KU LEUVEN

FACULTY OF SCIENCE

Generalized Linear Latent and Mixed Model:

method, bayesian estimation, advantages, and applications to educational data.

Jose Manuel Rivera Espejo

Supervisor: Prof. Geert Molenbegrhs
Affiliation *(optional)*

Co-supervisor: Prof. Wim Van den Noortgate *(optional)*
Affiliation *(optional)*

Thesis presented in fulfillment of the requirements for the degree of Master of Science in Statistics and Data Science for Social, Behavioral and Educational Sciences

Academic year 2020-2021

Dedication

To Manuel, for being my friend and father.
To Margarita, Susan, and Marysu, for their relentless encouragement.
To Ana, for showing me the value of family, here in this moorland.
To both of you, as you are always in my mind.
And to all that knowingly or not, help me to get here.
I am lucky due to all of you.
I hope I make you all proud.

A Manuel, por ser mi amigo y mi padre.
A Margarita, Susan y Marysu, por su incansable aliento.
A Ana, por mostrarme el valor de la familia, aquí en este páramo.
A ustedes dos, que siempre las tengo en mente.
Y a todos los que sabiendolo o no, me ayudaron a llegar aquí.
Soy un suertudo gracias todos ustedes.
Espero llenarlos de orgullo.

Abstract

(work in progress)
Keywords:

Contents

List of Figures

List of Tables

Abbreviations

CFA	Confirmatory Factor Analysis.
CLT	Central Limit Theorem.
DDM	Dual Dependence Models.
EFA	Exploratory Factor Analysis.
GLLAMM	Generalized Linear Latent and Mixed Model.
GLM	Generalized Linear Model.
GLMM	Generalized Linear Mixed Model.
HMC	Hamiltonian Monte Carlo.
iHMC	Interleaved Hamiltonian Monte Carlo.
IRT	Item Response Theory models.
MCMC	Markov Chain Monte Carlo.
ML	Maximum Likelihood.
MSEM	Multilevel Structural Equation Model.
SEM	Structural Equation Model.
VI	Variational Inference.

Chapter 1

Introduction

1.1 Preliminar considerations

Local independence is one of the key assumptions of Item Response Theory (IRT) models, and it is comprised of two parts: (i) local item independence and (ii) local individual independence [3, 24]. In the former case, the assumption entails that the individual's response to an item does not affect the probability of endorsing another item, after conditioning on the individual's ability. While in the case of the latter, the assumption considers that an individual's response to an item is independent of another person's response to that same item [58].

The literature has shown that IRT models are not robust to the violation of local independence. The transgression of the assumption affects model parameter estimates, inflates measurement reliabilities and test information, and underestimates standard errors (see Yen [69], Chen and Thissen [9], and Jiao et al. [28]).

However, item response data arising from educational assessments often display several types of dependencies, e.g. testlets, where items are constructed around a common stimulus [67]; the measurement of multiple latent traits within individuals [58]; cluster effects, where correlation among individuals results from the sampling and measurement mechanism used to gather the data [57]; among others. A good motivating example, that will permeate this research, is the reading comprehension sub-test, from the Peruvian public teaching career national assessment. The test is designed to measure three hierarchically nested sub-dimensions of reading comprehension: literal, inferential, and reflective abilities. Furthermore, the items are bundled together in testlets related to a common text or passage. Finally, multiple cluster effects are present, e.g. at the region, and district level, just to mention a few.

Recent studies have proposed IRT Dual Dependency Models (DDM) to deal with the testlets and individual clustering dependencies observed in the data [16, 15, 14, 28, 12, 13, 58, 7]. The majority of these representations have been developed under the Bayesian framework, and they are similar in parametrization to multilevel models. On the other hand, an almost independent line of research, the Generalized Linear Latent and Mixed Models (GLLAMM) [51, 53, 62, 54], have extended the capabilities of hierarchical models on the estimation of multiple latent traits at different hierarchical levels. These developments have been motivated mostly under the frequentist framework, and they are similar in parametrization to a Multilevel Structural Equation Model (MSEM).

While the initial sense is that both developments are independent of each other, follow-

ing their literature, one can easily notice that they share more than a resemblance. Both follow a multilevel/hierarchical multidimensional approach to account for the clustering of persons within the samples and the item bundles (DDM), or the latent structures within the individuals (GLLAMM). However, it is important to point out that in some cases the model parametrization between the two developments differs in a way, that some of them appear to be useful only under their specific contexts. Fortunately, their integration under the Bayesian framework is not only trivial but can be motivated under either type of model.

The benefits of the integration revolve around two facts: (i) the educational data often presents all of the aforementioned dependencies and more, as in the motivating example; and (ii) as it was hinted in the second paragraph, to reach appropriate conclusions from the parameter estimates, IRT models need to account for all of these dependencies. The latter is particularly important as, more often than not, a researcher is interested in producing inferences at the structural level of the model, i.e. how a different set of manifest variables explain the variability in the latent variables, or how the latent variables explain other manifest or latent variables, at different levels. As an example, one might be interested in finding evidence if the latent "abilities" of the teachers are explained by their initial educational conditions, i.e. if they were educated in an institute, university, or both. The main purpose of this would be to identify the type of teacher that might benefit more from the in-service training[1], offered by the national educational authorities, making the intervention cost-effective.

From the previous description, one can infer that the proposed IRT representation would be complex and highly dimensional. Moreover, as educational assessments are usually scored in a binary way (the individual either endorse or did not the item), and because not all individuals are assessed by all the items, the model will be trained on sparse data. From the modeling perspective, neither of the previous points presents a challenge for the bayesian framework. However, it has long been recognized that complex parametrizations, that allow this powerful modeling schemes, introduce pathologies that make Markov Chain Monte Carlo methods (MCMC) face performance challenges [17, 18, 44, 45, 5]. This is highly relevant because, in order to make inferences about the posterior distribution of the parameters, the chains need to achieve three requirements, highly related to the performance of the method: stationarity, convergence, and good mixing [38].

Throughout the bayesian IRT literature, one often finds that four solutions are offered to ensure the fulfillment of the previous requirements, and they can be classified into two broad groups: (i) solutions that involve changing the settings of the MCMC methods, and (ii) solutions that involve readjusting the Bayesian model.

In the first category, we find two proposals: (a) increasing the number of iterations per chain, with large burn-in and thinning processes, and (b) designing model-specific MCMC algorithms. The easiest to implement and more prevalent in the literature is the former, e.g. Fujimoto [15] used chains with $60,000$ iterations, where $15,000$ were discarded and the remaining were thinned in jumps of 3; while Fujimoto [14] used $225,000$ iterations, with burn-in of $30,000$ and thinning with jumps of 15. Among the drawbacks of this solution are the large computational times; the user involvement in deciding the specific setting for the process, which could be different for different parameters in the same model;

[1]Intervention designed with the purpose of potentiating specific abilities in teachers that are currently part of the public teaching career.

and finally, the lack of confidence that larger chain iterations actually produce a proper posterior investigation. On the other hand, several authors have developed high-tech MCMC algorithms that aim to optimize their performance within a particular class of models [45]. In these cases, the developers re-evaluate not only the use of the programming language, with the purpose of speeding and improving performance (e.g. Fujimoto [15]); but also the inclusion of ad-hoc model assumptions, like the ones used in staple software developments like Mplus [41] or Stata [52]. It is clear from the previous that this solution is not accessible to all researchers, either because of the lack of programming skills, or the restrictive cost of access involved in acquiring the software. But more importantly, these solutions are not always applicable to a wider framework of similar models [45].

In the second category, we also find two proposed solutions: (a) re-write the model in an alternative parametrization, and (b) encode prior information through the prior distributions. On both solutions, the purpose is to ensure the identification of the parameters within the model, which helps to stabilize the MCMC procedure [19]. An example of the former is Fujimoto [15], which decomposed the items' discriminatory parameters into overall and specific item discriminations. For the latter, Fujimoto [16] used informative priors also for the items' discrimination parameters.

More often than not, researchers use two or more of the aforementioned solutions to reach an acceptable performance in the chains. However, as point out by Betancourt and Girolami [5], even the most simple hierarchical models present formidable pathologies, that no simple correction can be performed to visit the posterior distribution properly. This is true no matter the rotation/rescaling of the parameter, or the amount of data. In this context, several authors [17, 18, 44, 45, 5] showed that prior information can be included in the model, not only through the prior distributions but also by encoding it in the model itself, changing the posterior sampling geometries by removing the dependence of the parameters on other sampled parameters, therefore favoring the performance of MCMC chains.

Given all of the above, the present research will focus its attention on showing how easy is to account for all of the dependencies that educational data often display, under the GLLAMM framework. Furthermore, given that only the literature related to gaussian hierarchical models have shown the benefits of changing the posterior sampling geometries, through the use of the non-centered parameterization [17, 18, 44, 45, 5], it seems sensible to provide a similar assessment for nonlinear hierarchical models, and in particular, the ones with latent stochastic processes like IRT [45]. Finally, the research will apply the newfound knowledge to data coming from a large Teacher's standardized educational assessments from Peru.

1.2 Objectives

As mentioned in the previous paragraph, the present research has a three-fold purpose:

1. Motivate the Bayesian GLLAMM for binary outcomes [51, 53, 62, 54]. The representation will emphasize the modeling of multiple hierarchical latent structures and testlets. This, in turn, will effectively blur the division between the GLLAMM framework and IRT models.

2. Empirically evaluate the benefits of changing the posterior sampling geometry. The

emphasis here will be on comparing the centered and non-centered parametrizations [17, 18] in terms of performance of the chains, the parameters' recovery, and the ability of the model to produce appropriate inferences.

3. Apply the model and its parametrization to a real data setting. Here the emphasis will be to assess the conclusions arrived from the application of the model and what they could imply for the educational authorities.

Given the aforementioned goals, the researcher believes the master's thesis contributes to the literature in two aspects:

1. In a theoretical and methodological sense, as the research is focused on describing a model that effectively controls for the multiple dependencies usually observed in educational datasets; and

2. In a more practical sense, as the study will provide empirical evidence if changing the sampling posterior geometries could (could not) benefit the performance of MCMC methods and therefore the inferences, under IRT models.

Finally, it is important to mention, that the computational implementation of the method will be developed in `Stan` [65] and `R` [48, 64].

1.3 Organization

Chapter 2 will motivate the GLLAMM for dichotomous outcomes, and define its components. Chapter 3 will describe the bayesian framework, its benefits and shortcomings. Furthermore, it will outline the evidence behind the change in posterior sampling geometries, and the computational implementation of the model. Chapter 4 will show the results of an empirical simulation study designed to assess the benefits of the re-parametrization, proposed in the previous chapter. Chapter 5 will describe the instruments, the data collection process, and scales under analysis, for a large standardized educational assessment. In addition, the chapter will show the conclusions achieved by the application of the model in the said data. Finally, Chapter 6, will discuss the conclusion for the research, and it will outline the path of future research topics that can be derived from the present effort.

Chapter 2

The GLLAMM for dichotomous outcomes

The Generalized Linear Latent and Mixed Model (GLLAMM) is a framework that unifies a wide range of latent variable models. Developed by Rabe-Hesketh and colleagues [51, 53, 52, 62, 54], the method was motivated by the need of a Multilevel Structural Equation Model (MSEM) that accommodated for unbalanced data, noncontinuous responses and cross-level effects among latent variables. The authors focused its development mainly from the frequentist perspective, however, they offered a general guidance on implementing the model under the bayesian framework (see Skrondal and Rabe-Hesketh [62]).

2.1 Model motivation

Consider a large standardized assessment composed of three sub-test designed to evaluate the reading comprehension, mathematical reasoning, and pedagogical knowledge of teachers; where each sub-test has several dichotomously scored items.

Focusing on the first sub-test, the items were designed to measure only one of the three hierarchically nested sub-dimensions of reading comprehension: literal, inferential, and reflective abilities. Furthermore, it is assumed the three sub-dimensions are all that is needed to measure the reading comprehension ability, effectively making this scale, the highest level latent variable in the model, similar to a hierarchical Confirmatory Factor Analysis (CFA). Finally, the items were bundled in groups of five to a common text or passage, i.e. testlets, that provided the stimulus over which the individual is assessed. Figure 2.1 shows the path diagram of the hypothesized dimensional structure, for the hierarchical cross-classified IRT model corresponding with the instrument.

With the purpose of providing an easier motivation of the model, we will not consider yet the cluster effects; however, later in the presentation we will show how easy is to introduce them in the model. Just for future reference, under this example, one expects to observe clustering effects, because individuals from different regions did not have the same educational opportunities, effectively causing differences among them at a regional level.

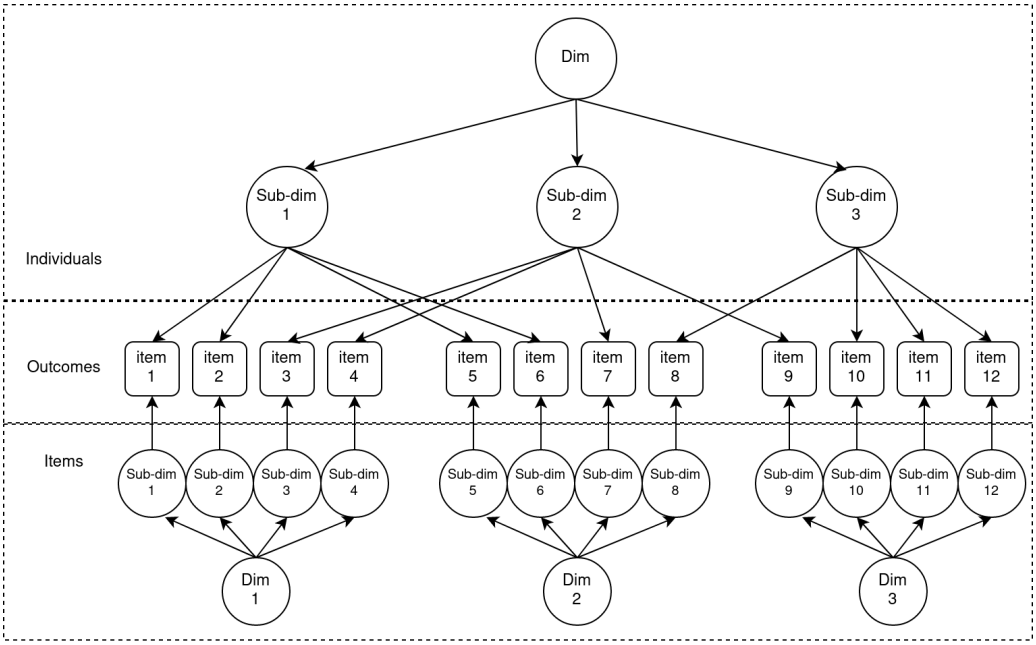

Figure 2.1: Path diagram of the dimensional structure for a hierarchical cross-classified IRT model. Squares represent dichotomous manifest variables, and circles represent latent variables. The figure is based on a reduced set of items, while the errors and scales of the latent variables are not represented. Different sub-dimensions at the individuals block represent the literal, inferential and reflective abilities, while at the items blocks represent the items' difficulties. The dimensions at the individuals block represent the reading comprehension ability, while at the items block represent the multiple testlets.

