PLAUSIBLE VALUES IN STATISTICAL INFERENCE

MAARTEN MARSMAN

Plausible Values in Statistical Inference

M. Marsman

November 19, 2014

Graduation Committee

Chair	Prof. Dr. Ir. A. J. Mouthaan
Promotors	Prof. Dr. C. A. W. Glas
	Prof. Dr. G. K. J. Maris
Assistant promotor	Dr. T. M. Bechger
Members	Prof. Dr. Ir. R. J. Boucherie
	Dr. M. von Davier
	Prof. Dr. Ir. G. J. A. Fox
	Prof. Dr. F. Tuerlinckx
	Prof. Dr. E. M. Wagenmakers

Marsman, Maarten
Plausible Values in Statistical Inference
PhD Thesis University of Twente, Enschede. - Met samenvatting in het Nederlands.
ISBN: 978-90-365-3744-5
doi: 10.3990/1.9789036537445
printed by: PrintPartners Ipskamp B.V., Enschede

PLAUSIBLE VALUES IN STATISTICAL INFERENCE

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday, November 19, 2014 at 16.45

by

Maarten Marsman

born April 2, 1982
in Hellendoorn

This dissertation is approved by the following promoters:

Promotor: Prof. Dr. C. A. W. Glas
Promotor: Prof. Dr. G. K. J. Maris
Assistant promotor: Dr. T. M. Bechger

To my most favourite random variables, Siem en Fem,
and my normalizing constant, Esther.

Acknowledgements

After a bit more than five years, the final product of my work is finally here. In this section I would like to thank all the people that helped to make this possible.

For the largest part, I worked on this thesis at the Psychometric Department of Cito. I would like to say thanks to all my colleagues at Cito for providing a stimulating work environment, and a clean desk. In particular, I would like to thank my two supervisors Gunter Maris and Timo Bechger. You guys have taught me a lot, and without you this thesis would not be here. Every PhD student wants to have Superman and Professor Prlwytzkofsky as supervisors, and I am certainly grateful I did.

I also worked on this thesis at the Department of Research Methodology, Measurement and Data Analysis (OMD) at the University of Twente in Enschede. I would like to say thanks to all my colleagues at OMD for keeping their door open for me, and making me feel welcome. In particular, I would like to thank my supervisor Cees Glas. You inspired me to work in psychometrics, and this thesis is the result of that. I thank you for all the opportunities that you provided me, and for the many interesting discussions we had.

To my friends and family, thanks for the necessary distractions from statistics. In particular, I thank ma en Gerard for always checking up on me, "oma Map" for the coffee and beers on Sunday mornings, Almie for the rockin', Werner for da Bounce and Maarten to ensure that I tap correctly.

Last, but certainly not least, I want to express my gratitude to Esther. I thank you for putting up with me, even when at times my work led me to antisocial behavior. I dedicate this work to you and our Oompa Loompas.

Arnhem, October 2014
Maarten Marsman

Contents

Chapter 1

Introduction

In educational surveys such as the *programme for international student assessment* (PISA)[1], the *national assessment of educational progress* (NAEP)[2] and the *European survey on language competences* (ESLC)[3], *plausible values* are used as a tool for secondary analyses, for researchers that lack the sophisticated resources to estimate the latent regression models. The theoretical underpinning of plausible value imputation was developed by Rubin (1987), and applied to large-scale assessment by Mislevy and colleagues (Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Beaton & Johnson, 1992; Mislevy, 1993; von Davier, Gonzalez, & Mislevy, 2009).

A plausible value for a pupil p is a draw from the posterior distribution of his or her (usually unidimensional) ability θ_p, given his or her vector of item responses $\mathbf{x}_p$ and any additional information available about the student that is encoded in a vector of covariates $\mathbf{y}_p$. The posterior distribution is

$$f(\theta|\mathbf{x}_p, \mathbf{y}_p) \propto P(\mathbf{x}_p|\theta, \boldsymbol{\delta}) f(\theta|\boldsymbol{\lambda}, \mathbf{y}_p) \tag{1.1}$$

where $P(\mathbf{x}_p|\theta, \boldsymbol{\delta})$ denotes an *item response theory* (IRT) model (Lord & Novick, 1968) with item parameters $\boldsymbol{\delta}$ and $f(\theta|\boldsymbol{\lambda}, \mathbf{y})$ denotes a population model with parameters $\boldsymbol{\lambda}$.

The primary aim of analyses of educational surveys is to estimate the population model, $f(\theta|\boldsymbol{\lambda}, \mathbf{y})$, usually the latent regression model

$$\theta = \mathbf{y}^T \boldsymbol{\beta} + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2). \tag{1.2}$$

Estimation of the population model needs specialized software and expertise. Note, however, that the population model can also be estimated using plausible values. For instance, the latent regression model (1.2) could be estimated repeatedly using plausible values as the dependent variables, and aggregating the results over replications. In this role, plausible values have been a source of confusion, leading to questions like: why would plausible values be needed if the population model

[1] `www.oecd.org/pisa/`

[2] `nces.ed.gov/nationsreportcard/`

[3] `www.surveylang.org`

has already been estimated to produce them? Or, put differently. *What can we learn from plausible values outside of what is already known from the population model*?

