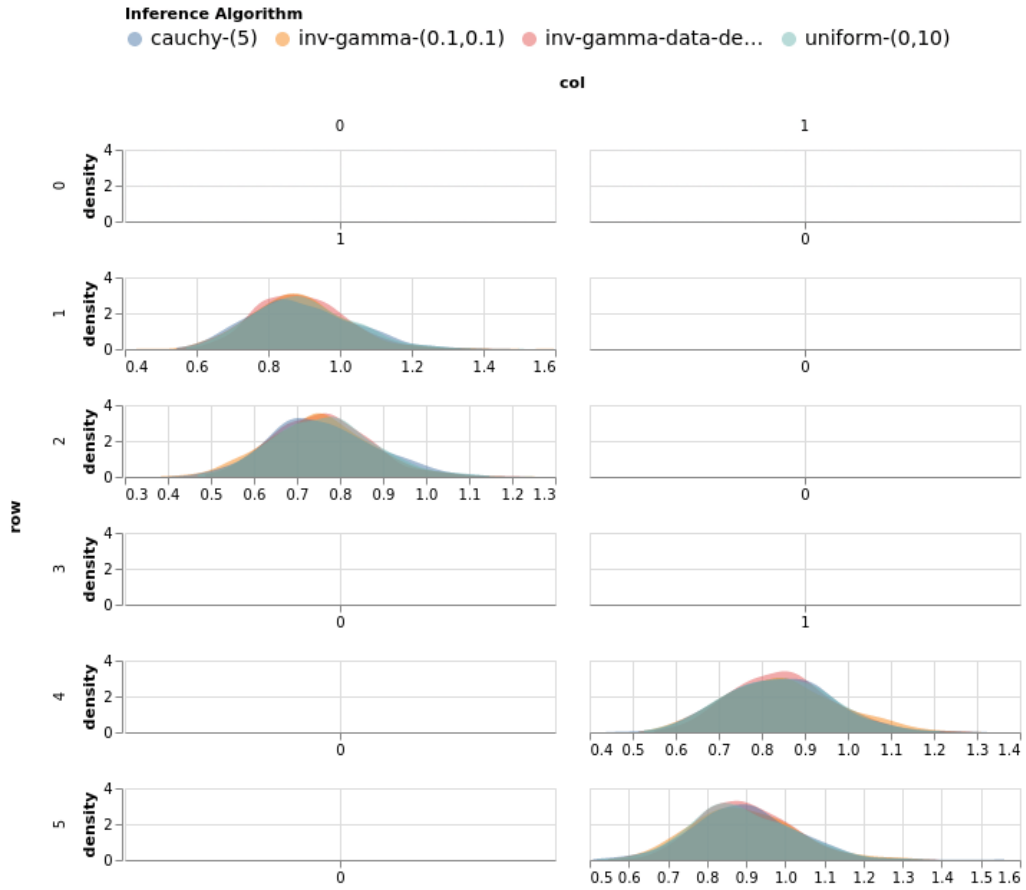

Figure A.1: Posterior density plots of the loading matrix parameters under 4 different prior choices. The model using a data-dependent prior (red) produces identical posterior density plots as three other models using priors independent of the data.

Appendix B

Sequential Latent Variable Modelling

B.1 HMC Implementation Details

One can also improve efficiency by providing good values Hamiltonian MCMC tuning parameters. In particular, the mass matrix needed to propose new values can be estimated accurately by running a longer chain for the first particle, which can then be reused as an initial value for the rest the particles' chains. Similarly, any other MCMC parameters that require a long adaptation phase, such as the step size in HMC, can be learned by running one long chain for the first particle, then used as initial values for the rest of the particles. With this approach, each time the criterion indicates a resample, we run one a long chain for the first particle and then we can afford to run short chains for the rest of the particles. In our experiments running the first particle for 500 steps produced sufficiently accurate initial values for the rest to be run for 10 steps. Further efficiencies can be achieved by running the particle chains in parallel as they are independent of each other.