2.2 Model definition

Following Rabe-Hesketh et al. [51, 53], we continue defining the GLLAMM in two parts: (i) the response model, and (ii) the latent structure.

In case the reader is interested in outcomes different than the dichotomous case, refer to Appendix A.1.

2.2.1 Response model

Conditional to all parameters $\mathbf{\Omega} = \{\boldsymbol{\beta}, \mathbf{\Lambda}, \mathbf{\Theta}, \mathbf{\Psi}, \mathbf{\Gamma}\}$, including all the latent variables, and the "stacked" vector of covariates for the first level ($\mathbf{X}$) and in the structural part ($\mathbf{W}$); the response model can be represented by a Generalized Linear Model (GLM) [43, 37] with a distributional and a systematic part. The latter composed of a linear predictor and a link function.

For the distributional part, the dichotomous items y_{jkd} are modeled at the first level by a Bernoulli probability mass function $f(\cdot)$, in the following form:

$$f\left(y_{jkd} = 1 \mid \mathbf{X}, \mathbf{W}, \mathbf{\Omega}\right) = \pi_{jkd}^{n}(1 - \pi_{jkd})^{1-n} \tag{2.1}$$

where n denotes the endorsement of the item in the Bernoulli trial. On the other hand, for the systematic part, the probability of endorsing the item π_{jkd} is linked to a linear

predictor v_{jkd} through an inverse-link function $h(\cdot)$, in the following form:

$$P\left(y_{jkd}=1 \mid \mathbf{X},\mathbf{W},\mathbf{\Omega}\right)=\pi_{jkd}=h(\tau_k+v_{jkd}) \tag{2.2}$$

where τ_k is k'th item threshold, assumed to be zero for the binary case [51], while the inverse-link function can be defined in three ways:

$$h(x)=\begin{cases}\exp(x)[1+\exp(x)]^{-1}\\ \Phi(x)\\ \exp(-\exp(x))\end{cases} \tag{2.3}$$

corresponding to the logistic, standard normal $\Phi(x)$, and Gumbel (extreme value type I) cumulative distributions, respectively. It is usual to report the last in terms of link functions $g(\cdot)=h^{-1}(\cdot)$. In that case, these corresponds to the well known logit, probit and complementary log-log, respectively. Finally, the linear predictor is defined by:

$$v_{jkd}=\sum_{p=1}^{P}x_{jp}\beta_p+\sum_{m=2}^{M+1}\sum_{k=1}^{K_{(m)}}\eta_k^{(m)}\alpha_k^{(m)}+\sum_{l=2}^{L+1}\sum_{d=1}^{D_{(l)}}\theta_{jd}^{(l)}\lambda_d^{(l)} \tag{2.4}$$

where individuals are indexed by $j=1,\ldots,J$, and J represents the total number of individuals in the sample. β_p denotes regression parameter for the x_{jp} explanatory variable with $p=1,\ldots,P$, and P denoting the total number of explanatory variables. $\eta_k^{(m)}$ is the kth item latent dimension at level m with loading $\alpha_k^{(m)}$, where $k=1,\ldots,K_{(m)}$, $K_{(m)}$ denotes the number of dimensions at level $m=2,\ldots,M+1$, and M represents the number of levels in the items block. $\theta_{jd}^{(l)}$ denotes the individual's dth latent dimension at level l with loading $\lambda_d^{(l)}$, where $d=1,\ldots,D_{(l)}$, $D_{(l)}$ represents the number of dimensions at level $l=2,\ldots,L+1$, and L denotes the number of levels in the individuals block.

Notice equation (2.4) can be re-written in matrix form in the following way:

$$v_{jkd}=\mathbf{X}_j\,\boldsymbol{\beta}+\sum_{m=2}^{M+1}\boldsymbol{\eta}^{(m)}\boldsymbol{\alpha}^{(m)}\mathbf{A}_j^{(m)}+\sum_{l=2}^{L+1}\boldsymbol{\theta}_j^{(l)}\boldsymbol{\lambda}^{(l)}\mathbf{B}_j^{(l)} \tag{2.5}$$

where $\mathbf{X}_j$ represents the individual's design matrix of explanatory variables that maps the parameter vector $\boldsymbol{\beta}$ to the linear predictor; and $\mathbf{X}=[\mathbf{X}_1^T,\ldots,\mathbf{X}_J^T]^T$ the "stacked" design matrix of $\mathbf{X}_j$. Moreover, $\boldsymbol{\eta}^{(m)}=[\eta_1^{(m)},\ldots,\eta_{K_{(m)}}^{(m)}]^T$, and $\boldsymbol{\alpha}^{(m)}=[\alpha_1^{(m)},\ldots,\alpha_{K_{(m)}}^{(m)}]^T$ are the vectors of the item's latent dimensions with corresponding loadings at level m, mapped by a block matrix $\mathbf{A}_j^{(m)}$. Similarly, $\boldsymbol{\theta}_j^{(l)}=[\theta_{j1}^{(l)},\ldots,\theta_{jD_{(l)}}^{(l)}]^T$, and $\boldsymbol{\lambda}^{(l)}=[\lambda_1^{(l)},\ldots,\lambda_{D_{(l)}}^{(l)}]^T$ are the vectors of the individual's latent dimensions with corresponding loadings at level l, mapped by a block matrix $\mathbf{B}_j^{(l)}$.

Finally, in order to have a more concise representation of the model, we can re-express equation (2.5) in the following way:

$$\begin{aligned}v_{jkd}&=\mathbf{X}_j\,\boldsymbol{\beta}+\boldsymbol{\eta}\,\boldsymbol{\alpha}\,\mathbf{A}_j+\boldsymbol{\theta}\,\boldsymbol{\lambda}\,\mathbf{B}_j\\&=\mathbf{X}_j\,\boldsymbol{\beta}+\mathbf{\Theta}\,\mathbf{\Lambda}\,\mathbf{H}_j\end{aligned} \tag{2.6}$$

where $\boldsymbol{\alpha}=[\boldsymbol{\alpha}^{(2)T},\ldots,\boldsymbol{\alpha}^{(M+1)T}]^T$ and $\boldsymbol{\lambda}=[\boldsymbol{\lambda}^{(1)T},\ldots,\boldsymbol{\lambda}^{(L+1)T}]^T$ represent all the loadings corresponding to the items and individuals dimensions at all levels; whereas $\boldsymbol{\eta}=$

$[\boldsymbol{\eta}^{(2)T}, \ldots, \boldsymbol{\eta}^{(M+1)T}]^T$ and $\boldsymbol{\theta} = [\boldsymbol{\theta}^{(2)T}, \ldots, \boldsymbol{\theta}^{(L+1)T}]^T$ represent the latent dimensions and sub-dimensions of items and individuals at all levels, respectively. Consequently, $\mathbf{A}_j$ and $\mathbf{B}_j$ are their mapping block matrices. Furthermore, $\boldsymbol{\Lambda} = [\boldsymbol{\alpha}^T, \boldsymbol{\lambda}^T]^T$ and $\boldsymbol{\Theta} = [\boldsymbol{\eta}^T, \boldsymbol{\theta}^T]^T$ to represent the "stacked" vector of loadings, and dimensions, respectively; and $\mathbf{H}_j$ its mapping block matrix.

Considering figure 2.1 as reference, we would have an empty level-1 covariates matrix $\mathbf{X}_j$, as we do not have explanatory variables ($P = 0$). Moreover, we have $M = 2$ levels at the items block, with $K_2 = 12$ and $K_3 = 3$. Consequently, $\boldsymbol{\eta}^{(2)} = [\eta_1^{(2)}, \ldots, \eta_{12}^{(2)}]^T$ and $\boldsymbol{\eta}^{(3)} = [\eta_1^{(3)}, \eta_2^{(3)}, \eta_3^{(3)}]^T$ denotes the items' difficulties and testlet effects, respectively. On the other hand, we have $L = 2$ levels in the individuals block, with $D_2 = 3$ and $D_3 = 1$. Therefore, $\boldsymbol{\theta}^{(2)} = [\theta_1^{(2)}, \theta_2^{(2)}, \theta_3^{(2)}]^T$ which denotes the literal, inferential and reflective abilities; whereas $\boldsymbol{\theta}^{(3)} = \theta_1^{(3)}$ represent the reading comprehension latent variable. Furthermore, since we are trying to express an IRT model, it makes sense to put some restrictions to the parameter set. In this case, we establish $\boldsymbol{\alpha}^{(2)} = -\boldsymbol{\lambda}^{(2)}$ with $\boldsymbol{\lambda}^{(2)} = [\lambda_1^{(2)}, \ldots, \lambda_{12}^{(2)}]^T$ representing the item's discriminatory parameter, where $\boldsymbol{\lambda}^{(2)} > \mathbf{0}$. This resemble to a multidimensional generalization of the linear predictor observed in the archetypical Rasch [56], or 2PL [33] models, i.e. $\lambda_d^{(2)}(\theta_{jd}^{(2)} - \eta_k^{(2)})$, where $|\lambda_d^{(2)}| = |\alpha_k^{(2)}|$ denotes the discriminatory power of item k, $\eta_k^{(2)}$ its difficulty, and $\theta_{jd}^{(2)}$ the ability of the individual at dimension d. In addition, $\boldsymbol{\lambda}^{(3)} = [\lambda_1^{(3)}, \lambda_2^{(3)}, \lambda_3^{(3)}]^T$ would represent the loadings from reading comprehension to their respective sub-dimensions; whereas $\boldsymbol{\alpha}^{(3)} = [\alpha_1^{(3)}, \alpha_2^{(3)}, \alpha_3^{(3)}]^T$ would represent the item-specific loadings from the testlets, usually set as $[1, 1, 1]^T$, indicating they explain directly the items difficulties at the lower level.

Finally, notice that trough the use of $\mathbf{A}_j$ and $\mathbf{B}_j$ design block matrices, and more generally with $\mathbf{H}_j$, the model departs from the traditional multivariate framework for formulating structural models, i.e. a wide data format; and adopts a univariate approach, i.e. a long data format. The former stores the subject's repeated outcomes in a single row, with multiple response vectors and explanatory variables appended column-wise to the outcome data. The later stores the subject's repeated outcomes in a single "stacked" response vector with as many rows as there are repeated measurements, and explanatory variables appended column-wise to the outcome data, distinguished from each other, by a design block matrix.

Cluster effects

Considering the previous, we can see that modeling clustering is as easier as to add more random effects to the linear predictor defined in equation (2.4), in the following form:

$$\begin{aligned} v_{jkdc} &= v_{jkd} + \sum_{c=1}^{C} \delta_c \\ &= v_{jkd} + \boldsymbol{\delta}\mathbf{Z}_j \end{aligned} \tag{2.7}$$

where $c = 1, \ldots, C$, which denotes the number of clusters, v_{jkd} is defined as in equation (2.4), and $\mathbf{Z_j}$ is a design block matrix.

2.2.2 Latent structure

The structural model for the latent variables is represented in the following form:

$$\underset{(S\times 1)}{\boldsymbol{\Theta}} = \underset{(S\times S)}{\boldsymbol{\Psi}}\ \underset{(S\times 1)}{\boldsymbol{\Theta}} + \underset{(S\times Q)}{\boldsymbol{\Gamma}}\ \underset{(Q\times 1)}{\mathbf{W}} + \underset{(S\times 1)}{\boldsymbol{\zeta}} \tag{2.8}$$

where $S = K + D$, $K = \sum_m K_m$, and $D = \sum_l D_l$. Furthermore, $\boldsymbol{\Psi}$ and $\boldsymbol{\Gamma}$ are parameter matrices that map the relationship between the latent variables $\boldsymbol{\Theta}$, and the "stacked" vector of covariates $\mathbf{W}$, respectively; while $\boldsymbol{\zeta}$ is a vector of errors or disturbances. It is important to indicate that $\mathbf{W}$ considers a different set of covariates from $\mathbf{X}$, as we hypothesize they explain the variability in the latent variables at different levels.

Notice equation (2.8) is the generalization of a single-level Structural Equation Models (SEM) to a multilevel setting, however, the main difference in the GLLAMM representation is that the latent variables may vary at different levels.

Additionally, considering that $\boldsymbol{\Theta}$ has no feedback effects, is permuted, and sorted according to the levels of interest, then $\boldsymbol{\Psi}$ will be a strictly upper triangular matrix. In this regard, (i) the absence of feedback loops imply the method deals with non-recursive models, i.e. none of the latent variables are specified as both causes and effects of each other [31]; and (ii) the strictly upper triangular structure reveals the GLLAMM does not allow latent variables to be regressed on lower level latent or observed variables, implying the method deals with reflective measurement[1][4]. For a more detailed explanation on the topic see Edwards and Bagozzi [11].

However, notice that because in the IRT framework $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ should be orthogonal to each other by design, we can further decompose equation (2.8) in the following form:

$$\underset{(K\times 1)}{\boldsymbol{\eta}} = \underset{(K\times K)}{\boldsymbol{\Psi}_\eta}\ \underset{(K\times 1)}{\boldsymbol{\eta}} + \underset{(K\times Q)}{\boldsymbol{\Gamma}_\eta}\ \underset{(Q\times 1)}{\mathbf{W}_\eta} + \underset{(K\times 1)}{\boldsymbol{\zeta}_\eta} \tag{2.9}$$

$$\underset{(D\times 1)}{\boldsymbol{\theta}} = \underset{(D\times D)}{\boldsymbol{\Psi}_\theta}\ \underset{(D\times 1)}{\boldsymbol{\theta}} + \underset{(D\times Q)}{\boldsymbol{\Gamma}_\theta}\ \underset{(Q\times 1)}{\mathbf{W}_\theta} + \underset{(D\times 1)}{\boldsymbol{\zeta}_\theta} \tag{2.10}$$

where $\boldsymbol{\Psi}_\eta$, $\boldsymbol{\Psi}_\theta$, $\boldsymbol{\Gamma}_\eta$, and $\boldsymbol{\Gamma}_\theta$ are the dimension-specific parameter matrices that map the relationship between the corresponding latent variables $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$; with $\boldsymbol{\zeta}_\eta$ and $\boldsymbol{\zeta}_\theta$ being the dimension-specific vector of disturbances. Finally, the matrices $\mathbf{W}_\eta$ and $\mathbf{W}_\theta$ would be the dimension-specific covariates.