In Chapter 2 it is argued that plausible values are more than just a simple tool to facilitate secondary analyses. In fact, plausible values should play the central role in population inferences from surveys, since they hold information that dominates the information in the population model. Here, the population model is properly seen as a prior distribution. It is shown that, given an IRT model and a population model, the marginal distribution of plausible values is a better estimator of the unknown ability distribution than the population model, and the population model is merely a vehicle to generate plausible values.

An interesting question in this context, is why plausible values are called plausible values and not "draws from the posterior distribution of ability". The reason can only be guessed, but it is probably due to the lack of acceptance of Bayesian methods in psychometrics in the late 1980's and early 1990's, when the concept of plausible values was first introduced. Nowadays, Bayesian methods are commonly applied and widely accepted, also in psychometrics. One of the reasons for this change is known as the *MCMC revolution* (Cappe & Robert, 2000; Brooks, 2003; Robert & Cassela, 2011), where it was recognized that complex models could be estimated in a Bayesian framework using Markov chain Monte Carlo (MCMC) computational methods. Although both the Bayesian approach and MCMC methods have been around for many decades, the increase of computing power made MCMC entirely feasible, and led to its popularization in statistics in the late 1980's and early 1990's (Gelfand & Smith, 1990; Zeger & Karim, 1991; Casella & George, 1992; Albert & Chib, 1993; Tierney, 1994). It was especially the Gibbs sampler (Geman & Geman, 1984) that received much attention, since it facilitates sampling from complex multivariate posterior distributions, and as such, solves estimation problems for models which lead to great difficulty in a (marginal) maximum likelihood framework (Bock & Lieberman, 1970; Bock & Aitken, 1981). The MCMC revolution has also found its way into psychometrics, where it has been popularized by Albert (1992) and Patz and Junker (1999b), with further applications to the estimation of models for testlets (Bradlow, Wainer, & Wang, 1999), latent classes (Hoijtink & Molenaar, 1997), multilevel IRT models (Fox & Glas, 2001), random item parameters (Janssen, Tuerlinckx, Meulders, & de Boeck, 2000), and multidimensional IRT models (Béguin & Glas, 2001).

The current *big-data revolution* poses new challenges for MCMC methods, as the size of applications are rapidly increasing. For instance, the 2012 PISA cycle required plausible values from over $500,000$ pupils which, combined with the complexity of the survey design, IRT and population model, poses a formidable task to any statistician. In Chapter 5, it is shown that certain existing methods that are available for such applications are not very efficient (or useful for that matter), and specialized algorithms need to be developed. To this aim, novel MCMC methods are introduced in Chapters 3 and 4. Specifically, two existing algorithms are adapted in Chapter 3, such that they become more efficient for big data. The methods introduced in Chapter 3 are particularly suited for large numbers of random effects, such as the pupils' abilities in educational surveys. A method that is also suited for small numbers of random effects, such as the

item parameters in educational surveys, is introduced in Chapter 4. The method in Chapter 4 becomes more efficient, the more observations there are about the random effect, and we typically have many observations on the items in educational surveys.

With the increase in the amount of data that becomes available, also comes an increase in the complexity of the models that are used. One recent development in this area is the use of network models (Barabási, 2012), which has already found its way into psychometrics (Cramer, Waldorp, van der Maas, & Borsboom, 2010; van Borkulo et al., in press). Of particular interest is the Ising (1925) network model to which we thank most of the MCMC methods that we know today[4]. In Chapter 4 it is shown that the Ising model can be characterized as a marginal IRT model using Mark Kac's (1968) latent variable representation. This relation opens the door to new lines of research, and is used in combination with a *full-data-information* estimation approach to estimate the Ising model. Estimating the Ising model has been known to be a notoriously difficult task, due to the intractability of its normalizing constant and to the large amount of parameters involved. The full-data-information approach circumvents computing the normalizing constant, which makes estimating the Ising model computationally feasible, even for large networks. However, in an n variable network there are $n(n+1)/2$ unknown parameters, and even for relatively small networks a large amount of replications of the network state are needed to estimating all parameters. In Chapter 4 low-rank approximations to the matrix of pairwise interactions are used, where all but the largest eigenvalues of the matrix are equated to zero. Such an approximation is ideally suited for the dense networks that are typically found in the social sciences, where the first few eigenvalues of the matrix capture most of the underlying structure.

1.1 Outline

The next four chapters in this thesis are written in an article style, intended to be self-contained, and some overlap could not be avoided. The methodology that is the topic of this thesis is highly practical and small pieces of GNU-R (R Core Team, 2014) code are provided in the appendices to help those who wish to try if the proposed methods work. In some cases, an appendix provides material that was simply to much fun to leave out.

In Chapter 2 it is shown that the marginal distribution of plausible values is a consistent estimator of the true latent variable distribution, and, furthermore, that convergence is *monotone* in an embedding in which the number of items tends to infinity. This result is used to clarify some of the misconceptions that exist about plausible values, and also to show how they can be used in the analyses of educational surveys.