B.2 IBIS with Laplace Approximation

In this section we derive the approximation to posterior $\pi(z_i|y_i, \theta)$ via the Laplace method. The goal is to use the approximating distribution, $p^L(z_i|y_i, \theta)$ in place of the proposal $q(.)$ Our model

assigns $N(0,4)$ priors to each component of z and also assumes a-priori independence between these components. More generally we can note, even though we do not need it in this derivation, that based on this prior and (3.8), conditional on θ, the z_i rows are independent a-posteriori. Thus, in order to approximate conditional posterior of z given θ, one can approximate the corresponding posteriors of each $z_{i\cdot}$ separately. For the exposition that follows we can drop the subscripts, since the focus of the approximation is the i-th point z_i only, so we adopt the running assumption that z and y represent z_i and y_i for the remaining of the section. We focus on approximating the posterior

$$\pi(z \,|\, y, \theta) \propto f(y \,|\, z, \theta) \exp(-\tfrac{1}{2} z z^T). \tag{B.1}$$

We then target the logarithm of (B.1),

$$\ell(z \,|\, y, \theta) = \log f(y \,|\, z, \theta) - \tfrac{1}{2} z z^T = \sum_{j=1}^{p} \left[y_j \log \pi_j(z, \theta) + (1 - y_j) \log \left\{ 1 - \pi_j(z, \theta) \right\} \right] - \tfrac{1}{2} \sum_{\ell=1}^{k} z_\ell^2 \tag{B.2}$$

In order to apply the Laplace approximation on (B.1) we need the first and second derivatives of (B.2) with respect each z_ℓ for $\ell = 1, \ldots, k$. These are

$$\begin{aligned} \tfrac{\partial}{\partial z_\ell} \ell(z \,|\, y, \theta) &= -z_\ell + \sum_{j=1}^{p} \left\{ \frac{y_j \tfrac{\partial}{\partial z_\ell} \pi_j(z, \theta)}{\pi_j(z, \theta)} - \frac{(1 - y_j) \tfrac{\partial}{\partial z_\ell} \pi_j(z, \theta)}{1 - \pi_j(z, \theta)} \right\} \\ &= -z_\ell + \sum_{j=1}^{p} \tfrac{\partial}{\partial z_\ell} \pi_j(z, \theta) \left\{ \frac{y_j}{\pi_j(z, \theta)} - \frac{1 - y_j}{1 - \pi_j(z, \theta)} \right\} \end{aligned} \tag{B.3}$$

$$\begin{aligned} \tfrac{\partial^2}{\partial z_\ell^2} \ell(z \,|\, y, \theta) &= -1 + \sum_{j=1}^{p} \tfrac{\partial^2}{\partial z_\ell^2} \pi_j(z, \theta) \left\{ \frac{y_j}{\pi_j(z, \theta)} - \frac{1 - y_j}{1 - \pi_j(z, \theta)} \right\} \\ &+ \sum_{j=1}^{p} \tfrac{\partial}{\partial z_\ell} \pi_j(z, \theta) \left[- \frac{y_j \tfrac{\partial}{\partial z_\ell} \pi_j(z, \theta)}{\pi_j(z, \theta)^2} - \frac{(1 - y_j) \tfrac{\partial}{\partial z_\ell} \pi_j(z, \theta)}{\left\{ 1 - \pi_j(z, \theta) \right\}^2} \right] \\ &= -1 + \sum_{j=1}^{p} \tfrac{\partial^2}{\partial z_\ell^2} \pi_j(z, \theta) \left\{ \frac{y_j}{\pi_j(z, \theta)} - \frac{1 - y_j}{1 - \pi_j(z, \theta)} \right\} \\ &- \sum_{j=1}^{p} \left\{ \tfrac{\partial}{\partial z_\ell} \pi_j(z, \theta) \right\}^2 \left[\frac{y_j}{\pi_j(z, \theta)^2} + \frac{1 - y_j}{\left\{ 1 - \pi_j(z, \theta) \right\}^2} \right] \end{aligned}$$

and for $m = 1, \ldots, k$, with $\ell \neq m$

$$\begin{aligned}\tfrac{\partial^2}{\partial z_\ell \partial z_m}\ell(z \,|\, y, \theta) &= \sum_{j=1}^{p} \tfrac{\partial^2}{\partial z_\ell \partial z_m}\pi_j(z,\theta)\left\{\frac{y_j}{\pi_j(z,\theta)} - \frac{1-y_j}{1-\pi_j(z,\theta)}\right\} \\ &- \sum_{j=1}^{p} \tfrac{\partial}{\partial z_\ell}\pi_j(z,\theta)\tfrac{\partial}{\partial z_m}\pi_j(z,\theta)\left[\frac{y_j}{\pi_j(z,\theta)^2} + \frac{1-y_j}{\{1-\pi_j(z,\theta)\}^2}\right]\end{aligned}$$