Considering figure 2.1 as reference, we did not specified any structural relationships $\boldsymbol{\Psi}$; nor declare any additional covariates $\mathbf{W}$. However, chapter 4 and 5 will show an implementation where we hypothesized a set of covariates explain the variability on some of the latent variables.

2.3 Model assumptions

Following Skrondal and Rabe-Hesketh [62], the framework has two main assumptions:

(M1) Complete latent space. The latent space is considered complete if all the latent variables, that we hypothesize affect the outcomes, are considered in the model [23]. In the GLLAMM representation, these would be all latent variables $\boldsymbol{\Theta}$ at levels $l > 1$ and $m > 1$.

[1] A latent variable is considered reflective when it is thought to be the cause of the lower level latent or manifest variables.

(M2) Local Independence, also known as conditional independence, is defined in the following form:

$$f\left(\mathbf{y}=\mathbf{1}\mid\mathbf{X},\mathbf{W},\mathbf{\Omega}\right)=\prod_{j=1}^{J}\prod_{d=1}^{D}\prod_{k=1}^{K}f\left(y_{jkd}=1\mid\mathbf{X},\mathbf{W},\mathbf{\Omega}\right) \tag{2.11}$$

where $f\left(\mathbf{y}=\mathbf{1}\mid\mathbf{X},\mathbf{W},\mathbf{\Omega}\right)$ denotes the level-1 conditional likelihood, and $\mathbf{y}$ the vector of all responses at the first level. It is important to mention that equation (2.11) results from the union of two parts [58, 3, 24]:

(a) **Local item independence,** which entails the individual's response to an item does not affect the probability of endorsing another item, after conditioning on the individual's ability. This is expressed in the following mathematical form:

$$f\left(y_{j..}=1\mid\mathbf{X},\mathbf{W},\mathbf{\Omega}\right)=\prod_{d=1}^{D}\prod_{k=1}^{K}f\left(y_{jkd}=1\mid\mathbf{X},\mathbf{W},\mathbf{\Omega}\right) \tag{2.12}$$

where $y_{j..}=[y_{j11},\ldots,y_{jKD}]$ is the vector of all items for individual j and, as previously defined, $K=\sum_m K_m$ and $D=\sum_l D_l$.

(b) **Local individual independence,** which entails that an individual's response to an item is independent of another person's response to that same item. The assumption is expressed in the following form:

$$f\left(y_{.kd}=1\mid\mathbf{X},\mathbf{W},\mathbf{\Omega}\right)=\prod_{j=1}^{J}f\left(y_{jkd}=1\mid\mathbf{X},\mathbf{W},\mathbf{\Omega}\right) \tag{2.13}$$

where $y_{.kd}=[y_{1kd},\ldots,y_{Jkd}]$ is the vector of all individuals endorsing item k from dimension d.

Chapter 3

Bayesian estimation

The practical use the GLLAMM developed in chapter 2 requires the estimation of all of the items and individuals' dimensions, loadings, regression and structural parameters. These can be obtained within two frameworks: the Frequentist and Bayesian.

The current chapter center its attention on describing the Bayesian estimation procedure, using the Markov Chain Monte Carlo method (MCMC). For a full development of GLLAMM under the Frequentist framework refer to Rabe-Hesketh and colleagues [51, 53, 62, 54].

3.1 Benefits and shortcomings

3.1.1 Why Bayesian?

The reasons on why Bayesian statistics is attractive to perform the parameters' estimation of any model, and especially for the GLLAMM developed in chapter 2, are:

1. It is built on a simulation-based estimation method, therefore, it can handle all kinds of priors and data-generating processes [13]. This is especially useful with highly complex and over-parameterized models, where other methods are unfeasible or work poorly [2, 30].

2. While the likelihood functions are used to define the posterior sampling distributions, they can also be used in a generative way. The likelihood for the data, and priors for the parameters, form the basis to produce samples from the posterior distribution. However, they can also be used to simulate observations, allowing us to test the ability of the method/data to recover the parameters of interest [38].

3. The Bayesian estimates are at least as good as its Frequentist counterparts [2, 68, 27]. This is true when the method uses uninformative 'flat' priors. However, because the procedure allow us to integrate prior knowledge about the parameters, beyond the observed responses, it can produce results even in scenarios where the Maximum Likelihood methods (ML) have issues of non-convergence or improper estimation [62, 13, 38]. Examples of such are:

   (a) when we have small sample sizes.

(b) when individuals have null scores or aberrant response patterns [24, 1]. The latter happens when examinees answered some relatively difficult and discriminating items correctly, while answering some of the easiest incorrectly.

(c) when parameters need to be confined to a permitted space, e.g. the estimation of positive unique factors variances [36].

(d) when we need to estimate parameters under sparse data, where the asymptotic theory is unlikely to hold [13];

3.1.2 Are you sure there's nothing wrong?

Off course the Bayesian framework has shortcomings, among them:

1. It exposes the user to arbitrary decisions about the running of the chains, in order to ensure a proper performance, e.g. how many iterates does the chain need to achieve precise estimates?, what is the right size for the burn-in and warm-up phases?, how should the thinning procedure be performed, if any?, should we follow the same procedure for all parameters of interest?, among others [62].

2. The user has multiple options to evaluate the performance of the method, and most of them are visual, making it hard to assess if a proper posterior investigation have been made [20]. More specifically, the user has multiple options to assess if the chain achieves the three requirements of good performance: stationarity, convergence, and good mixing [38].

3. The procedure makes it hard to discover parameters' lack of identification [62]. Inadequate mixing of the chain could lead us to think unidentified parameters have been estimated with precision, when in fact what we have are 'flat' posteriors [29].

4. Oftentimes the posterior sampling geometry of the model makes it hard to find proper solutions for the parameter space, no matter the rotation/rescaling of the parameter, or the amount of data [5]. This is especially true in complex hierarchical models.

5. The greater the complexity of the model, the harder it is to communicate/share the implementation with other scientist. This is especially true, when researcher re-parameterize the model to solve the previous shortcoming [38].

6. The procedure usually requires more time to achieve a proper solution, compared to the classical methods. This is especially true in models with high complexity [66, 59].

Although some of the previous shortcomings have made the Bayesian procedure a "controversial" implementation, most of them already have acceptable solutions.

For the first point, a popular approach is to use a large number of iterates, or multiple chains with different initial states. This is mostly applicable under the Metropolis-Hastings and Gibbs sampling methods. However, recent solutions like the Hamiltonian Monte Carlo method (HMC) [5] is less reliant on these decisions, as it implements a different sampling mechanism (see section 3.4).

About the second point, it is well accepted that the visual assessment of stationarity and convergence is easier, and this procedure usually has additional support from statistics like `Rhat` [19]. On the contrary, a visual evaluation of 'good' mixing remains as a hard task. A recent approach to increase the possibility of a well mixed chain is to change the posterior sampling geometry of the model [44, 45, 5, 38] (see section 3.5).

On the third point, the most common solution is to use regularizing priors, i.e. priors that are more 'skeptical' of wider parameter spaces [38]. However, it is important to mention, there are scenarios where one achieves poor parameter estimates, even in the presence of 'enough' data and regularizing priors, but this shortcoming is also applicable to the classical estimation procedures, e.g. the estimation of the variance parameters in random effects models [62].

About the fourth point, as it was mentioned in previous paragraphs, a recent approach to solve this issue is to change the posterior sampling geometry of the model. This means to re-parameterize the model in a way that removes the dependence of the parameters on other sampled parameters, or even, produce a sample mechanism that is located in a continuous between a centered (CP) and non-centered parametrization (NCP), e.g. Interleaved HMC (iHMC) or Variational Inference (VI) [17, 18, 44, 45, 5, 22] (see section 3.5).

Finally, the fifth and sixth points can be considered the 'price' a scientist has to pay, to be able to fit models that conforms better with the observed data generating processes. Although, recent developments are striving to improve upon reducing its required time.

3.2 Bayesian GLLAMM for dichotomous outcomes

3.2.1 Posterior distribution

Denoting $\mathbf{Y}$ as the observed data and $\mathbf{\Omega} = \{\boldsymbol{\beta}, \mathbf{\Lambda}, \mathbf{\Theta}, \mathbf{\Psi}, \mathbf{\Gamma}\}$, i.e. all the parameters declared in section 2.2.1 and 2.2.2; the posterior distribution is obtained using the Bayes theorem, in the following way:

$$P(\mathbf{\Omega} \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \mathbf{\Omega})\ P(\mathbf{\Omega})}{\int P(\mathbf{Y} \mid \mathbf{\Omega})\ P(\mathbf{\Omega})\ d\mathbf{\Omega}} \tag{3.1}$$

this is possible since the Bayesian approach makes no distinction between latent variables and parameters. All of them are considered random quantities [62]. Furthermore, since inference only requires representing the likelihood of the data $P(\mathbf{Y} \mid \mathbf{\Omega})$ and the prior distribution $P(\mathbf{\Omega})$; and because the denominator is just a (hard to calculate) constant, the posterior can be represented in the following form, without loss of generality:

$$P(\mathbf{\Omega} \mid \mathbf{Y}) \propto P(\mathbf{Y} \mid \mathbf{\Omega})\ P(\mathbf{\Omega}) \tag{3.2}$$

3.2.2 Prior distributions

Similar to Patz and Junker [46], we use an independent distributional structure for the joint priors of the parameters, at the highest level of the GLLAMM:

$$\begin{aligned} P(\mathbf{\Omega}) &= P(\boldsymbol{\beta}, \mathbf{\Lambda}, \mathbf{\Theta}, \mathbf{\Psi}, \mathbf{\Gamma}) \\ &= P(\boldsymbol{\beta})\ P(\mathbf{\Lambda})\ P(\mathbf{\Theta})\ P(\mathbf{\Psi})\ P(\mathbf{\Gamma}) \\ &= P(\boldsymbol{\beta})\ [P(\boldsymbol{\alpha})\ P(\boldsymbol{\lambda})]\ [P(\boldsymbol{\eta})\ P(\boldsymbol{\theta})]\ [P(\mathbf{\Psi}_\eta)\ P(\mathbf{\Psi}_\theta)]\ [P(\mathbf{\Gamma}_\eta)\ P(\mathbf{\Gamma}_\theta)] \end{aligned} \tag{3.3}$$

However, giving the hierarchical and cross-classified structure of the model, the lower level priors will also depend on further parameters[1], e.g. variances and covariances of the latent variables. These are known as *hyperparameters*, and their prior distributions as *hyperpriors*. Because of the previous, we are not going to detail their lower level representation, as they are highly dependent on the specifics of the model.

Finally, as recommended by several authors the priors will be mildly informative [38, 16, 66, 28, 1, 68], model-specific and determined by a prior predictive investigation.

Chapter 4 and 5 will show the process of prior elicitation in the context of a simulated and real dataset, respectively.

3.2.3 Likelihood

Following Rabe-Hesketh et al. [51], the likelihood function is build in a recursive way. First, we replace the structural model (2.8) into the linear predictor (2.6):

$$v_{jkd} = \mathbf{X}_j\boldsymbol{\beta} + (\mathbf{I} - \boldsymbol{\Psi})^{-1}\left[\boldsymbol{\Gamma}\mathbf{W} + \boldsymbol{\zeta}\right]\ \boldsymbol{\Lambda}\,\mathbf{H}_j \tag{3.4}$$

Second, the linear predictor and inverse-link function (2.3) are used to construct the systematic part of the response (2.2), i.e. its expected value. Here we use a logit link for pedagogical purposes:

$$\pi_{jkd} = \frac{\exp(\tau_k + v_{jkd})}{1 + \exp(\tau_k + v_{jkd})} = \frac{\exp(v_{jkd})}{1 + \exp(v_{jkd})} \tag{3.5}$$

Third, the expected value is used in the distributional part (2.1), in the following form:

$$f\left(y_{jkd} = 1 \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right) = \pi_{jkd}^{n}(1 - \pi_{jkd})^{1-n} \tag{3.6}$$

Fourth, considering assumptions (M1) and (M2) described in section 2.3, we produce the level-1 likelihood:

$$f\left(\mathbf{y} = \mathbf{1} \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right) = \prod_{j=1}^{J}\prod_{d=1}^{D}\prod_{k=1}^{K} f\left(y_{jkd} = 1 \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right) \tag{3.7}$$

Fifth, we define the marginal likelihood at levels l and m, conditional on the latent variables at levels $l+1$ and $m+1$, in the following form:

$$f_{(m)}^{(l)}\left(\mathbf{y} = \mathbf{1} \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right) = \int \left[\prod f_{(m-1)}^{(l-1)}\left(\mathbf{y} = \mathbf{1} \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right)\right]\ P(\boldsymbol{\Theta}_{(m)}^{(l)})\, d\boldsymbol{\Theta}_{(m)}^{(l)} \tag{3.8}$$

where the product inside the brackets is over all units in level $(l-1)$ and $(m-1)$, the first level is defined in equation (3.7), and $P(\boldsymbol{\Theta}_{(m)}^{(l)})$ are the prior distributions for the latent variables at level l and m, with $\boldsymbol{\Theta}_{(m)}^{(l)} = [\boldsymbol{\eta}^{(m)}, \boldsymbol{\theta}^{(l)}]$. Finally, we define the likelihood as:

$$\mathcal{L}(\mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}) = \prod_{m=2}^{M+1}\prod_{l=2}^{L+1} f_{(m)}^{(l)}\left(\mathbf{y} = \mathbf{1} \mid \mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}\right) \tag{3.9}$$

and the log-likelihood as:

$$\ell(\mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}) = \log \mathcal{L}(\mathbf{X}, \mathbf{W}, \boldsymbol{\Omega}) \tag{3.10}$$

[1]Because of this sequential model specification, it is said the priors also have a "hierarchical" structure.