Chapter 3 is about two recently published algorithms that can be used to sample from conditional distributions, such as the posterior distribution of abil-

[4]MCMC methods such as Metropolis-Hastings (Metropolis, Rosenbluth, Rosenbluth, & Teller, 1953; Hastings, 1970), Gibbs sampling (Creutz, 1980; Rubinstein, 1981), Perfect Sampling (Propp & Wilson, 1996), and Data Augmentation (Swendsen & Wang, 1987), have been developed purely with the aim to generate data for the Ising model.

ity (i.e. plausible values). It is shown how the efficiency of the two algorithms can be improved when a sample is required from many conditional distributions. Using simulated-data and real-data examples from educational measurement, it is shown how the algorithms can be used to sample from intractable full-conditional distributions of the person and item parameters in an application of the Gibbs sampler.

Estimating the structure of networks is a notoriously difficult problem. In Chapter 4 it is shown that using a latent variable representation of the Ising network, it is possible to employ a full data information approach to uncover the network structure, thereby, only ignoring information encoded in the prior distribution (of the latent variables). The full data information approach avoids having to compute the partition function, and is thus computationally feasible, even for networks with many nodes. We illustrate the full data information approach to the estimation of dense networks, thereby complementing recent approaches based on regularization for nearest neighbour networks (van Borkulo et al., in press).

In Chapter 5 an example is used to demonstrate that the autocorrelation in MCMC sampling methods using data-augmentation (Tanner & Wong, 1987) may depend on sample size. This means that, in some cases, the Markov chain will eventually stop mixing when more and more data are observed and thus becomes useless.

Chapter 2

A Non-Parametric Estimator of Latent Variable Distributions

2.1 Introduction

In educational surveys an *item response theory* (IRT) model is used to model the conditional distribution of a vector of item responses $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ as a function of a latent random variable (ability) Θ, where the item response functions, that is, the expected responses as function of ability, are monotonically increasing in ability. The IRT model characterizes the latent variable Θ, and the goal of educational surveys is to estimate the distribution of Θ, which we denote by f. Together, the IRT model and the ability distribution induce the following statistical model:

$$P(\mathbf{X}_f = \mathbf{x}) = \int_{\mathbb{R}} P(\mathbf{X} = \mathbf{x}|\theta) f(\theta) d\theta,$$

where $P(\mathbf{X}_f)$ is the true data distribution of which we obtain a sample. Throughout this chapter we assume that the IRT model is given, and focus on the unknown f. We furthermore assume that the random variables X_i are discrete and have a finite number of possible realizations. Our results remain unchanged for the case that the X_i are continuous and the reader may replace corresponding sums by integrals.

There are four possible approaches to estimate f from the observed data. The first entails the derivation of a function T such that $T(\mathbf{X}) \sim \Theta$. Usually, $\mathbf{X}$ is discrete, and thus, the realizations of $T(\mathbf{X})$ are discrete as well. This implies that the first approach only works when ability is discrete, as in the extended Rasch model (Cressie & Holland, 1983) and the semi-parametric Rasch model (Tjur, 1982), but it doesn't work for continuous Θ. The second approach entails the derivation of a function T such that $T(\mathbf{X}) \xrightarrow{\mathcal{L}} \Theta$, i.e., a random variable that,

Adapted from: Marsman, M., Maris, G., Bechger, T. & Glas, C. (2014). A Non-Parametric Estimator of Latent Variable Distributions. Submitted for publication.

asymptotically, has the same distribution as Θ. This can be any function T that is a consistent estimator of Θ. Common examples are the Maximum Likelihood (ML) or Weighted ML (WML) estimator (Warm, 1989). The third approach entails using the realizations of $\mathbf{X}$ to generate a random variable Θ^* such that $\Theta^* \perp\!\!\!\perp \Theta|\mathbf{X}$ and $\Theta^* \sim \Theta$. That is, Θ and Θ^* are *exchangeable* and their joint density can be written as:

$$f(\theta^*, \theta) = \sum_{\mathbf{x}} f(\theta^*|\mathbf{X} = \mathbf{x}) f(\theta|\mathbf{X} = \mathbf{x}) P(\mathbf{X}_f = \mathbf{x}),$$

where the summation is over all possible realizations of $\mathbf{X}$. Note that the conditional distributions $f(\theta|\mathbf{X})$ are posterior distributions, and in order to construct them we need to know f: the one thing we do not know. We could formulate prior beliefs about what f could be and choose a prior distribution g, and use g to construct posteriors $g(\theta|\mathbf{X})$. Draws from these posteriors are called *plausible values* in the psychometric literature (Mislevy, 1991, 1993). In this chapter we prove that, under mild regularity conditions, plausible values are random variables of the form $\tilde{\Theta}|\mathbf{X}$ such that $\tilde{\Theta} \xrightarrow{\mathcal{L}} \Theta$, which constitutes the fourth and final approach. That is, we will show that the marginal distribution of plausible values is a consistent estimator of f, and, furthermore, that convergence is *monotone* in an embedding where the number of items, n, tends to infinity.