The above can also provide the Fisher's information matrix $\mathcal{I}(z \,|\, \theta)$, as for $\ell = 1, \ldots, k$, we get

$$\begin{aligned}[\mathcal{I}]_{\ell\ell}(z \,|\, \theta) &= -E\left\{\tfrac{\partial^2}{\partial z_\ell^2}\ell(z \,|\, y, \theta)\right\} = 1 + \sum_{j=1}^{p}\left\{\tfrac{\partial}{\partial z_\ell}\pi_j(z,\theta)\right\}^2\left\{\frac{1}{\pi_j(z,\theta)} + \frac{1}{1-\pi_j(z,\theta)}\right\} \\ &= 1 + \sum_{j=1}^{p}\frac{\left\{\tfrac{\partial}{\partial z_\ell}\pi_j(z,\theta)\right\}^2}{\pi_j(z,\theta)\{1-\pi_j(z,\theta)\}} \qquad \text{(B.4)}\end{aligned}$$

and for $m = 1, \ldots, k$, with $\ell \neq m$

$$[\mathcal{I}]_{\ell m}(z \,|\, \theta) = -E\left\{\tfrac{\partial^2}{\partial z_\ell \partial z_m}\ell(z \,|\, y, \theta)\right\} = \sum_{j=1}^{p}\frac{\tfrac{\partial}{\partial z_\ell}\pi_j(z,\theta)\tfrac{\partial}{\partial z_m}\pi_j(z,\theta)}{\pi_j(z,\theta)\{1-\pi_j(z,\theta)\}} \qquad \text{(B.5)}$$

It remains to calculate $\pi_j(z,\theta)$ and $\frac{\partial}{\partial z_\ell}\pi_j(z,\theta)$. Based on the model in (3.9) we get

$$\pi_j(z,\theta) = \frac{\exp\left(\alpha_j + \sum_{\ell=1}^{k} z_\ell \Lambda_{\ell j}\right)}{1 + \exp\left(\alpha_j + \sum_{\ell=1}^{k} z_\ell \Lambda_{\ell j}\right)}, \quad \tfrac{\partial}{\partial z_\ell}\pi_j(z,\theta) = \frac{\exp\left(\alpha_j + \sum_{\ell=1}^{k} z_\ell \Lambda_{\ell j}\right)\Lambda_{\ell j}}{\left\{1+\exp\left(\alpha_j + \sum_{\ell=1}^{k} z_\ell \Lambda_{\ell j}\right)\right\}^2},$$

which we can plug in to (B.3), (B.4) and (B.5) to obtain the Laplace approximation

$$z \;\sim\; N\left\{\arg\max_z \ell(z \,|\, y, \theta),\; \mathcal{I}(z \,|\, \theta)^{-1}\right\} \qquad \text{(B.6)}$$

where $\arg\max_z \ell(z \,|\, y, \theta)$ can be obtained via the Fisher's Scoring algorithm.

Alternatively we can compute the Observed Information matrix which is negative the Hessian of the log-likelihood evaluated at the mode. For that we will need the second derivative of

$\frac{\partial^2}{\partial^2 z_\ell}\pi_j(z,\theta)$ as follows:

$$\frac{\partial}{\partial z_\ell}\pi_j(z,\theta) = \frac{-\Lambda_{\ell j}^2 \exp\left(\alpha_j + \sum_{\ell=1}^{k} z_\ell \Lambda_{\ell j}\right)}{\left\{1 + \exp\left(\alpha_j + \sum_{\ell=1}^{k} z_\ell \Lambda_{\ell j}\right)\right\}^2}$$

which we can plug into equation (B.3) to compute the matrix.

References

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Algina, J. (1980). A note on identification in the oblique and orthogonal factor analysis models. *Psychometrika 45*(3), 393–396.

Anderson, T. W. and H. Rubin (1956). Statistical inference in factor analysis. In J. Newman (Ed.), *Proceedings of the third Berkley Symposium on Mathematical Statistics and Probability*, Volume 5, pp. 111–150. University of California Press.

Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(3), 269–342.

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics 37*(2), 697–725.

Ansari, A. and K. Jedidi (2000). Bayesian factor analysis for multilevel binary observations. *PSYCHOMETRIKA 65*, 475–496.

Arminger, G. and B. O. Muthén (1998). A bayesian approach to nonlinear latent variable models using the gibbs sampler and the metropolis-hastings algorithm — enhanced reader.

Asparouhov, T. and B. Muthén (2020). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal 0*(0), 1–14.

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

Bartholomew, D. J., M. M. Knott, and I. Moustaki (2011). *Latent variable models and factor*

analysis : a unified approach. Wiley.