3.3 Model identification

As indicated by Rabe-Hesketh et al. [51], in order to fully specify the model and provide a scale for the latent variables, we have to make assumptions for either the distribution of the disturbances $\boldsymbol{\zeta}$, or the distribution of one or more of the latent variables $\boldsymbol{\Theta}$. On the other hand, as point out by Fujimoto [14] and several others authors, we could also set restrictions for one or more of the loadings in $\boldsymbol{\Lambda}$. The last is more prevalent in frequentist software packages like Winsteps [32].

Furthermore, as it is hinted by equation (3.9), it is assumed the latent variables at different levels are independent from each other, whereas latent variables at the same level may present dependency. In case of the latter, we presume the latent variables at the same level have a multivariate normal distribution with a mean and covariance structure determined by equations (2.9) and (2.10). In the case of the covariance matrices, these are determined by the covariance of the specific disturbances $\boldsymbol{\zeta}$.

Chapter 4 and 5 will show the model identification strategy in the context of a simulated and real dataset, respectively.

3.4 Computational implementation

3.4.1 Hamiltonian Monte Carlo

For years the state-of-the-art platform for Bayesian statistical modeling have been the BUGS[2] project [35, 34], with its WinBUGS and OpenBUGS implementations. However, a more recent participant JAGS[3] [47] has gain traction, due to its cross-platform capabilities. In any case, both platforms are not that different form each other. Both use two of the most popular and successful algorithms for performing MCMC: the Metropolis-Hastings [39, 25] and Gibbs Sampling [21] algorithms.

In general lines, the Metropolis algorithm follows a two-step procedure: (i) it produces a new proposal for a parameter's value, and (ii) it evaluates the proposal against the current, accepting the new value if it is more likely under the posterior distribution. On the other hand, the Gibbs sampling uses a similar procedure, but it gains efficiency by exploiting the knowledge behind the target distribution. The latter improvement catapulted the Gibbs sampler into the workhorse of high dimensional Bayesian computing.

However, both methods remain as highly random procedures, and because of it, they still have important limitations. The most important of them all is that they do not yield independent and identically distributed samples (iid). This in turn means, the exploration of the posterior is made through a series of (sometimes highly dependent) values, causing the MCMC chain to converge to the target distribution only in the long run, i.e. when the number of iterations approach infinity ($t \to \infty$) [19].

In this setting two solutions have been proposed within these algorithms, and oftentimes the aforementioned software packages use them in conjunction. The first consist on improving the sampling mechanism by using a Hybrid MCMC method. A good example is the "Metropolis-within-Gibbs" algorithm [40]. On the other hand, the second consist

[2]Bayesian inference Using Gibbs Sampling
[3]Just Another Gibbs Sampling

on defining the number of iterates, burn-in and thinning processes in a way that it reduces the serial dependencies in the samples.

Nevertheless, the aforementioned solutions, while useful, do not solve all issues related to the MCMC chains performance, as these can remain biased in subtle ways, that are more harder to identify [38]. Therefore, researcher have decided instead to propose new improved algorithms.

Is in this context were the Hamiltonian Monte Carlo or Hybrid Monte Carlo (HMC) was proposed by Duane et al. [10], using the theory of Hamiltonian Dynamics. While the full representation of method is out of the scope of this research, we can still make a high-level description of its process, with purpose of point out its benefits and shortcomings.

Assuming we have continuous distributions and the partial derivatives of the log-density function (the gradient) exists, the method iterates the updating of two vector spaces of a "particle" [42, 38]. First, it samples new values for the momentum vector of the particle, independent of the current values of its position vector, i.e. it gives the particle a random "flick". In the second iteration, using Hamiltonian dynamics, a Metropolis update is performed to propose a new position vector for the particle. In this sense, one can say the HMC runs a small physics simulation, where the position of the particle denotes the current parameter values in the Markov chain, the momentum denotes the proposed path to a new set of values, and the log-posterior provides the "friction-less" surface over which the particle moves, where the gradient defines its curvature [38].

In principle, because HMC produce "more informed" proposals, it will accept all of them. In practice however, HMC uses a rejection criterion, that inform us when the method was not able to maintain the "balance" of the system, producing a bad numeric approximation (see section 3.5 for an example). For more on the background theory on Hamiltonian Dynamics and specifics about the HMC algorithm refer to Neal [42] and Betancourt and Girolami [5].

All of this just tells us that HMC is a highly complex algorithm. However, as McElreath [38] point out, the Gibbs strategy got its improvement over the Metropolis-Hastings by being less random, not more. And in that sense, HMC pushes this principle to a greater length, but manages to improve the efficiency of the MCMC procedure. This is especially true in cases where ordinary Metropolis or Gibbs sampling cannot make a proper exploration of the parameter space, e.g. with highly correlated parameters or with models with hundreds or thousands of parameters, like complex hierarchical models. Moreover, as HCM is more efficient, it requires less number of iterations to make a proper posterior investigation [38, 19]

Off course, this comes with a cost. The methods is more computationally intensive than the previous. However, because of the efficiency improvement and the need of less number of iterations, the method requires less computer time in total, even when each individual sample needs more.

3.4.2 Where can I find this magical software?

`Stan` [65] is the software package that will provide us with the machinery of HMC. Furthermore, the results produced from the software will be analyzed with `R` [48, 64], and its integration packages.

3.4.3 What about burn-in and thinning?

Since HMC uses a different approach and produces more informative proposals, setting specific burn-in and thinning processes is no longer necessary.

However, the method does require to perform a warm-up procedure to adapt sampling, and to tune in two parameters of the algorithm: the number of steps, and the step size [65]. These parameters are used to make a discrete approximation of the momentum vector, i.e. the path of the particle [42, 5].

Notice, this phase is similar to a procedure performed in JAGS, where the proposal distribution gets tuned up to improve the posterior investigation. However, the similarities end there. The warm-up samples are not representative of the target posterior distribution, no matter how long the iterations continue [38]. Finally, after the warm-up is finished, and assuming the adaption was successful, the method will sample directly from the target distribution.

In the current research, we will use the default values set in the software, i.e. a total of $2,000$ iterations, where $1,000$ of them will be spend on warm-up. Notice this departs (by far) from most of the literature on Bayesian IRT models, e.g. Fujimoto [15] used chains with $60,000$ iterations, where $15,000$ were discarded and the remaining were thinned in jumps of 3; while Fujimoto [14] used $225,000$ iterations, with burn-in of $30,000$ and thinning with jumps of 15; just to mention a few.

3.4.4 Initial starts

Since the method requires starting values for the parameters, the author decided to allow the software to sample these from the priors defined in the model.

3.5 To center or not to center

As it was point out by Betancourt and Girolami [5], even the most simple hierarchical models present formidable pathologies, that no simple correction can be performed to visit the posterior distribution properly. This is true no matter the rotation/rescaling of the parameter, or the amount of data.

A great example of this is what McElreath [38] dubbed the Devil's funnel, where the author shows that you do not need a complex model to start observing these issues. Consider the following joint distribution:

$$\begin{aligned} v &\sim N(0, 3) \\ \theta &\sim N(0, \exp(v)) \end{aligned} \tag{3.11}$$

where v can be interpreted as the *hyperprior* of θ, and θ as dependent on the samples of v. The latter is what the literature dubbed as the centered parametrization (CP).

This joint distribution might seem familiar, as Bayesian hierarchical model practitioners often use it to ensure the standard deviation remains into the permitted parameter space ($\sigma \geq 0$). Similar types of requirements can be observed in IRT models. Moreover, it seems that any MCMC procedure would not have any issues exploring the posterior distribution of these parameters, as they are normally distributed and there are only two of them, but we would be wrong. The reader can find the `stan` and `jags` implementations in Appendix C.1.1.

Figure 3.1 show the chains resulting from implementing the model on equation (3.11), through HMC (`stan`). For pedagogical purposes, the example was run without data, as the author wanted to emphasize the pathologies are present even before we feed the data to our model. However, its easy to extrapolate that these issues will remain present when data is available.

As one can see from the figure, the joint posterior distribution is not explored properly. The chains show no sign of being "healthy", i.e. they do not show stationarity or convergence (top panels), nor a good mixing (middle and bottom panels). This is further supported by the `Rhat` [19], and effective sample sizes estimates `n_eff` [19]. The `Rhat` for all parameters were above one (recommended threshold), indicating the chains did not achieve convergence: the between-chain variability was larger than the within-chain variability. Furthermore, the effective samples sizes were 34 and 293 for v and θ, respectively; indicating the values of the chains remained highly correlated. This is striking as one could expect effective sample sizes closer to the actual number of iterations $3,000$ (3 chains with $1,000$ samples each, after warm-up).

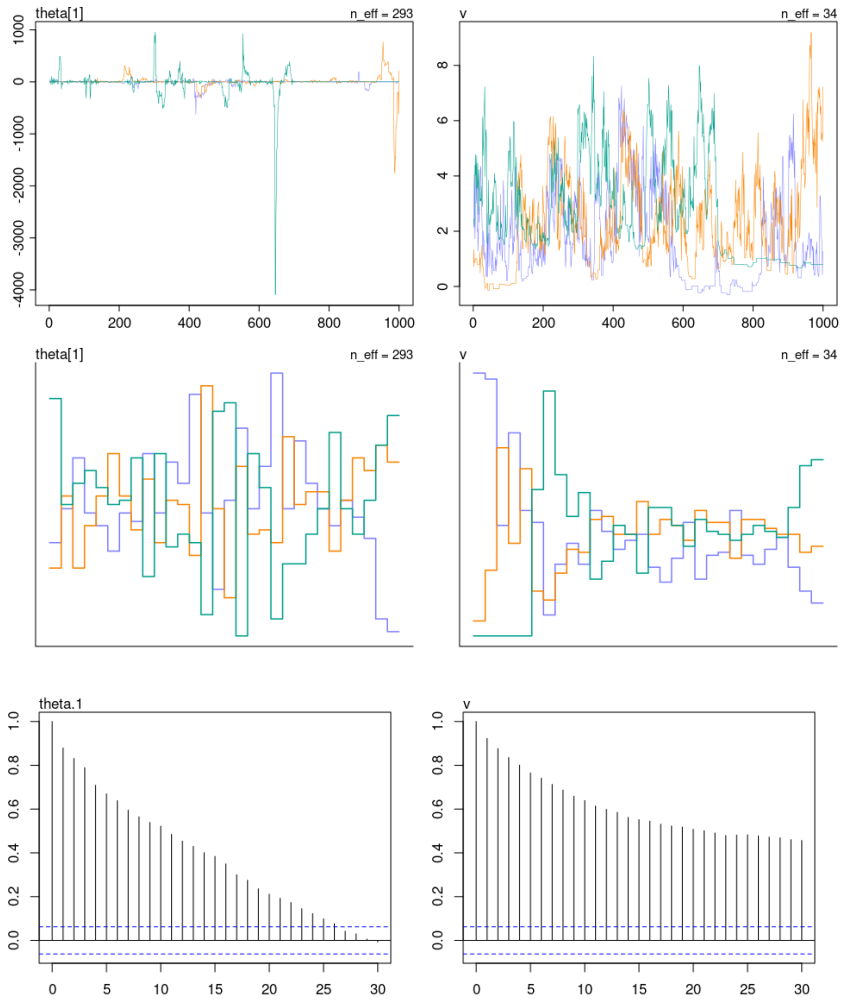

Figure 3.1: The Devil's funnel. Centered Parametrization. (Top) It shows the traceplot for the iterations on three chains. (Middle) It shows the trankplot for the same data. (Bottom) It shows the Auto-correlation Functions (ACF) of the iterations.

3.5.1 Wasn't HMC the solution to this?

HMC is a powerful MCMC algorithm, and its benefits stand the test of other scenarios, as detailed by multiple authors [38, 19]. However, the issue with this example goes a little bit farther than what the algorithm can actually overcome, and this is true also for the Metropolis and Gibbs sampler. Nevertheless, we will see HMC still manages to be efficient within the constrains of the problem.

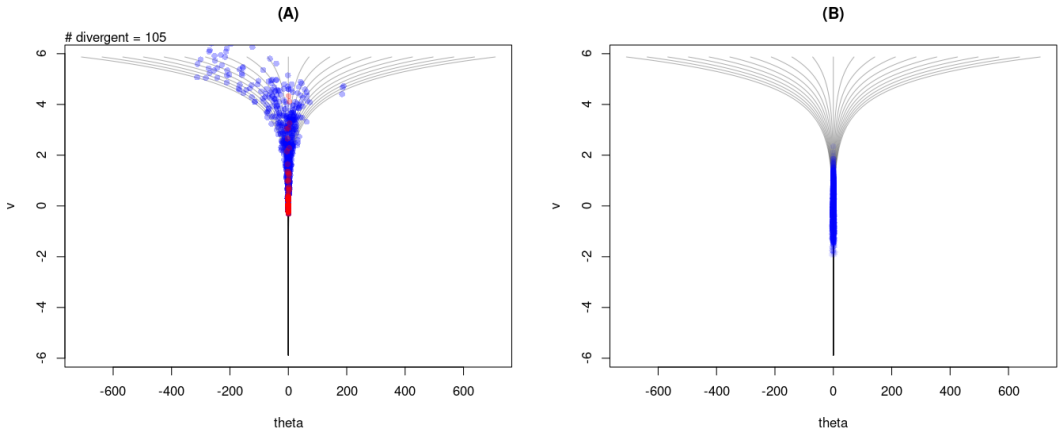

Figure 3.2: Posterior sampling geometry. Centered Parametrization. (A) HMC exploration, blue points are accepted samples, red points denote divergent transitions. (B) Metropolis/Gibbs exploration, blue points are samples.

Figure 3.2 shows the posterior sampling geometry of the model in equation (3.11), for the HMC algorithm (`stan`) and Metropolis/Gibbs sampler (`jags`).

The first thing to notice is the complexity of the joint posterior geometry, with a steep funnel shape as the values of v gets smaller and smaller. This makes sense, as the exponential of larger negative values, force the standard deviation of θ to be narrower and narrower around its mean (zero).

The second thing to notice is that HMC still manages to successfully explore the steep parameter space of v and θ, albeit some parts are less visited ($\theta > 0$). This does not happen under the Metropolis-Hastings or Gibbs sampler algorithms. In fact, `jags` did not managed to escape the narrow funnel shape, no matter the number of iterations used the adaptation, burn-in and sampling procedures[4] (panel B). Moreover, as shown in figure B.1 (appendix), the traceplots indicate the chains are "healthy", ignoring the clear lack of exploration of the posterior distribution.