Let $\tilde{g}$ denote the marginal distribution of plausible values, i.e.,

$$\tilde{g}(\theta) = \sum_{\mathbf{x}} g(\theta|\mathbf{X} = \mathbf{x}) P(\mathbf{X}_f = \mathbf{x}).$$

The marginal distribution of plausible values is intractable (due in part to the unknown $P(\mathbf{X}_f)$), yet it is easily sampled from using Monte Carlo integration: we obtain realizations from $P(\mathbf{X}_f)$ during data collection, and we use these realizations to sample plausible values. The empirical distribution of the plausible values thus obtained is the non-parametric estimator of $\tilde{g}$, and if $\tilde{g}$ converges to f, it is also the non-parametric estimator of f. Thus, our goal in this chapter is to prove that $\tilde{g}$ converges in law to f (i.e. $\tilde{\Theta} = \Theta_{\tilde{g}} \xrightarrow{\mathcal{L}} \Theta_f$).

The plan is as follows. Instead of proving the convergence in law directly, we instead prove the stronger result that $\tilde{g}$ converges to f in Kullback-Leibler (KL; Kullback & Leibler, 1951) divergence. This is sufficient for our purpose, since convergence in KL divergence implies convergence in law (DasGupta, 2008, p. 21). To prove that $\tilde{g}$ converges to f in KL divergence, we instead prove an even stronger result. Namely, that $\tilde{g}$ converges to f in Expected KL (EKL) divergence, see Definition 1. Moreover, we prove that convergence in EKL divergence is monotone as the number of items n tends to infinity.

Definition 1. *The Expected Kullback-Leibler (EKL) divergence between $\Theta_f|\boldsymbol{X}$ and $\Theta_g|\boldsymbol{X}$, w.r.t. $f(\Theta|\boldsymbol{X})$ and $P(\boldsymbol{X}_f)$ is:*

$$\begin{aligned}\mathbb{E}(\Delta(\Theta_f; \Theta_g|\boldsymbol{X}_f)) &= \sum_{\boldsymbol{x}} \Delta(\Theta_f; \Theta_g|\boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X}_f = \boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}} \left[\int_{\mathbb{R}} \ln\left(\frac{f(\theta|\boldsymbol{X} = \boldsymbol{x})}{g(\theta|\boldsymbol{X} = \boldsymbol{x})} \right) f(\theta|\boldsymbol{X} = \boldsymbol{x}) d\theta \right] P(\boldsymbol{X}_f = \boldsymbol{x}),\end{aligned}$$

where $\Delta(\Theta_f;\Theta_g|\boldsymbol{X})$ denotes the KL divergence of $f(\Theta|\boldsymbol{X})$ and $g(\Theta|\boldsymbol{X})$ with respect to $f(\Theta|\boldsymbol{X})$.

To see why convergence in EKL divergence is stronger than convergence in KL divergence, we consider the following Theorem:

Theorem 1. *KL divergence of $\Theta_{\tilde{g}}$ w.r.t. Θ_f, i.e.,*

$$\Delta(\Theta_f;\Theta_{\tilde{g}}) = \int_{\mathbb{R}} \ln \frac{f(\theta)}{\tilde{g}(\theta)} f(\theta) d\theta,$$

is always smaller than or equal to EKL divergence (see Definition 1). That is,

$$\Delta(\Theta_f;\Theta_{\tilde{g}}) \leq \mathbb{E}(\Delta(\Theta_f;\Theta_{\tilde{g}}|\boldsymbol{X}_f)).$$

Proof. We start with rewriting the logarithm of the ratio of $\tilde{g}$ over f:

$$\begin{aligned}
\ln \frac{\tilde{g}(\theta)}{f(\theta)} &= \ln \left\{ \frac{\sum_{\mathbf{x}} g(\theta|\mathbf{X}=\mathbf{x}) P(\mathbf{X}_f=\mathbf{x})}{\sum_{\mathbf{x}} f(\theta|\mathbf{X}=\mathbf{x}) P(\mathbf{X}_f=\mathbf{x})} \right\} \\
&= \ln \left\{ \sum_{\mathbf{x}} \frac{g(\theta|\mathbf{X}=\mathbf{x}) P(\mathbf{X}_f=\mathbf{x})}{f(\theta|\mathbf{X}=\mathbf{x}) P(\mathbf{X}_f=\mathbf{x})} \frac{f(\theta|\mathbf{X}=\mathbf{x}) P(\mathbf{X}_f=\mathbf{x})}{\sum_{\mathbf{x}} f(\theta|\mathbf{X}=\mathbf{x}) P(\mathbf{X}_f=\mathbf{x})} \right\} \\
&= \ln \left\{ \sum_{\mathbf{x}} \frac{g(\theta|\mathbf{X}=\mathbf{x})}{f(\theta|\mathbf{X}=\mathbf{x})} P(\mathbf{X}=\mathbf{x}|\theta) \right\} \\
&\geq \sum_{\mathbf{x}} \ln \frac{g(\theta|\mathbf{X}=\mathbf{x})}{f(\theta|\mathbf{X}=\mathbf{x})} P(\mathbf{X}=\mathbf{x}|\theta),
\end{aligned}$$

using Jensen's inequality. Thus, we obtain

$$\ln \frac{f(\theta)}{\tilde{g}(\theta)} \leq \sum_{\mathbf{x}} \ln \frac{f(\theta|\mathbf{X}=\mathbf{x})}{g(\theta|\mathbf{X}=\mathbf{x})} P(\mathbf{X}=\mathbf{x}|\theta).$$