Bartlett, M. S. (1957). '{C}omment on {D.V. L}indley's Statistical Paradox'. *Biometrika 44*, 533–534.

Beguin, A. and C. Glas (2001). {MCMC} estimation and some model-fit analysis of multidimensional {IRT} models. *Psychometrika 66*(4), 541–561.

Bekker, P. (1986). A note on the identification of restricted factor loading matrices. *Psychometrika 51*(4), 607–611.

Berger, J. O. and M. Delampady (1987). Testing Precise Hypotheses.

Berger, J. O. and T. Sellke (1987, 3). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association 82*(397), 112–122.

Betancourt, M. (2017). Identifying Bayesian Mixture Models.

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

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association 112*(518), 859–877.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

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

Browne, M. W. (2001). An overview of analytic rotation in Exploratory Factor Analysis. *Multivariate Behavioral Research 36*, 111–150.

Buse, A. (1982, 8). The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. *The American Statistician 36*(3a), 153–157.

Can, S., R. van de Schoot, and J. Hox (2015). Collinear latent variables in multilevel confirmatory factor analysis: A comparison of maximum likelihood and Bayesian estimations. *Educational and Psychological Measurement 75*(3), 406–427.

Carlisle, B., N. Demko, G. Freeman, A. Hakala, N. MacKinnon, T. Ramsay, S. Hey, A. J. London, and J. Kimmelman (2016). Benefit, risk, and outcomes in drug development: a systematic review of sunitinib. *JNCI: Journal of the National Cancer Institute 108*(1).

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of statistical software 76*(1).

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

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

Cattell, R. B. (1966, 4). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research 1*(2), 245–276.

Chib, S. and E. Greenberg (1998a, 6). Analysis of multivariate probit models. *Biometrika 85*(2), 347–361.

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

Chopin, N. (2002, aug). A sequential particle filter method for static models. *Biometrika 89*(3), 539–552.

Chopin, N. (2004, dec). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics 32*(6), 2385–2411.

Chopin, N., P. E. Jacob, and O. Papaspiliopoulos (2012, oct). SMC 2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(3), 397–426.

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

Costa, M. J., W. He, Y. Jemiai, Y. Zhao, and C. Di Casoli (2017). The case for a bayesian approach to benefit-risk assessment: overview and future directions. *Therapeutic innovation & regulatory science 51*(5), 568–574.

Cowles, M. K. (1996). Accelerating monte carlo markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing 6*(2), 101–111.

Cox, D. R. and D. V. Hinkley (1979). *Theoretical statistics*. CRC Press.

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

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(3), 411–436.

Depaoli, S. and J. P. Clifton (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal 22*(3), 327–351.

Dienes, Z. (2016). How bayes factors change scientific practice. *Journal of Mathematical Psychology 72*, 78–89.

Dodgson, J. S., M. Spackman, A. Pearman, and L. D. Phillips (2009). Multi-criteria analysis: a manual.

Doucet, A., N. De Freitas, N. J. Gordon, et al. (2001). *Sequential Monte Carlo methods in practice*, Volume 1. Springer.

Dunn, J. (1973). A note on a sufficiency condition for uniqueness of a restricted factor matrix. *Psychometrika 38*, 141–143.

Dunson, D. B., J. Palomo, and K. Bollen (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.

Elffers, H., J. Bethlehem, and R. Gill (1978). Indeterminacy problems and the interpretation of factor analysis results. *Statistica Neerlandica 32*(4), 181–199.

Engle, R. F. (1984, 1). Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of Econometrics 2*, 775–826.

Erosheva, E. A. and M. S. Curtis (2017). Dealing with reflection invariance in Bayesian factor analysis. *Psychometrika 82*(2), 295–307.

Fong, E. and C. Holmes (2020). On the marginal likelihood and cross-validation. *Biometrika 107*(2), 489–496.

Fox, J. and C. A. W. Glas (2001). Bayesian estimation of a multilevel IRT model using Gibbs

sampling. *Psychometrika 66*(2), 271–288.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Fruhwirth-Schnatter, S. and H. F. Lopes (2010, 7). Parsimonious Bayesian Factor Analysis when the Number of Factors is Unknown. pp. 1–37.

Frühwirth-Schnatter, S. and H. F. Lopes (2018). Parsimonious Bayesian factor analysis when the number of factors is unknown. *Unpublished Working Paper*.

Garnier-Villarreal, M. and T. D. Jorgensen (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods 25*(1), 46.