Finally, although HMC did not fully explore the posterior, at least the method let us know which parts of the distribution did not manage to visit appropriately (see red points indicating divergent transitions). The last is important, because HMC gives us the opportunity to identify where can we improve our parametrization, something that is not available on Metropolis or Gibbs.

[4] `stan` used 3 chains with $1,000$ iterations for adaptation, and $1,000$ iterations for sampling. `jags` used 3 chains with $5,000$ iterations for adaptation, $5,000$ iterations for burn-in, and $5,000$ iterations for sampling.

3.5.2 So, how can we solve this?

There are two proposed solutions for this issue: (i) use regularizing priors, and (ii) change the posterior sampling geometry. This section will outline the benefits of the first, as it is already a well study result; while is going to advocate for the benefits of the second, in the context of IRT models.

Regularizing priors

It turns out that the benefits of using regularizing priors can also be explained from a geometrical perspective. Consider the following change to equation (3.11):

$$\begin{aligned} v &\sim N(0, 1) \\ \theta &\sim N(0, \exp(v)) \end{aligned} \tag{3.12}$$

notice we chose a mildly informative prior for v.

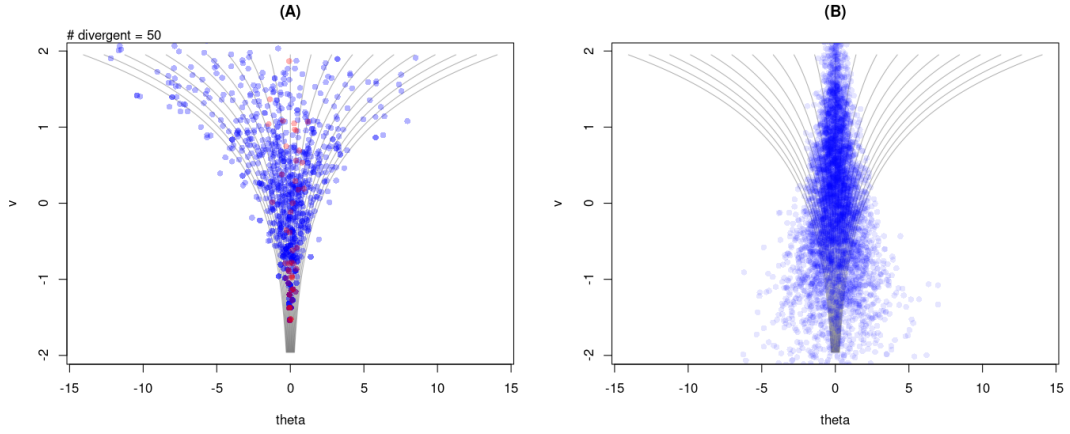

Figure 3.3: Posterior sampling geometry. Centered Parametrization with mildly informative priors. (A) HMC exploration, blue points are accepted samples, red points denote divergent transitions. (B) Metropolis/Gibbs exploration, blue points are samples.

Figure 3.3 shows the joint posterior sampling geometry of equation (3.12), for the HMC algorithm (`stan`) and Metropolis/Gibbs sampler (`jags`), respectively.

The first thing to notice from the figure is that the geometry available for exploration seems more broad. However, what is actually happening is that by adding information about v, we are "zooming" into the posterior geometry observed in figure 3.2, according to the range of the new priors. This further benefits the exploration of the posterior, as extreme ranges don't have to be visited from the start, e.g. the narrow funnel in the range $[-6, -2]$ in figure 3.2.

The second thing one can notice is that HMC does not "loose time" exploring sections of the posterior distribution that are not needed. Surprisingly in contrast, the Gibbs sampler explore non-useful sections and leaves the useful ones unexplored. Again, this is true no matter the number of iterations used the adaptation, burn-in and sampling procedures[5]. Moreover, the chains in figure B.2 (appendix) seem "healthy", ignoring the clear lack of exploration of the posterior, as in the previous example.

[5]the author used the same settings as in the previous example.

Third, we see that using regularizing priors reduced the number of `stan`'s divergent transitions from 150 (in the previous example) to 50. Although we still observe them in the steepest part of the geometry. Such information is not provided in `jags`.

Finally, we notice a mild improvement on the trace, tranks and ACF plots in figure 3.4 compared to figure 3.1, in the previous example.

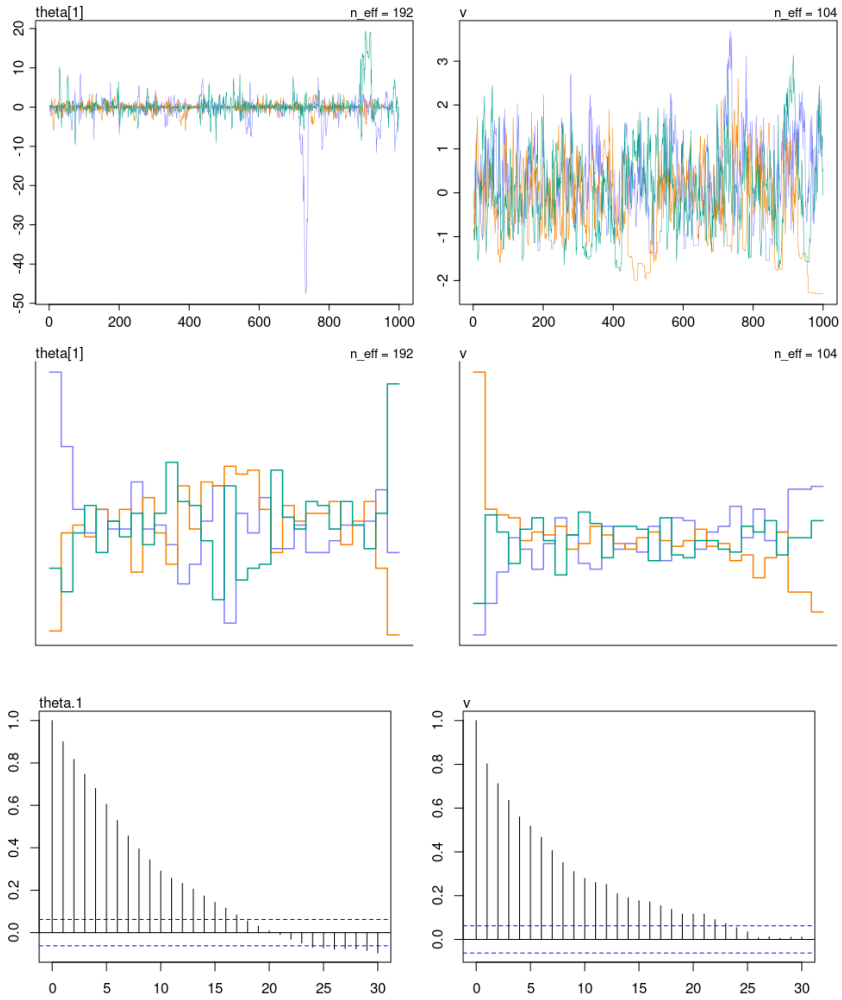

Figure 3.4: The Devil's funnel. Centered Parametrization with mildly informative priors. (Top) It shows the traceplot for the iterations on three chains. (Middle) It shows the trankplot for the same data. (Bottom) It shows the Auto-correlation Functions (ACF) of the iterations.

Non-Centered Parametrization

It would seem that, increasing the regularizing power of the priors is the answer to get rid of the sampling issues, raised by a complex posterior sampling geometry. However, there is a limit on the information a prior can contain, without imposing strong assumptions about the parameters in a model. Furthermore, sometimes researchers want the information on the data dominates the posterior parameter space. In those cases, imposing even mildly informative priors is a solution that is out of the picture.

In this context, Betancourt and Girolami [5] indicated that prior information can be included in the model, not only through the prior distributions, but also by encoding it in the model itself, i.e. changing the posterior sampling geometries, taking advantage of the hierarchical structure explicitly.

Under the Bayesian framework, a change in geometry consist on a re-parameterization of the model, that seeks to remove the dependence of the parameters on other sampled parameters, therefore favoring the performance of the MCMC chains. This is what the literature has dubbed as the non-centered parametrization (NCP).

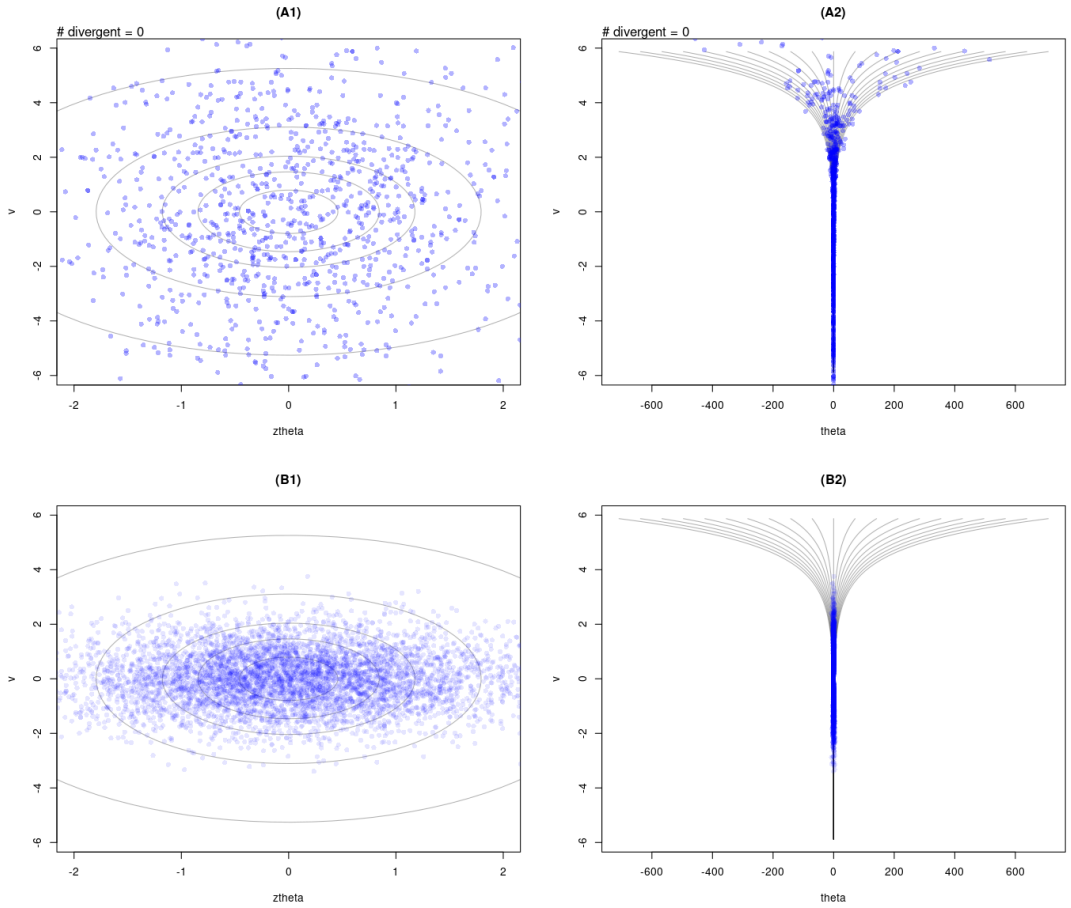

Figure 3.5: Posterior sampling geometry. Non-Centered Parametrization. (A1, B1) geometry defined by the independent parameters v and z. (A2, B2) Original sampling geometry defined by v and θ. (A1, A2) HMC algorithm (`stan`). (B1, B2) Metropolis/Gibbs sampler (`jags`). Blue points are accepted samples. Red points denote divergent transitions.

In our case, changing the posterior sampling geometry means to modify equation (3.11) in the following way:

$$\begin{aligned} v &\sim N(0, 3) \\ z &\sim N(0, 1) \\ \theta &= \exp(v) * z \end{aligned} \tag{3.13}$$

notice v has the original assumed variability, but now θ is defined in a way, that is no

longer sample dependent on v, i.e. we no longer sample θ directly but z, and transform it back.

The motivation for this re-parametrization have seeds in the calculation of the standard score (z-sore). Notice that if $\theta \sim N(\mu_\theta, \sigma_\theta)$ where $\sigma_\theta = \exp(v)$, then $(\theta - \mu_\theta)/\sigma_\theta = z$ and $z \sim N(0, 1)$. Using this process in reverse, we notice $\theta = \mu_\theta + \sigma_\theta z = \exp(v)\ z$ when $\mu_\theta = 0$, then $\theta \sim N(0, \exp(v))$. As the previous, there are transformation that can be done for other distributions, and they can even be extended to the multivariate normal case, through the use of the Cholesky decomposition [38].

For our current application however, figure 3.5 show the posterior sampling geometry for the HMC and Gibbs sampler. Notice both algorithms manage to explore more successfully the posterior distribution, although HMC does it a bit better on the extremes of v. Moreover, HMC shows no divergent transitions, meaning the posterior is "visited" without issues; as seen on figure 3.6, where the chains show clear signs of stationarity, convergence and good mixing.