Integrating both sides of this expression w.r.t. f gives the desired result:

$$\begin{aligned}
\int_{\mathbb{R}} \ln \frac{f(\theta)}{\tilde{g}(\theta)} f(\theta) d\theta &\leq \int_{\mathbb{R}} \sum_{\mathbf{x}} \ln \frac{f(\theta|\mathbf{X}=\mathbf{x})}{g(\theta|\mathbf{X}=\mathbf{x})} P(\mathbf{X}=\mathbf{x}|\theta) f(\theta) d\theta \\
&= \int_{\mathbb{R}} \sum_{\mathbf{x}} \ln \frac{f(\theta|\mathbf{X}=\mathbf{x})}{g(\theta|\mathbf{X}=\mathbf{x})} f(\theta|\mathbf{X}=\mathbf{x}) d\theta P(\mathbf{X}_f=\mathbf{x}).
\end{aligned}$$

□

In this chapter we assume that all divergences exist, meaning that they are finite, which is true if the support of g contains that of f (i.e. f is absolutely continuous w.r.t. g), except possible for data points $\mathbf{X}$ of which the asymptotic probability becomes arbitrarily small (null sets). Note that the (E)KL divergences that we use in this chapter are non-symmetric in their arguments, yet their values are always non-negative and equal zero, if and only if, the compared probability distributions are the same a.e. (see Theorem 9.6.1 in Cover & Thomas, 1991, p. 232).

The outline of this chapter is as follows. First, we prove that EKL divergence is monotonically non-increasing function in n. Then, we prove that the EKL divergence tends to zero as the number of items n tends to infinity using standard results from Bayesian theory. Once we have established that the sequence of EKL divergences converges to zero, and thus that $\tilde{g}$ converges to f, we discuss a number of implications of our results for educational surveys. The chapter ends with a discussion of our main findings.

2.2 Monotone Convergence of Plausible Values

Using the definition of the posterior, and given the IRT model $P(\mathbf{X}|\theta)$, we rewrite the EKL divergence as follows:

$$\begin{aligned}\mathbb{E}(\Delta(\Theta_f;\Theta_g|\mathbf{X}_f)) &= \sum_{\mathbf{x}} \int_{\mathbb{R}} \ln\left(\frac{\frac{P(\mathbf{X}=\mathbf{x}|\theta)f(\theta)}{P(\mathbf{X}_f=\mathbf{x})}}{\frac{P(\mathbf{X}=\mathbf{x}|\theta)g(\theta)}{P(\mathbf{X}_g=\mathbf{x})}}\right) f(\theta|\mathbf{X}=\mathbf{x})d\theta P(\mathbf{X}_f=\mathbf{x}) \\ &= \sum_{\mathbf{x}} \int_{\mathbb{R}} \ln\left(\frac{f(\theta)}{g(\theta)}\frac{P(\mathbf{X}_g=\mathbf{x})}{P(\mathbf{X}_f=\mathbf{x})}\right) f(\theta|\mathbf{X}=\mathbf{x})d\theta P(\mathbf{X}_f=\mathbf{x}),\end{aligned}$$

where $P(\mathbf{X}_g)$ is the distribution of the data under the prior g. Using properties of the logarithm we obtain

$$\begin{aligned}\mathbb{E}(\Delta(\Theta_f;\Theta_g|\mathbf{X}_f)) = &\sum_{\mathbf{x}} \int_{\mathbb{R}} \ln\left(\frac{f(\theta)}{g(\theta)}\right) f(\theta|\mathbf{X}=\mathbf{x})d\theta P(\mathbf{X}_f=\mathbf{x}) \\ &+ \sum_{\mathbf{x}} \int_{\mathbb{R}} \ln\left(\frac{P(\mathbf{X}_g=\mathbf{x})}{P(\mathbf{X}_f=\mathbf{x})}\right) f(\theta|\mathbf{X}=\mathbf{x})d\theta P(\mathbf{X}_f=\mathbf{x}).\end{aligned}$$

If we sum over the possible values of $\mathbf{X}$ in the first term and integrate over Θ in the second term, respectively, we obtain

$$\begin{aligned}\mathbb{E}(\Delta(\Theta_f;\Theta_g|\mathbf{X}_f)) &= \int_{\mathbb{R}} \ln\left(\frac{f(\theta)}{g(\theta)}\right) f(\theta)d\theta + \sum_{\mathbf{x}} \ln\left(\frac{P(\mathbf{X}_g=\mathbf{x})}{P(\mathbf{X}_f=\mathbf{x})}\right) P(\mathbf{X}_f=\mathbf{x}) \\ &= \int_{\mathbb{R}} \ln\left(\frac{f(\theta)}{g(\theta)}\right) f(\theta)d\theta - \sum_{\mathbf{x}} \ln\left(\frac{P(\mathbf{X}_f=\mathbf{x})}{P(\mathbf{X}_g=\mathbf{x})}\right) P(\mathbf{X}_f=\mathbf{x}) \\ &= \Delta(\Theta_f;\Theta_g) - \Delta(\mathbf{X}_f;\mathbf{X}_g).\end{aligned}$$

It follows that EKL divergence of the posterior distribution is equal to the difference between *prior divergence* $\Delta(\Theta_f;\Theta_g)$ and *marginal divergence* $\Delta(\mathbf{X}_f;\mathbf{X}_g)$ (i.e. divergence of $P(\mathbf{X}_g)$ w.r.t. $P(\mathbf{X}_f)$). For future reference, we state this as Lemma 1.