Garrison Jr, L. P., A. Towse, and B. W. Bresnahan (2007). Assessing a structured, quantitative health outcomes approach to drug risk-benefit analysis. *Health Affairs 26*(3), 684–695.

Geldhof, G. J., K. J. Preacher, and M. J. Zyphur (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods 19*(1), 72.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian data analysis*, Volume 2. CRC press Boca Raton, FL.

Gelman, A., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing 24*(6), 997–1016.

Gelman, A. and X.-L. Meng (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, 163–185.

Gelman, A., X.-L. Meng, and H. Stern (1996). POSTERIOR PREDICTIVE ASSESSMENT OF MODEL FITNESS VIA REALIZED DISCREPANCIES.

Gelman, A., D. Simpson, and M. Betancourt (2017). The prior can often only be understood in the context of the likelihood. *Entropy 19*(10), 555–497.

Geweke, J. and G. Zhou (1996a, 4). Measuring the Pricing Error of the Arbitrage Pricing Theory. *The Review of Financial Studies 9*(2), 557–587.

Geweke, J. and G. Zhou (1996b). Measuring the Pricing Error of the Arbitrage Pricing Theory. *The Review of Financial Studies 9*(2), 557–587.

Ghosh, J. and D. B. Dunson (2009). Default prior distributions and efficient posterior computa-

tion in Bayesian factor analysis. *Journal of Computational and Graphical Statistics 18*(2), 306–320.

Gifford, J. A. and H. Swaminathan (1990). Bias and the Effect of Priors in {B}ayesian Estimation of Parameters of {I}tem {R}esponse {M}odels. *Applied Psychological Measurement 14*(1), 33–43.

Glas, C. A. W. and R. R. Meijer (2003). A Bayesian approach to Person Fit Analysis in Item Response Theory models. *Applied Psychological Measurement 27*(3), 217–233.

Glasziou, P. P. and L. M. Irwig (1995). An evidence based approach to individualising treatment. *Bmj 311*(7016), 1356–1359.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Gordon, N. J., D. J. Salmond, and A. F. Smith (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F-radar and signal processing*, Volume 140, pp. 107–113. IET.

Green, P. J. (1995, 12). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Guo, J. J., S. Pandey, J. Doyle, B. Bian, Y. Lis, and D. W. Raisch (2010). A review of quantitative risk–benefit methodologies for assessing drug safety and efficacy—report of the ispor risk–benefit management working group. *Value in Health 13*(5), 657–666.

Hauck, W. W. and A. Donner (1977, 12). Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association 72*(360a), 851–853.

Heatherton, T., L. Kozlowski, R. Frecker, and K. Fagerstrom (1991). The Fagerstrom Test for Nicotine Dependence : a revision of the Fagerstrom Tolerance Questionnaire. *British Journal of Addiction 86*, 1119–1127.

Hoijtink, H. and R. van de Schoot (2018a). Testing small variance priors using prior-posterior predictive p values. *Psychological Methods 23*(3), 561.

Hoijtink, H. and R. van de Schoot (2018b). Testing small variance priors using prior-posterior predictive p values. *Psychological Methods 23*(3), 561.

Holden, W. L. (2003). Benefit-risk analysis. *Drug safety 26*(12), 853–862.

Hox, J. J., R. van de Schoot, and S. Matthijsse (2012). How few countries will do? comparative survey analysis from a Bayesian perspective. In *Survey Research Methods*, Volume 6, pp. 87–93.

Janssen, R., F. Tuerlinckx, M. Meulders, and P. De Boeck (2000). A hierarchical {IRT} model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics 25*, 285–306.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005, 2). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science 20*(1), 50–67.

Johnson, V. E. and J. H. Albert (2006). *Ordinal data modeling*. Springer Science & Business Media.

Jong, N. K. and P. Stone (1976). Keeney, rl &raiffa, h. decisions with multiple objectives: Preferences and value tradeoffs. In *In Proceedings of the ICML-06 Workshop on Kernel Methods in Reinforcement Learning*. Citeseer.

Jordan, A., F. Krüger, and S. Lerch (2019). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software, Articles 90*(12), 1–37.

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika 34*(2), 183–202.

Jöreskog, K. G. and I. Moustaki (2001). Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research 36*, 347–387.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Kang, T. and A. S. Cohen (2007). {IRT} Model Selection Methods for Dichotomous Items. *Applied Psychological Measurement 31*(4), 331–358.