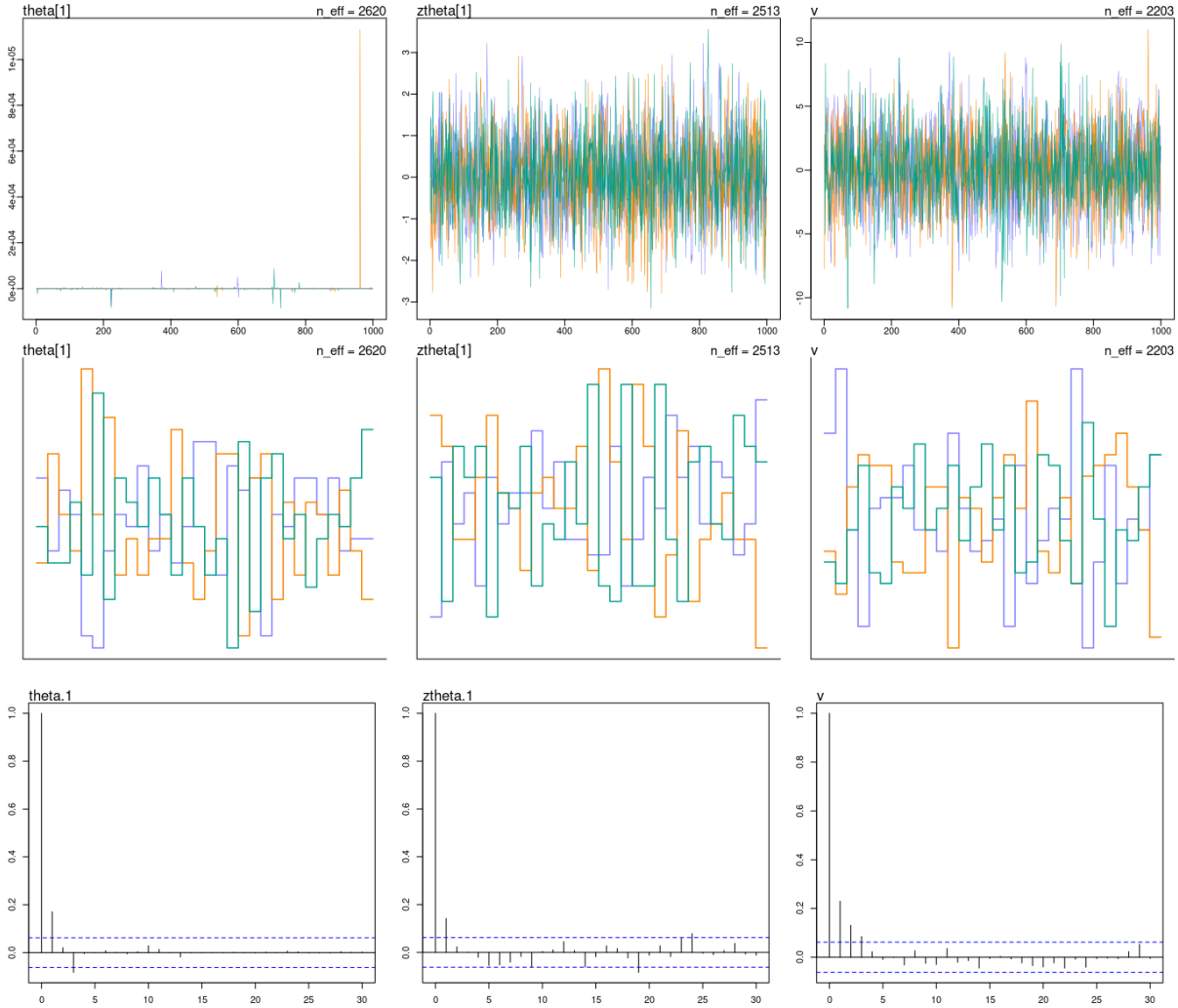

Figure 3.6: The Devil's funnel. Non-centered Parametrization. (Top) It shows the traceplot for the iterations on three chains. (Middle) It shows the trankplot for the same data. (Bottom) It shows the Auto-correlation Functions (ACF) of the iterations.

Finally, Papaspiliopoulos et al. [45] indicated that the success of the NCP strategy is largely dependent on the specifics of the model, i.e. neither strategy is more "optimal" than the other. The author pointed out that the natural CP showed better performance when conditional conjugacy was present, i.e. when the posterior distribution belonged to the same parametric family as the likelihood or prior distribution. On the other hand, the NCP worked as its complement, excelling when the previous requirement was not present. Therefore it makes sense that recent advancements on the topic are focused on producing sample mechanisms located in a continuous between CP and NCP, as greater performance improvements can be obtained by leveraging on the "best of both worlds", e.g. Interleaved HMC (iHMC) or Variational Inference (VI) [17, 18, 44, 45, 22].

Chapter 4 and 5 will show the application of non-centered parametrization in the context of an IRT model.

Chapter 4

Simulation Study

4.1 Simulated data

4.2 Models

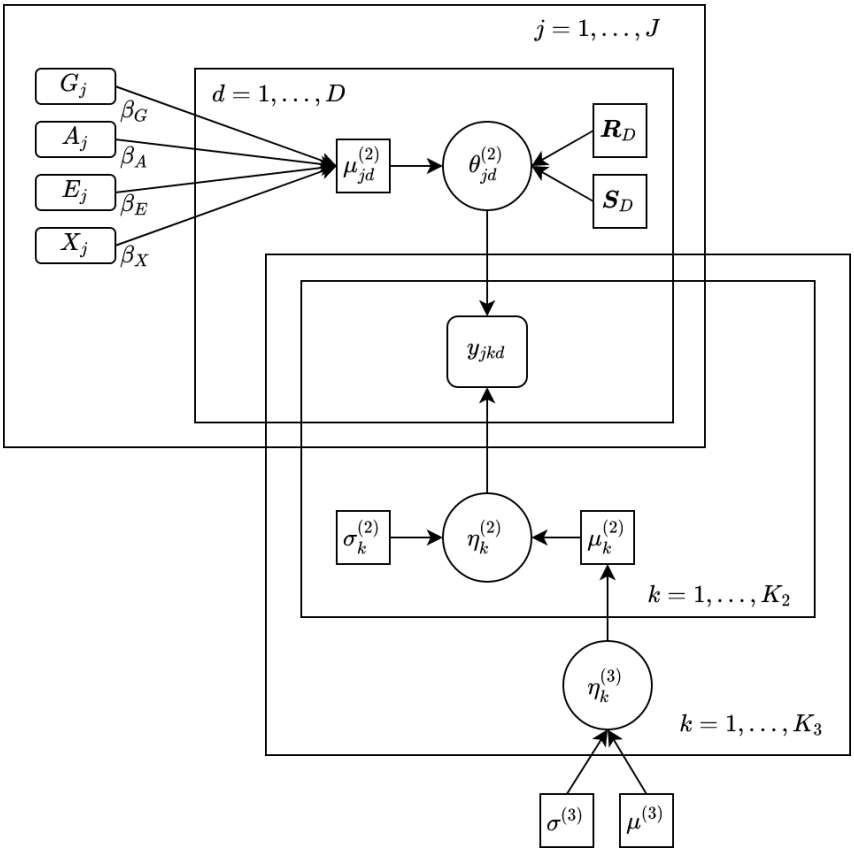

Figure 4.1: Directed Acyclig Graph (DAG). First Order Latent Variables model (FOLV). Circles represent latent variables. Squares represent parameters for priors. Large Squares represent nesting in specific units.

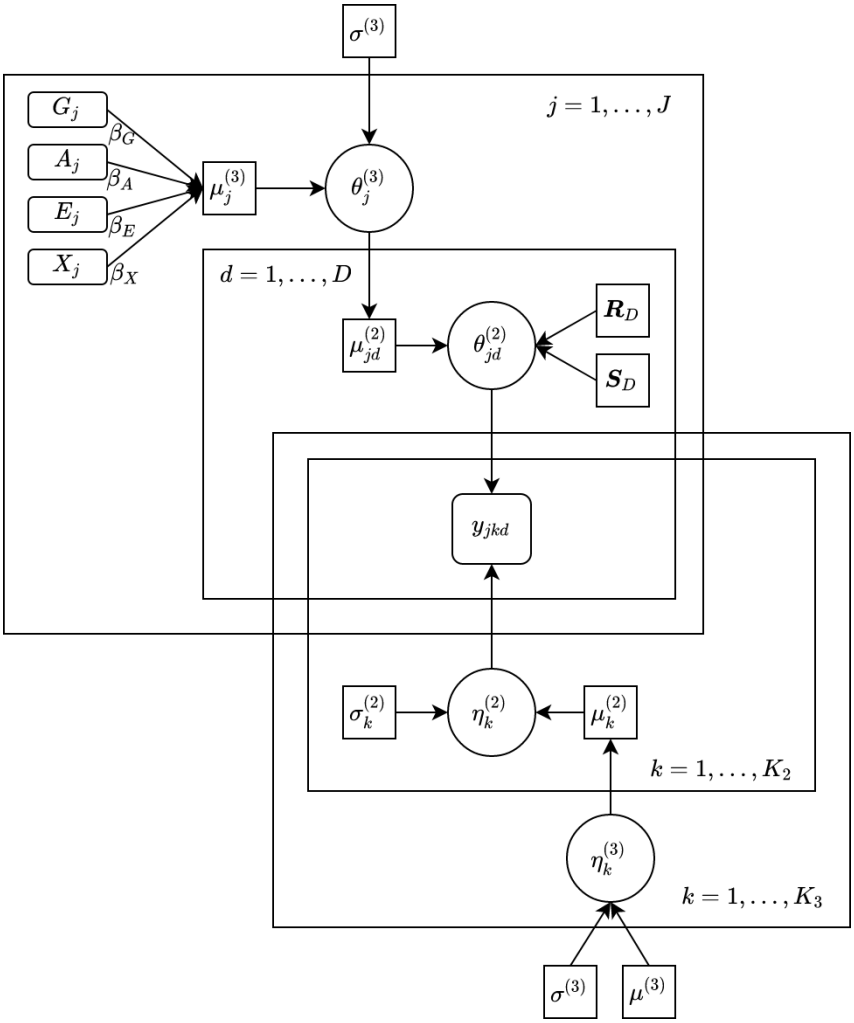

Figure 4.2: Directed Acyclig Graph (DAG). Second Order Latent Variables model (SOLV). Circles represent latent variables. Squares represent parameters for priors. Large Squares represent nesting in specific units.

4.3 Results

Chapter 5

Application

5.1 Instruments

5.2 Data

5.2.1 Collection

5.2.2 Sample scheme

5.3 Results

Chapter 6

Conclusion and Discussion

6.1 Discussion

6.2 Conclusions

6.3 Future development

Appendix A

Additional Theory

A.1 Other links and distributions

As stated in chapter 2, the GLLAMM is a unifying framework for a wide range of latent models. This is possible thanks to the flexibility in the modeling of the outcomes, coming from defining the response model as a GLM [37] (see section 2.2).

In the current section the author will try to briefly describe the most important link functions and outcomes distributions, that can be accommodated within the framework. For a more detailed approach on either of these, refer to Rabe-Hesketh et al. [51, 53, 52], Skrondal and Rabe-Hesketh [62], and Rabe-Hesketh et al. [54].

1. **Continuous:**
   It results form selecting an identity link function for the scaled mean response,

$$\begin{aligned} \mu^* &= E[y^*|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= v \end{aligned} \tag{A.1}$$

   where $\mu^* = \mu\sigma^{-1}$, $y^* = y\sigma^{-1}$, and σ denotes the standard deviation of the errors.

   On the other hand, the distributional part is defined by a Standard Normal distribution $\phi(x) = (2\pi)^{-1/2}exp(-x^2/2)$,

$$\begin{aligned} f(y^*|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}) &= \phi(\mu^*)\sigma^{-1} \\ &= \phi(v)\sigma^{-1} \end{aligned} \tag{A.2}$$

   Notice that the same parametrization can be achieved considering $y^* = v + \epsilon^*$, and $\epsilon^* \sim N(0, 1)$. Additionally, the decision to standardize the response variables has been made with the purpose of making the estimation process easier, as such distribution is free of unknown parameters.

2. **Polytomous:**
   It results from selecting a generalized logistic inverse-link function [6] for the expected value of the response, which in this case, describe the probability of endorsing

one of the S unordered available categories,

$$\begin{aligned} \mu_s &= E[y = y_s|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= P[y = y_s|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= \pi_s \\ &= h(v_s) \end{aligned} \tag{A.3}$$

where v_s is the linear predictor for category s ($s = 1, \ldots, S$), and $h(\cdot)$ is defined as:

$$h(x) = exp(x) \cdot \left[\sum_{s=1}^{S} exp(x)\right]^{-1} \tag{A.4}$$

It is important to note that, as in the dichotomous case, the same parametrization can be achieved using the concept of underlying continuous responses in the form $y_s^* = v_s + \epsilon_s$, where $y = s$ if $y_s^* > y_k^* \; \forall s, s \neq k$, ϵ_s have a Gumbel (extreme value type I) distribution, as the one defined in equation (2.3), and y_s denotes the random utility for the s category.

Finally, the distributional part is defined by a Multinomial distribution,

$$\begin{aligned} f[y = \{y_1, \cdots, y_S\}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] &= \frac{n!}{y_1! \cdots y_S!} \prod_{s=1}^{S} \mu_s^{y_s} \\ &= \frac{n!}{y_1! \cdots y_S!} \prod_{s=1}^{S} \pi_s^{y_s} \end{aligned} \tag{A.5}$$

where y_s denotes the number of "successes" in category s.

3. **Ordinal and discrete time duration:**
For the ordinal case, the linear predictor is "linked" to the probability of endorsing category s, against all previous categories, in the following form:

$$\begin{aligned} \mu_s &= E[y = y_s|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= P[y \leq y_s|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] - P[y \leq y_{s-1}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= h(\kappa_s - v_s) - h(\kappa_{s-1} - v_{s-1}) \end{aligned} \tag{A.6}$$

where κ_s denotes the thresholds for category s. For discrete time duration, the linear predictor is "linked" to the probability of survival, in the sth time interval, as follows:

$$\begin{aligned} \mu_s &= E[t_{s-1} \leq T \leq t_s|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= P[T \leq t_s|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] - P[T \leq t_{s-1}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= h(v_s + t_s) - h(v_{s-1} + t_{s-1}) \end{aligned} \tag{A.7}$$

where T is the unobserved continuous time, and t_s its observed discrete realization. Additionally, for both type of responses, $h(\cdot)$ can be defined as the logistic, standard normal, and Gumbel (extreme value type I) *cumulative distributions*, as in equation (2.3).

Similar to the dichotomous and polytomous case, the same parametrization can be achieved using the concept of underlying latent variables with $y_s^* = v_s + \epsilon_s$, where $y = s$ if $\kappa_{s-1} < y_s^* \leq \kappa_s$, $\kappa_0 = -\infty$, $\kappa_1 = 0$, $\kappa_S = +\infty$, ϵ_s has one of the distributions in equation (2.3), and y_s denotes the random utility for the s category.

It is important to note, for discrete time duration responses, the logit link corresponds to a *Proportional-Odds model*, while the complementary log-log link to a *Discrete Time Hazards model* [55]. Other models for ordinal responses, such as the *Baseline Category Logit* or the *Adjacent Category Logit* models can be specified as special cases of the generalized logistic response function, defined in equation (A.4).

Finally, the distributional part is defined by a Multinomial distribution, as the one defined in equation (A.5).