Lemma 1. *Expected Kullback-Leibler divergence of $\Theta_f|\boldsymbol{X}$ and $\Theta_g|\boldsymbol{X}$, w.r.t. $f(\Theta|\boldsymbol{X})$ and $P(\boldsymbol{X}_f)$, equals prior divergence minus marginal divergence, that is,*

$$\mathbb{E}(\Delta(\Theta_f;\Theta_g|\boldsymbol{X}_f)) = \Delta(\Theta_f;\Theta_g) - \Delta(\boldsymbol{X}_f;\boldsymbol{X}_g).$$

Lemma 1 implies that $\mathbb{E}(\Delta(\Theta_f;\Theta_g|\mathbf{X}_f))$ equals zero iff prior divergence is equal to marginal divergence. Since the divergences are finite and non-negative, we find that

$$\Delta(\Theta_f;\Theta_g) \geq \Delta(\mathbf{X}_f;\mathbf{X}_g).$$

We are now going to prove that $\Delta(\mathbf{X}_f;\mathbf{X}_g)$ is a monotone non-decreasing sequence in n, for which $\Delta(\Theta_f;\Theta_g)$ is an upper bound. To this aim, we consider what happens to marginal divergence when an item is added, i.e., when n is increased to $n+1$. To fix the notation, let $X_1, X_2, ...,$ denote an infinite sequence of item responses, with X_n the n-th element and $\mathbf{X}_n$ a vector consisting of the first n elements of this sequence. The marginal divergence for $n+1$ items is

$$\Delta(\mathbf{X}_{f,n+1};\mathbf{X}_{g,n+1}) = \sum_{\mathbf{x}_{n+1}} \ln\left(\frac{P(\mathbf{X}_{f,n+1}=\mathbf{x}_{n+1})}{P(\mathbf{X}_{g,n+1}=\mathbf{x}_{n+1})}\right) P(\mathbf{X}_{f,n+1}=\mathbf{x}_{n+1}).$$

Conditioning on the first n observations and factoring the distribution, we obtain

$$\begin{aligned}\Delta(\mathbf{X}_{f,n+1};\mathbf{X}_{g,n+1}) =& \sum_{\mathbf{x}_n}\sum_{x_{n+1}} \ln\left(\frac{P(X_{f,n+1}=x_{n+1}|\mathbf{X}_n=\mathbf{x}_n)}{P(X_{g,n+1}=x_{n+1}|\mathbf{X}_n=\mathbf{x}_n)}\frac{P(\mathbf{X}_{f,n}=\mathbf{x}_n)}{P(\mathbf{X}_{g,n}=\mathbf{x}_n)}\right)\\ &\times P(X_{f,n+1}=x_{n+1}|\mathbf{X}_n=\mathbf{x}_n)P(\mathbf{X}_{f,n}=\mathbf{x}_n).\end{aligned}$$

This is equal to

$$\begin{aligned}\Delta(\mathbf{X}_{f,n+1};\mathbf{X}_{g,n+1}) =& \sum_{\mathbf{x}_n}\sum_{x_{n+1}} \ln\left(\frac{P(X_{f,n+1}=x_{n+1}|\mathbf{X}_n=\mathbf{x}_n)}{P(X_{g,n+1}=x_{n+1}|\mathbf{X}_n=\mathbf{x}_n)}\right)\\ &\times P(X_{f,n+1}=x_{n+1}|\mathbf{X}_n=\mathbf{x}_n)P(\mathbf{X}_{f,n}=\mathbf{x}_n)\\ &+\sum_{\mathbf{x}_n}\ln\left(\frac{P(\mathbf{X}_{f,n}=\mathbf{x}_n)}{P(\mathbf{X}_{g,n}=\mathbf{x}_n)}\right)P(\mathbf{X}_{f,n}=\mathbf{x}_n)\\ =&\mathbb{E}(\Delta(X_{f,n+1};X_{g,n+1}|\mathbf{X}_{f,n})) + \Delta(\mathbf{X}_{f,n};\mathbf{X}_{g,n}),\end{aligned}$$

a result that is closely related to the *chain rule* of KL divergence (Cover & Thomas, 1991, p. 23). Since, $\mathbb{E}(\Delta(X_{f,n+1};X_{g,n+1}|\mathbf{X}_{f,n})) \geq 0$, we see that

$$\Delta(\mathbf{X}_{f,n+1};\mathbf{X}_{g,n+1}) \geq \Delta(\mathbf{X}_{f,n};\mathbf{X}_{g,n}).$$

For future reference we will state this as Lemma 2.