Kantas, N., A. Beskos, and A. Jasra (2014). Sequential monte carlo methods for high-dimensional inverse problems: a case study for the navier-stokes equations. *SIAM/ASA Journal on Uncertainty Quantification 2*, 464–489.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Publications.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the american statistical association 90*(430), 773–795.

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association 91*(435), 1343–1370.

Keeney, R. L., H. Raiffa, and R. F. Meyer (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.

Kim, J. S. and D. M. Bolt (2007, 12). Estimating Item Response Theory Models Using Markov Chain Monte Carlo Methods. *Educational Measurement: Issues and Practice 26*(4), 38–51.

Kingma, D. P. and M. Welling (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krüger, F., S. Lerch, T. Thorarinsdottir, and T. Gneiting (2020). Predictive inference based on markov chain monte carlo output. *International Statistical Review*.

Kucukelbir, A., D. Tran, R. Ranganath, A. Gelman, and D. M. Blei (2017a). Automatic differentiation variational inference. *The Journal of Machine Learning Research 18*(1), 430–474.

Kucukelbir, A., D. Tran, R. Ranganath, A. Gelman, and D. M. Blei (2017b). Automatic differentiation variational inference. *The Journal of Machine Learning Research 18*(1), 430–474.

Lahdelma, R., J. Hokkanen, and P. Salminen (1998). Smaa-stochastic multiobjective acceptability analysis. *European Journal of Operational Research 106*(1), 137–143.

Lawley, D. N. and A. E. Maxwell (1963). Factor Analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician) 12*(3), 209–229.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*, Volume 711. John Wiley & Sons.

Lee, S.-Y. and X.-Y. Song (2003, 10). Bayesian analysis of structural equation models with dichotomous variables. *Statistics in Medicine 22*, 3073–3088.

Lee, S.-Y. and X.-Y. Song (2010). Evaluation of the bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes.

Lee, S.-Y., X.-Y. Song, and J.-H. Cai (2010, 4). A bayesian approach for nonlinear structural equation models with dichotomous variables using logit and probit links.

http://dx.doi.org.gate3.library.lse.ac.uk/10.1080/10705511003659425 17, 280–302.

Lee, S.-Y. and Y.-M. Xia (2008). A robust Bayesian approach for structural equation models with missing data. *Psychometrika 73*(3), 343.

Levy, R., R. J. Mislevy, and S. Sinharay (2009, 10). Posterior Predictive Model Checking for Multidimensionality in Item Response Theory. *Applied Psychological Measurement 33*(7), 519–537.

Lewandowski, D., D. Kurowicka, and H. Joe (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis 100*(9), 1989 – 2001.

Li, K., S. Luo, S. Yuan, and S. Mt-Isa (2019). A bayesian approach for individual-level drug benefit-risk assessment. *Statistics in medicine 38*(16), 3040–3052.

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. and M. West (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 41–67.

Lu, Z.-H., S.-M. Chow, and E. Loken (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research 51*(4), 519–539.

Lüdtke, O., A. Robitzsch, D. A. Kenny, and U. Trautwein (2013). A general and flexible approach to estimating the social relations model using Bayesian methods. *Psychological Methods 18*(1), 101.

Lynd, L. D. and B. J. O'brien (2004). Advances in risk-benefit evaluation using probabilistic simulation methods: an application to the prophylaxis of deep vein thrombosis. *Journal of clinical epidemiology 57*(8), 795–803.

MacCallum, R. C., M. C. Edwards, and L. Cai (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods*.

MacCallum, R. C., M. Roznowski, and L. B. Necowitz (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin 111*(3),

490.

Maydeu-Olivares, A. and H. Joe (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. (1994a). Posterior predictive p-values. *The Annals of Statistics 22*(3), 1142–1160.

Meng, X.-L. (1994b). Posterior Predictive p-Values.

Merkle, E. C., D. Furr, and S. Rabe-Hesketh (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika 84*(3), 802–829.

Merkle, E. C. and Y. Rosseel (2015). blavaan: Bayesian structural equation models via parameter expansion. *arXiv preprint arXiv:1511.05604*.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British journal of mathematical and statistical psychology 49*(2), 313–334.

Moustaki, I. and M. Knott (2000). Generalized latent trait models. *Psychometrika 65*(3), 391–411.

Mt-Isa, S., C. E. Hallgreen, N. Wang, T. Callréus, G. Genov, I. Hirsch, S. F. Hobbiger, K. S. Hockley, D. Luciani, L. D. Phillips, et al. (2014). Balancing benefit and risk of medicines: a systematic review and classification of available methodologies. *Pharmacoepidemiology and drug safety 23*(7), 667–678.