4. **Counts and continuous time duration:**
It results from selecting an exponential inverse-link function (log link) for the expected value of the response,

$$\begin{aligned} \mu &= E[y|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= \lambda \\ &= exp(v) \end{aligned} \tag{A.8}$$

and a Poisson conditional distribution for the counts,

$$\begin{aligned} f[y|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] &= exp(-\mu)\mu^y(y!)^{-1} \\ &= exp(-\lambda)\lambda^y(y!)^{-1} \end{aligned} \tag{A.9}$$

It is important to mention that unlike the models for dichotomous, polytomous and ordinal responses, model for counts cannot be written under the random utility framework.

5. **Rankings and pairwise comparisons:**
Following Skrondal and Rabe-Hesketh [60], the parametrization for polytomous responses can serve as the building block for the conditional distribution of rankings. Selecting a "exploded logit" inverse-link function [8] for the expected value of the response, which describes the probability of the full rankings of category s,

(work in progress)

$$\begin{aligned} \mu_s &= P[\mathbf{R}_s = \{r_s^1, \dots r_s^1\}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= \pi_s \\ &= h(v_s) \end{aligned} \tag{A.10}$$

where v_s is the linear predictor for category s ($s = 1, \dots, S$), and $h(\cdot)$ is defined as:

$$h(x) = \prod_{s=1}^{S} exp(x^s) \left[\sum_{s=1}^{S} exp(x^s) \right]^{-1} \tag{A.11}$$

Again, as in specific previous cases, the same parametrization can be achieved using the concept of underlying latent variables.

Finally, the distributional part is defined by a Multinomial distribution,

$$\begin{aligned} f[y = \{y_1, \cdots, y_S\}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] &= \frac{n!}{y_1! \cdots y_S!} \prod_{s=1}^{S} \mu_s^{y_s} \\ &= \frac{n!}{y_1! \cdots y_S!} \prod_{s=1}^{S} \pi_s^{y_s} \end{aligned} \tag{A.12}$$

where y_s denotes the number of "success cases" in category s.

6. **Mixtures:**
Given the previous definitions, the framework easily lends itself to model five additional settings:

   (a) **Different links and distributions for different latent variables**. This can be easily achieved by setting different links and distributions for each of the M_2 latent variables located at level 2.

   (b) **Left- or right-censored continuous responses**. Common in selection models (e.g. Heckman [26]), they can be achieved by specifying an identity link and Normal distribution for the uncensored scaled responses, as in equations (A.1) and (A.2); and a scaled probit link and Binomial distribution otherwise, as in equations (2.3) and (**??**).

   (c) **zero-inflated count responses**. where a log link and a Poisson distribution is set for the counts, as in equations (A.8) and (A.9); and a logit link and Binomial distribution is specified to model the zero center of mass, as in equations (**??**) and (**??**).

   (d) **Measurement error in covariates**. this setting occurs when standard models use variables, with measurement error, as covariates, e.g. a logistic regression with a continuous covariate that presents measurement error. For more details on this type of setting see Rabe-Hesketh, Skrondal and Pickles [50], Rabe-Hesketh, Pickles and Skrondal [49], and Skrondal and Rabe-Hesketh [61].

   (e) **Composite links**. Useful for specifying proportional odds models for right-censored responses, for handling missing categorical covariates and many other model types. For more details on this type of settings see Skrondal and Rabe-Hesketh [63].

Heteroscedasticity and over-dispersion in the response

Much like the Generalized Linear Mixed Model framework (GLMM), the GLLAMM allows to model heteroscedasticity, and over- or under-dispersion by adding random effects to the linear predictor, at level 1. The types of responses, in which such characteristics can be modeled, are the following:

1. **Continuous:**
We model **heteroscedasticity** in the following form:

$$\sigma = exp(\boldsymbol{\alpha}^T \mathbf{Z}^{(\mathbf{1})}) \tag{A.13}$$

Notice that the previous formula implies that equation (A.2) can be re-written in the following form:

$$f(y^*|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}) = \phi(v + \boldsymbol{\alpha}^T \mathbf{Z}^{(\mathbf{1})}) \tag{A.14}$$

where $\mathbf{Z}^{(\mathbf{1})}$ is the design matrix that maps the random effects $\boldsymbol{\alpha}$. Notice that equation (A.14) effectively corresponds to a model that includes random intercepts at level 1.

2. **Ordinal, and discrete time duration:**
Similar to the dichotomous case, by including random intercepts at level 1 in equation (A.6), we can model over- or under-dispersion:

$$\begin{aligned} \mu_s &= P[y \leq y_s|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] - P[y \leq y_{s-1}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= h(\kappa_s - v_s + \boldsymbol{\alpha}^T \mathbf{Z}^{(\mathbf{1})}) - h(\kappa_{s-1} - v_{s-1} + \boldsymbol{\alpha}^T \mathbf{Z}^{(\mathbf{1})}) \end{aligned} \tag{A.15}$$

A similar parametrization can be used for discrete time duration.

3. **Counts, and continuous time duration:**
Finally, modifying equation (A.8) allow us to model over- or under-dispersion under a counts model:

$$\begin{aligned} \mu &= E[y|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}] \\ &= \lambda \\ &= exp(v + \boldsymbol{\alpha}^T \mathbf{Z}^{(\mathbf{1})}) \end{aligned} \tag{A.16}$$

A.2 Sampling scheme

Appendix B

Additional plots

B.1 Chapter 3: Bayesian estimation

B.1.1 To center or not to center

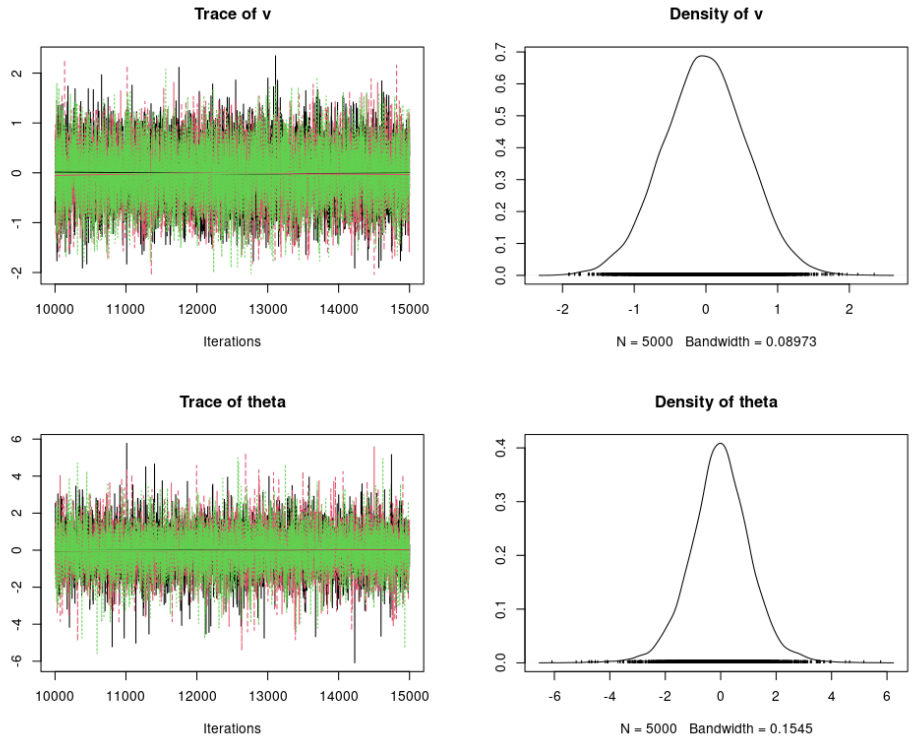

Figure B.1: The Devil's funnel. Centered Parametrization implemented in JAGS. It shows the traceplot and distribution of the parameters of interest.

B.2 Chapter 4: Simulation study

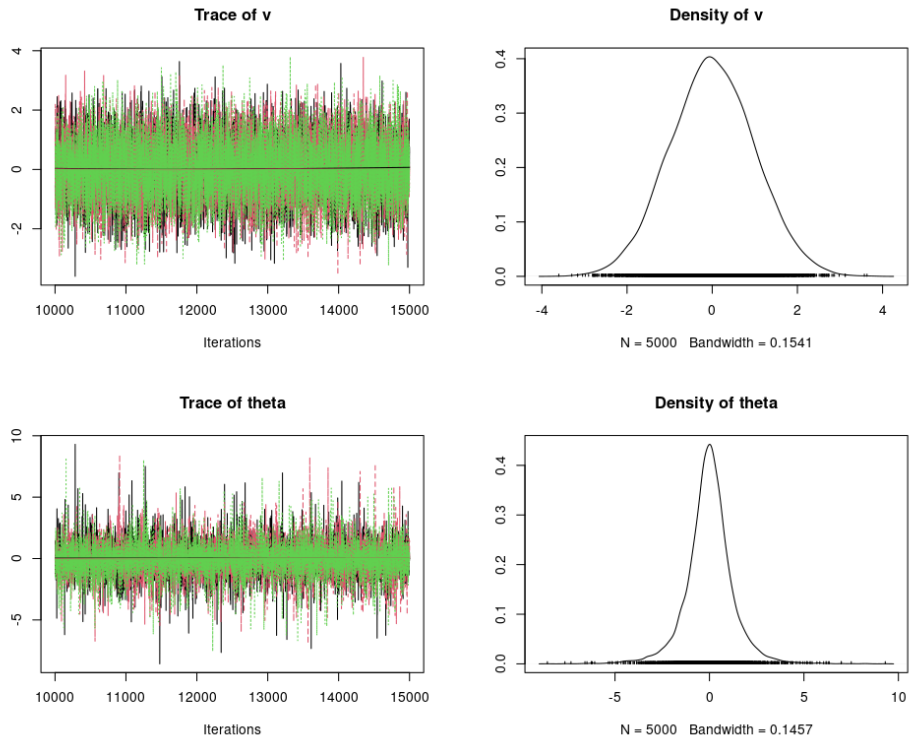

Figure B.2: The Devil's funnel. Centered Parametrization with mildly informative priors implemented in JAGS. It shows the traceplot and distribution of the parameters of interest.

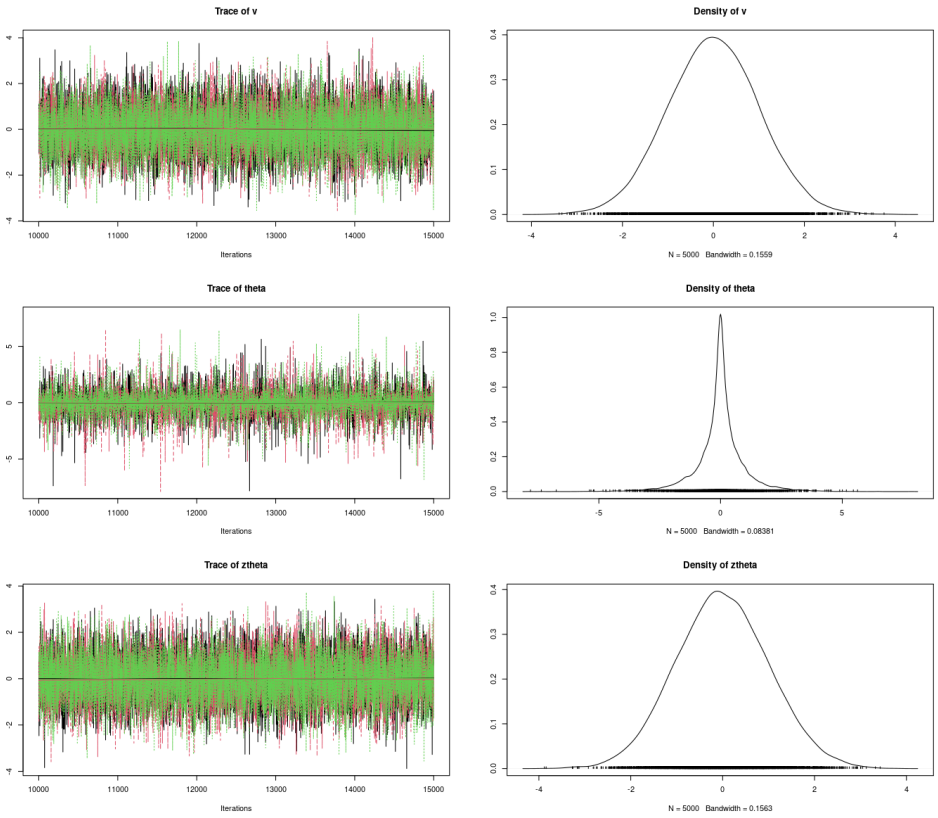

Figure B.3: The Devil's funnel. Non-Centered Parametrization implemented in JAGS. It shows the traceplot and distribution of the parameters of interest.

Appendix C

Code

C.1 Chapter 3: Bayesian estimation

C.1.1 To center or not to center

The devil's funnel, centered parametrization.

Stan

```
transformed data {
        int<lower=0> J;
        J = 1;
}
parameters {
        real theta[J];
        real v;
}
model {
        v ~ normal(0, 3);
        theta ~ normal(0, exp(v));
}
```

JAGS

```
model{
        v ~ dnorm(0,3)
        theta ~ dnorm(0, exp(v))
}
```

The devil's funnel, centered parametrization with priors.

Stan

```
transformed data {
        int<lower=0> J;
        J = 1;
```

```
}
parameters {
        real theta[J];
        real v;
}
model {
        v ~ normal(0, 1);
        theta ~ normal(0, exp(v));
}
```

JAGS

```
model{
        v ~ dnorm(0,1)
        theta ~ dnorm(0, exp(v))
}
```

The devil's funnel, non-centered parametrization.

Stan

```
transformed data {
        int<lower=0> J;
        J = 1;
}
parameters {
        real ztheta[J];
        real v;
}
transformed parameters{
        vector[J] theta;
        theta = exp(v) * to_vector( ztheta );
}
model {
        v ~ normal(0, 3);
        ztheta ~ normal(0, 1);
}
```

JAGS

```
model{
        v ~ dnorm(0,1)
        ztheta ~ dnorm(0,1)
        theta = v * ztheta
}
```

C.2 Chapter 4: Simulation study

C.2.1 Simulated data

C.2.2 Models

FOLV

SOLV

Bibliography

[1] Azevedo, C. [2003]. *Métodos de estimação na teoria de resposta ao item*, Master's thesis, Universidade de São Paulo (USP).
url: https://teses.usp.br/teses/disponiveis/45/45133/tde-05102004-163906/pt-br.php.