Lemma 2. *Marginal divergence for $n+1$ observations is larger than or equal to marginal divergence for n observations:*

$$\Delta(\boldsymbol{X}_{f,n+1};\boldsymbol{X}_{g,n+1}) \geq \Delta(\boldsymbol{X}_{f,n};\boldsymbol{X}_{g,n}).$$

Using Lemma 1 and Lemma 2 we can now state our first Theorem.

Theorem 2 (Monotonicity Theorem)**.** *Given an IRT model $P(\boldsymbol{X}|\theta)$ and the assumption that prior divergence is finite, $\mathbb{E}(\Delta(\Theta_f;\Theta_g|\boldsymbol{X}_{f,n}))$ is monotone non-increasing in the number of items n.*

Proof. From Lemma 1 and Lemma 2 we obtain

$$\begin{aligned}\mathbb{E}(\Delta(\Theta_f;\Theta_g|\mathbf{X}_{f,n+1})) =& \Delta(\Theta_f;\Theta_g) - \Delta(\mathbf{X}_{f,n+1};\mathbf{X}_{g,n+1}) \\ =& \Delta(\Theta_f;\Theta_g) - \mathbb{E}(\Delta(X_{f,n+1};X_{g,n+1}|\mathbf{X}_{f,n})) \\ & - \Delta(\mathbf{X}_{f,n};\mathbf{X}_{g,n}),\end{aligned}$$

and Lemma 1 shows that the first and last term are equal to the EKL divergence for n items. Thus, we have

$$\mathbb{E}(\Delta(\Theta_f;\Theta_g|\mathbf{X}_{f,n+1})) = \mathbb{E}(\Delta(\Theta_f;\Theta_g|\mathbf{X}_{f,n})) - \mathbb{E}(\Delta(X_{f,n+1};X_{g,n+1}|\mathbf{X}_{f,n})).$$

This implies a sequence of EKL divergences which adheres to the (in-)equality:

$$0 \leq \mathbb{E}(\Delta(\Theta_f;\Theta_g|\mathbf{X}_{f,n+1})) \leq \mathbb{E}(\Delta(\Theta_f;\Theta_g|\mathbf{X}_{f,n})) \leq \Delta(\Theta_f;\Theta_g),$$

i.e., a monotone non-increasing sequence in n with lower bound 0. Since prior divergence is finite, it is an upper bound for this sequence. Since a bounded monotone sequence converges, the result follows. □

2.3 Large Sample Properties of Plausible Values

The Monotonicity Theorem shows that the sequence of EKL divergences converges in an embedding in which $n \to \infty$. This does not complete the proof that the marginal distribution of plausible values converges to f, since the sequence of EKL divergences may converge to a number that is strictly larger than zero. Thus, to prove that the marginal distribution of plausible values converges to f, we must show that the sequence of EKL divergences converges to zero. Since the EKL divergence is equal to the difference between prior and marginal divergence, we have to show that

$$\Delta(\Theta_f;\Theta_g) \geq \Delta(\mathbf{X}_{f,n};\mathbf{X}_{g,n}), \tag{2.1}$$

becomes an equality as $n \to \infty$.

We start with a direct proof of (2.1) (suppressing the dependence on n). Note first that,

$$\begin{aligned}\forall \mathbf{x} : \ln \frac{P(\mathbf{X}_f = \mathbf{x})}{P(\mathbf{X}_g = \mathbf{x})} &= -\ln \frac{P(\mathbf{X}_g = \mathbf{x})}{P(\mathbf{X}_f = \mathbf{x})} \\ &= -\ln \frac{\int_{\mathbb{R}} P(\mathbf{X} = \mathbf{x}|\theta) g(\theta) d\theta}{\int_{\mathbb{R}} P(\mathbf{X} = \mathbf{x}|\theta) f(\theta) d\theta} \\ &= -\ln \int_{\mathbb{R}} \frac{P(\mathbf{X} = \mathbf{x}|\theta) g(\theta)}{P(\mathbf{X} = \mathbf{x}|\theta) f(\theta)} \frac{P(\mathbf{X} = \mathbf{x}|\theta) f(\theta)}{\int_{\mathbb{R}} P(\mathbf{X} = \mathbf{x}|\theta) f(\theta) d\theta} d\theta \\ &= -\ln \int_{\mathbb{R}} \frac{g(\theta)}{f(\theta)} f(\theta|\mathbf{X} = \mathbf{x}) d\theta.\end{aligned}$$

We can now use Jensen's Inequality to obtain:

$$\forall \mathbf{x} : \ln \frac{P(\mathbf{X}_f = \mathbf{x})}{P(\mathbf{X}_g = \mathbf{x})} \leq -\int_{\mathbb{R}} \ln \frac{g(\theta)}{f(\theta)} f(\theta|\mathbf{X} = \mathbf{x}) d\theta = \int_{\mathbb{R}} \ln \frac{f(\theta)}{g(\theta)} f(\theta|\mathbf{X} = \mathbf{x}) d\theta. \tag{2.2}$$

Taking expectations w.r.t. $P_f(\mathbf{X})$ gives the inequality in (2.1). This shows that marginal divergence equals prior divergence, whenever Jensen's inequality in (2.2) becomes an equality. This happens, if and only if, $f(\theta|\mathbf{X}_n)$ becomes degenerate at a single point, which means that all mass of $f(\theta|\mathbf{X}_n)$ becomes concentrated at a single point as $n \to \infty$. Thus, the sequence of EKL divergences converges to zero, when $f(\theta|\mathbf{X}_n)$ has a single mode and a variance that tends to zero sufficiently fast as $n \to \infty$.