Mühlbacher, A. C., C. Juhnke, A. R. Beyer, and S. Garner (2016). Patient-focused benefit-risk analysis to inform regulatory decisions: the european union perspective. *Value in Health 19*(6), 734–740.

Murphy, K. P. (2012). *Machine learning : a probabilistic perspective*. MIT Press.

Mussen, F., S. Salek, and S. Walker (2007). A quantitative approach to benefit-risk assessment of medicines–part 1: the development of a new model using multi-criteria decision analysis. *Pharmacoepidemiology and drug safety 16*(S1), S2–S15.

Mussen, F., S. Salek, and S. Walker (2009). Benefit-risk appraisal of medicines. *A systematic approach to decision-making*.

Muthén, B. and T. Asparouhov (2012). Bayesian Structural Equation Modeling: A more flexible representation of substantive theory. *Psychological Methods 17*, 313–335.

Muthén, L. K. and B. Muthén (2017). *Mplus user's guide: Statistical analysis with latent variables, user's guide*. Muthén & Muthén.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo 2*(11), 2.

Nissen, S. E. and K. Wolski (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine 356*(24), 2457–2471.

Oberski, D. L., G. H. van Kollenburg, and J. K. Vermunt (2013). A monte carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification 7*(3), 267–279.

Oravecz, Z., F. Tuerlinckx, and J. Vandekerckhove (2011). A hierarchical latent stochastic differential equation model for affective dynamics. *Psychological Methods 16*(4), 468.

Patz, R. J. and B. W. Junker (1999a). A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models. *Journal of Educational and Behavioral Statistics 24*(2), 146.

Patz, R. J. and B. W. Junker (1999b, 8). Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses. *Journal of Educational and Behavioral Statistics 24*(4), 342–366.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Pinson, P. and R. Girard (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy 96*, 12–20.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics 9*(3), 523–539.

Pokropek, A., E. Davidov, and P. Schmidt (2019). A monte carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal 26*(5), 724–744.

Pokropek, A., P. Schmidt, and E. Davidov (2020). Choosing priors in bayesian measurement invariance modeling: A monte carlo simulation study. *Structural Equation Modeling: A*

Multidisciplinary Journal 27(5), 750–764.

Ponce, R. A., S. M. Bartell, E. Y. Wong, D. LaFlamme, C. Carrington, R. C. Lee, D. L. Patrick, E. M. Faustman, and M. Bolger (2000). Use of quality-adjusted life year weights with dose-response models for public health decisions: A case study of the risks and benefits of fish consumption. *Risk Analysis 20*(4), 529–542.

Pramanik, S., V. E. Johnson, and A. Bhattacharya (2021). A modified sequential probability ratio test. *Journal of Mathematical Psychology 101*, 102505.

Richardson, C. G. and P. A. Ratner (2005). A confirmatory factor analysis of the Fagerstrom Test for Nicotine Dependence. *Addictive Behaviors 30*(4), 697 – 709.

Rubin, D. B. (1984, 12). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics 12*(4), 1151–1172.

Rue, H., S. Martino, and N. Chopin (2009, 4). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Sahu, S. K. (2002). Bayesian Estimation and Model Choice in Item Response Models. *Journal of Statistical Computation and Simulation 72*(3), 217–232.

Sato, M. (1991). A study of an identification problem and substitute use of principal component analysis in factor analysis. *Hiroshima Mathematical Journal 22*(3), 607–611.

Schäfer, C. and N. Chopin (2013). Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing 23*(2), 163–184.

Scheines, R., H. Hoijtink, and A. Boomsma (1999a). Bayesian estimation and testing of structural equation models. *Psychometrika 64*(1), 37–52.

Scheines, R., H. Hoijtink, and A. Boomsma (1999b). Bayesian estimation and testing of structural equation models. *Psychometrika 64*(1), 37–52.

Scheuerer, M. and T. M. Hamill (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review 143*(4), 1321 – 1334.

Schnuerch, M. and E. Erdfelder (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological methods 25*(2), 206.

Schönbrodt, F. D., E.-J. Wagenmakers, M. Zehetleitner, and M. Perugini (2017). Sequential

hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological methods 22*(2), 322.

Shaffer, M. L. and K. L. Watterberg (2006). Joint distribution approaches to simultaneously quantifying benefit and risk. *BMC medical research methodology 6*(1), 1–8.

Shakespeare, T. P., V. J. Gebski, M. J. Veness, and J. Simes (2001). Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. *The Lancet 357*(9265), 1349–1353.