[2] Baker, F. [1998]. An investigation of the item parameter recovery characteristics of a gibbs sampling procedure, *Applied Psychological Measurement* **22**(22): 153–169.
doi: https://doi.org/10.1177/01466216980222005.

[3] Baker, F. [2001]. The basic of item response theory, *Technical report*, ERIC Clearinghouse on Assessment and Evaluation.

[4] Beaujean, A. [2014]. *Latent Variable Modeling Using R. A Step-by-Step Guide.*, Routledge.

[5] Betancourt, M. and Girolami, M. [2012]. Hamiltonian monte carlo for hierarchical models.
url: arxiv.org/abs/1312.0906v1.

[6] Bock, R. [1972]. Estimating item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika* **37**(1).
doi: https://doi.org/10.1007/BF02291411.

[7] Bradlow, E., Wainer, H. and Wang, X. [1999]. A bayesian random effects model for testlets, *Psychometrika* **64**(2): 153–168.
doi: https://doi.org/10.1007/BF02294533.

[8] Chapaaan, R. and Staelin, R. [1982]. Exploiting rank ordered choice set data within the stochastic utility model, *Journal of Marketing Research* **19**(3): 288–301.
doi: https://www.doi.org/10.1177/002224378201900302.

[9] Chen, W. and Thissen, D. [1997]. Local dependence indexes for item pairs using item response theory, *Journal of Educational and Behavioral Statistics* **22**(3): 265–289.
doi: https://doi.org/10.3102/10769986022003265.

[10] Duane, S., Kennedy, A., Pendleton, B. and Roweth, D. [1987]. Hybrid monte carlo, *Physics Letters B* **195**(2): 216–222.
doi: https://doi.org/10.1016/0370-2693(87)91197-X.
url: https://www.sciencedirect.com/science/article/pii/037026938791197X.

[11] Edwards, J. and Bagozzi, R. [2000]. On the nature and direction of relationships between constructs and measures, *Psychological Methods* **5**(2): 155–174.
doi: https://www.doi.org/10.1037/1082-989X.5.2.155.

[12] Flores, S. [2012]. *Modelos testlet logísticos y logísticos de exponente positivo para pruebas de compresión de textos*, Master's thesis, Pontificia Universidad Católica del Perú.

[13] Fox, J. [2010]. *Bayesian Item Response Modeling, Theory and Applications*, Statistics for Social and Behavioral Sciences, fienberg, s. and van der linden, w. edn, Springer Science+Business Media, LLC.

[14] Fujimoto, K. [2018a]. The bayesian multilevel trifactor item response theory model, *Educational and Psychological Measurement* **79**(3): 462–494.
doi: https://doi.org/10.1177/0013164418806694.

[15] Fujimoto, K. [2018b]. A general bayesian multilevel multidimensional irt model for locally dependent data, *Br J Math Stat Psychol* **71**(3): 536–560.
doi: https://doi.org/10.1111/bmsp.12133.

[16] Fujimoto, K. [2020]. A more flexible bayesian multilevel bifactor item response theory model, *Journal of Educational Measurement* **57**(2): 255–285.
doi: https://doi.org/10.1111/jedm.12249.

[17] Gelfand, A., Sahu, S. and Carlin, B. [1995]. Efficient parametrisations for normal linear mixed models, *Biometrika* **82**(3): 479–488.
doi: https://doi.org/10.1093/biomet/82.3.479.

[18] Gelfand, A., Sahu, S. and Carlin, B. [1996]. Efficient parameterizations for generalised linear models (with discussion), *in* J. Bernardo, J. Berger, A. Dawid and a. Smith (eds), *Bayesian Statistics*, Vol. 5, pp. 165–180.

[19] Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. [2014]. *Bayesian Data Analysis*, Texts in Statistical Science, third edn, Chapman and Hall/CRC.

[20] Gelman, A. and Rubin, D. [1996]. Markov chain monte carlo methods in biostatistics, *Statistical Methods in Medical Research* **5**(4): 339–355.
doi: https://doi.org/10.1177/096228029600500402.

[21] Geman, S. and Geman, D. [1984]. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**(6): 721–741.
doi: https://doi.org/10.1109/TPAMI.1984.4767596.

[22] Gorinova, M., Moore, D. and Hoffman, M. [2019]. Automatic reparameterisation of probabilistic programs.
url: https://arxiv.org/abs/1906.03028.

[23] Hambleton, R. and Swaminathan, H. [1991]. *Item Response Theory*, Evaluation in Education and Human Services series, Springer Science+Business Media, LLC.

[24] Hambleton, R., Swaminathan, H. and Rogers, H. [1991]. *Fundamentals of Item Response Theory*, SAGE Publications Inc.

[25] Hastings, W. K. [1970]. Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**(1): 97–109.
doi: https://doi.org/10.1093/biomet/57.1.97.
url: https://academic.oup.com/biomet/article-pdf/57/1/97/23940249/57-1-97.pdf.

[26] Heckman, J. [1979]. Sample selection bias as a specification error, **47**(1): 153–161.
doi: https://www.doi.org/10.2307/1912352.
url: https://www.jstor.org/stable/1912352.

[27] Hsieh, M., Proctor, T., Hou, J. and Teo, K. [2010]. A comparison of bayesian mcmc and marginal maximum likehood methods in estimating the item parameters for the 2pl irt model, *International Journal of Innovative Management, Information and Production* **1**(1): 81–89.
url: http://ismeip.org/IJIMIP/contents/imip1011/10IN15T.pdf.

[28] Jiao, H., Kamata, A., Wang, S. and Jin, Y. [2012]. A multilevel testlet model for dual local dependence, *Journal of Educational Measurement* **49**: 82–100.
doi: https://doi.org/10.1111/j.1745-3984.2011.00161.x.

[29] Keane, M. [1992]. A note on identification in the multinomial probit model, *Journal of Business and Economic Statistics* **10**(2): 193–200.
doi: https://doi.org/10.2307/1391677.
url: https://www.jstor.org/stable/1391677.

[30] Kim, S. and Cohen, A. [1999]. Accuracy of parameter estimation in gibbs sampling under the two-parameter logistic model, *Annual Meeting of the American Educational Research Association*, American Educational Research Association.
url: https://eric.ed.gov/?id=ED430012.

[31] Kline, R. [2012]. Assumptions in structural equation modeling, *in* R. Hoyle (ed.), *Handbook of Structural Equation Modeling*, The Guilford Press, chapter 7, pp. 111–125.

[32] Linacre, J. [2021]. *Winsteps® (Version 5.1.0) [Computer Software]*, Portland, Oregon.
url: https://www.winsteps.com/.

[33] Lord, F. and Novik, M. [2008]. *Statistical Theories of Mental Test Scores*, Information Age Publishing.

[34] Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. [2009]. The bugs project: Evolution, critique and future directions, *Statistics in Medicine* **28**(25): 3049–3067.

[35] Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. [2000]. Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility, *Statistics and Computing* (10): 325–337.
doi: https://www.doi.org/10.1023/A:1008929526011.

[36] Martin, J. and McDonald, R. [1975]. Bayesian estimation in unrestricted factor analysis: A treatment for heywood cases, *Psychometrika* (40): 505–517.
doi: https://doi.org/10.1007/BF02291552.

[37] McCullagh, P. and Nelder, J. [1989]. *Generalized Linear Models*, Monographs on Statistics Applied Probability, Chapman Hall/CRC Press.

[38] McElreath, R. [2020]. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Texts in Statistical Science, 2 edn, Chapman and Hall/CRC.
doi: https://doi.org/10.1201/9780429029608.

[39] Metropolis, N., Rosenbluth, A., Rosenbluth, M. and Teller, A. [1953]. Equation of state calculations by fast computing machines, *The Journal of Chemical Physics* **21**: 1087–1092.
doi: https://doi.org/10.1063/1.1699114.

[40] Muller, P. [1991]. A generic approach to posterior integration and gibbs sampling, *Technical Report 91-09*, Department of Statistics, Purdue University.
url: https://www.stat.purdue.edu/docs/research/tech-reports/1991/tr91-09.pdf.

[41] Muthén, L. and Muthén, B. [1998-2011]. *Mplus User's Guide*, CA: Muthén Muthén.

[42] Neal, R. [2011]. Mcmc using hamiltonian dynamics, *in* S. Brooks, A. Gelman, G. Jones and X. Meng (eds), *Handbook of Markov Chain Monte Carlo*, Chapman Hall/CRC Press, chapter 5, pp. 113–162.

[43] Nelder, J. and Wedderburn, W. [1972]. Generalized linear models, *Royal Statistical Society* **135**(3): 370–384.
doi: https://doi.org/10.2307/2344614.
url: https://www.jstor.org/stable/2344614.

[44] Papaspiliopoulos, O., Roberts, G. and Skold, M. [2003]. Non-centered parameterisations for hierarchical models and data augmentation, *Bayesian Statistics* **7**: 307–326.
url: http://econ.upf.edu/ omiros/papers/val7.pdf.

[45] Papaspiliopoulos, O., Roberts, G. and Skold, M. [2007]. A general framework for the parametrization of hierarchical models, *Statistical Science* **22**(1): 59–73.
doi: https://www.doi.org/10.1214/088342307000000014.

[46] Patz, R. J. and Junker, B. W. [1999]. A straightforward approach to markov chain monte carlo methods for item response models, *Journal of Educational and Behavioral Statistics* **24**(2): 146–178.
doi: 10.3102/10769986024002146.

[47] Plummer, M. [2003]. Jags: A program for analysis of bayesian graphical models using gibbs sampling.

[48] R Core Team [2015]. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
url: http://www.R-project.org/.

[49] Rabe-Hesketh, S., Pickles, A. and Skrondal, A. [2003]. Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation, *Statistical Modelling* **3**(3): 215–232.
doi: https://www.doi.org/10.1191/1471082X03st056oa.

[50] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. [2003]. Maximum likelihood estimation of generalized linear models with covariate measurement error, *The Stata Journal* **3**(4): 386–411.
doi: https://www.doi.org/10.1177/1536867X0400300408.
url: https://journals.sagepub.com/doi/pdf/10.1177/1536867X0400300408.

[51] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. [2004a]. Generalized multilevel structural equation modeling, *Psychometrika* **69**(2): 167–190.
doi: https://www.doi.org/10.1007/BF02295939.

[52] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. [2004b]. *GLLAMM Manual*, UC Berkeley Division of Biostatistics.
url: http://www.biostat.jhsph.edu/ fdominic/teaching/bio656/software-gllamm.manual.pdf.

[53] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. [2004c]. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects, *Journal of Econometrics* **128**(2): 301–323.
doi: https://www.doi.org/10.1016/j.jeconom.2004.08.017.
url: http://www.sciencedirect.com/science/article/pii/S0304407604001599.

[54] Rabe-Hesketh, S., Skrondal, A. and Zheng, X. [2012]. Multilevel structural equation modeling, *in* R. Hoyle (ed.), *Handbook of Structural Equation Modeling*, The Guilford Press, chapter 30, pp. 512–531.

[55] Rabe-Hesketh, S., Yang, S. and Pickles, A. [2001]. Multilevel models for censored and latent responses, *Statistical Methods in Medical Research* **10**(6): 409–427.
doi: https://www.doi.org/10.1177/096228020101000604.

[56] Rasch, G. [1980]. *Probabilistic Models for Some Intelligence and Attainment Tests*, University of Chicago Press.

[57] Raudenbush, S. and Bryk, A. [2002]. *Hierarchical linear models: Applications and data analysis methods (Vol. 1)*, Advanced Quantitative Techniques in the Social Sciences, SAGE Publications Inc.

[58] Reckase, M. [2009]. *Multidimensional Item Response Theory*, Statistics for Social and Behavioral Sciences, Springer Science+Business Media, LLC.

[59] Rivera, J. [2019]. *El modelo de respuesta nominal: Aplicación a datos educacionales*, Master's thesis, Pontificia Universidad Católica del Peru.
url: http://hdl.handle.net/20.500.12404/14600.

[60] Skrondal, A. and Rabe-Hesketh, S. [2003a]. Multilevel logistic regression for polytomous data and rankings, *Psychometrika* **68**: 267–287.
doi: https://www.doi.org/10.1007/BF02294801.

[61] Skrondal, A. and Rabe-Hesketh, S. [2003b]. Some applications of generalized linear latent and mixed models in epidemiology: Repeated measures, measurement error and multilevel modeling, *Norsk Epidemiologi* **13**(2): 265–278.

[62] Skrondal, A. and Rabe-Hesketh, S. [2004a]. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Interdisciplinary Statistics, Chapman Hall/CRC Press.

[63] Skrondal, A. and Rabe-Hesketh, S. [2004b]. Generalized linear latent and mixed models with composite links and exploded likelihoods, *in* BiggeriA., E. Dreassi, C. Lagazio and M. Marchi (eds), *Proceedings of the 19th International Workshop on Statistical Modeling*, Firenze University Press, Florence, Italy, pp. 27–39.
url: http://www.gllamm.org/composite$_c$*onf.pdf*.

[64] Stan Development Team [2020a]. RStan: the R interface to Stan. R package version 2.21.2.
url: http://mc-stan.org/.

[65] Stan Development Team. [2020b]. *Stan Modeling Language Users Guide and Reference Manual, version 2.26*, Vienna, Austria.
url: https://mc-stan.org.

[66] Tarazona, E. [2013]. *Modelos alternativos de respuesta graduada con aplicaciones en la calidad de servicios*, Master's thesis, Pontificia Universidad Católica del Perú (PUCP).
url: http://hdl.handle.net/20.500.12404/6175.

[67] Wainer, H., Bradlow, E. and Wang, X. [2007]. *Testlet response theory and its applications*, Cambridge University Press.

[68] Wollack, J. A., Bolt, D. M., Cohen, A. S. and Lee, Y.-S. [2002]. Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and markov chain monte carlo estimation, *Applied Psychological Measurement* **26**(3): 339–352.
doi: https://www.doi.org/10.1177/0146621602026003007.

[69] Yen, W. [1984]. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, *Applied Psychological Measurement* **8**(2): 125–145.
doi: https://doi.org/10.1177/014662168400800201.

AFDELING
Straat nr bus 0000
3000 LEUVEN, BELGIË
tel. + 32 16 00 00 00
fax + 32 16 00 00 00
www.kuleuven.be