Sheng, Y. (2008). A {MATLAB} Package for {Markov} {Chain} {Monte Carlo} with a Multi-Unidimensional {IRT} Model. *Journal of Statistical Software 28*(10), 1–20.

Sinharay, S. (2005a). Assessing Fit of Unidimensional Item Response Theory Models Using a Bayesian Approach. *Journal of Educational Measurement 42*(4), 375–394.

Sinharay, S. (2005b). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement 42*(4), 375–394.

Sinharay, S. (2005c). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement 42*(4), 375–394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology 59*(2), 429–449.

Sinharay, S., M. S. Johnson, and H. S. Stern (2006, 7). Posterior Predictive Assessment of Item Response Theory Models. *Applied Psychological Measurement 30*(4), 298–321.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized latent variable modeling : multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC.

Sobel, M. E. and G. W. Bohrnstedt (1985). Use of Null Models in Evaluating the Fit of Covariance Structure Models. *Sociological Methodology 15*, 152.

Song, X.-Y., Z.-H. Lu, J.-H. Cai, and E. H.-S. Ip (2013). A Bayesian modeling approach for generalized semiparametric structural equation models. *Psychometrika 78*(4), 624–647.

Steiger, J. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika 44*(2), 157–167.

Stromeyer, W. R., J. W. Miller, R. Sriramachandramurthy, and R. DeMartino (2015). The prowess and pitfalls of Bayesian structural equation modeling: Important considerations

for management research. *Journal of Management 41*(2), 491–520.

Sutton, A. J., N. J. Cooper, K. R. Abrams, P. C. Lambert, and D. R. Jones (2005). A bayesian approach to evaluating net clinical benefit allowed for parameter uncertainty. *Journal of clinical epidemiology 58*(1), 26–40.

Talhouk, A., A. Doucet, and K. Murphy (2012, 7). Efficient Bayesian Inference for Multivariate Probit Models With Sparse Inverse Correlation Matrices. *Journal of Computational and Graphical Statistics 21*(3), 739–757.

Tervonen, T. (2014). Jsmaa: open source software for smaa computations. *International Journal of Systems Science 45*(1), 69–81.

Tervonen, T. and J. R. Figueira (2008). A survey on stochastic multicriteria acceptability analysis methods. *Journal of Multi-Criteria Decision Analysis 15*(1-2), 1–14.

Tervonen, T., G. Van Valkenhoef, E. Buskens, H. L. Hillege, and D. Postmus (2011). A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Statistics in Medicine 30*(12), 1419–1428.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(3), 611–622.

Vamvourellis, K., K. Kalogeropoulos, and I. Moustaki (2021). Generalised bayesian structural equation modelling. *arXiv preprint arXiv:2104.01603*.

Van De Schoot, R., A. Kluytmans, L. Tummers, P. Lugtig, J. Hox, and B. Muthén (2013). Facing off with scylla and charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in psychology 4*, 770.

Van De Schoot, R., S. D. Winter, O. Ryan, M. Zondervan-Zwijnenburg, and S. Depaoli (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods 22*(2), 217–239.

Van Erp, S., J. Mulder, and D. L. Oberski (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods 23*(2), 363.

Vitoratou, S., I. Ntzoufras, and I. Moustaki (2016). Explaining the behavior of joint and marginal monte carlo estimators in latent variable models with independence assumptions. *Statistics and Computing 26*(1), 333–348.

Vitoratou, V., I. Ntzoufras, and I. Moustaki (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.

Waddingham, E., S. Mt-Isa, R. Nixon, and D. Ashby (2016). A bayesian approach to probabilistic sensitivity analysis in structured benefit-risk assessment. *Biometrical Journal 58*(1), 28–42.

Wallach, J. D., K. Wang, A. D. Zhang, D. Cheng, H. K. G. Nardini, H. Lin, M. B. Bracken, M. Desai, H. M. Krumholz, and J. S. Ross (2020). Updating insights into rosiglitazone and cardiovascular risk through shared data: individual patient and summary level meta-analyses. *bmj 368*.

Williams, J. (1978). A definition for the common-factor analysis model and the elimination of problems of factor score indeterminacy. *Psychometrika 43*(3), 293–306.

Yang, M. and D. B. Dunson (2010). Bayesian semiparametric structural equation models with latent variables. *Psychometrika 75*(4), 675–693.

Yuan, Y. and D. P. MacKinnon (2009). Bayesian mediation analysis. *Psychological Methods 14*(4), 301.