Instead of proving that $f(\theta|\mathbf{X}_n)$ becomes degenerate, we can make use of a stronger result. In the context of IRT, Chang and Stout (1993) proved that under mild regularity conditions the posterior $f(\theta|\mathbf{X}_n)$ almost surely converges to $\mathcal{N}(\hat{\theta}, \mathcal{I}_n(\hat{\theta})^{-1})$, where $\hat{\theta}$ is the MLE of θ and $\mathcal{I}_n(\hat{\theta})$ is the test (or Fisher) information evaluated at $\hat{\theta}$. If test information tends to infinity as the number of items increases, the posterior variance tends to zero and the posterior becomes a degenerate distribution at $\hat{\theta}$, in which case marginal divergence equals prior divergence. The assumption that test information increases when items are added to the test is a reasonable assumption to make for standard applications in which monotone IRT models are used. Although the conditions of Chang and Stout (1993) are not necessary, and more than we require to proof that marginal divergence tends to prior divergence as $n \to \infty$, they are sufficient. This completes the proof of Theorem 3.

Theorem 3 (Convergence Theorem)**.** *Under mild regularity conditions,*

$$\lim_{n\to\infty} \mathbb{E}(\Delta(\Theta_f; \Theta_g|\boldsymbol{X}_{f,n})) = 0.$$

Together, the Monotonicity Theorem and the Convergence Theorem imply Theorem 4, which we state for future reference.

Theorem 4 (Monotone Convergence Theorem)**.** *Under mild regularity conditions, the marginal distribution of plausible values converges monotonically to the true latent variable distribution in Expected Kullback-Leibler divergence and an embedding in which $n \to \infty$.*

In summary, the Monotone Convergence Theorem states that (under mild regularity conditions) the marginal distribution of plausible values $\tilde{g}$ is a *consistent* estimator of the true ability distribution f.

2.4 Implications

In plain words, the Monotone Convergence Theorem implies that we can use plausible values to learn about the true distribution of ability. This result has important practical implications that we discuss and illustrate in this section. We start with addressing some misconceptions that exist about plausible values, and then discuss how our results are relevant for the design and analyses of educational surveys. We remind the reader that g is a prior distribution, f the true distribution and $\tilde{g}$ the marginal distribution of plausible values.

2.4.1 What can we learn from Plausible Values?

In educational surveys usually an estimated population model is used as prior distribution g to generate plausible values. One misconception about plausible values is that nothing can be learned from plausible values over that which is already known from the population model (prior distribution), see, for instance, the article of Kreiner and Christensen (2014). We can use our results to clarify this misconception.

If we assume that nothing can be learned from plausible values over that which is known from the population model, we assume that $\tilde{g} = g$. Our results show that this is true, if and only if, the population model is the true ability distribution (i.e. $g = f$). Moreover, if $\tilde{g} \neq g$, hence $g \neq f$, the distribution of plausible values $\tilde{g}$ comes closer to f than g is. In view of this, it is important to note that plausible values are often introduced merely as a kindness to the analysts conducting secondary analyses; analysts who often lack the sophisticated resources to estimate the latent regression IRT models used in educational surveys. Our results suggest that it is not the latent regression models but the plausible values that should play the central role in the analyses of educational survey data. The latent regression models figure merely as vehicles to obtain plausible values.

Empirical example

To illustrate that the plausible value distribution may diverge from the prior in applications, we analyse data from the 2006 PISA cycle. More specifically, we used data from $n = 26$ items intended to assess reading ability in booklet 6 made by $N = 1,738$ Canadian students (see Appendix A for details of this analyses).

We used the *One Parameter Logistic Model* (OPLM; Verhelst & Glas, 1995) as IRT model, and a standard normal distribution as prior. Using the IRT model and prior distribution, a single plausible value was generated for each of the N persons. The ecdf of N draws from the prior distribution g, and the ecdf of generated plausible values are shown in Figure 2.1. The marginal distribution of plausible values is clearly different from the specified prior distribution.

2.4.2 How can we tell that we have the "wrong" prior distribution?

Another misconception about plausible values is that if the population model (or prior distribution) of Θ is wrong (i.e. $g \neq f$), there is nothing to be learned from looking at the plausible values. Kreiner and Christensen (2014) explicitly voice this opinion:

> [...] a [...] reason to suspect plausible values may be less than plausible is the assumption that the distribution of Θ is conditionally Gaussian. If this assumption is false, it follows that the density of distribution of plausible values $\tilde{\Theta}$ is not the same as the density of distribution of Θ; and thus, there is no reason to be interested in the distribution of $\tilde{\Theta}$ at all. *[Section 3.3]*

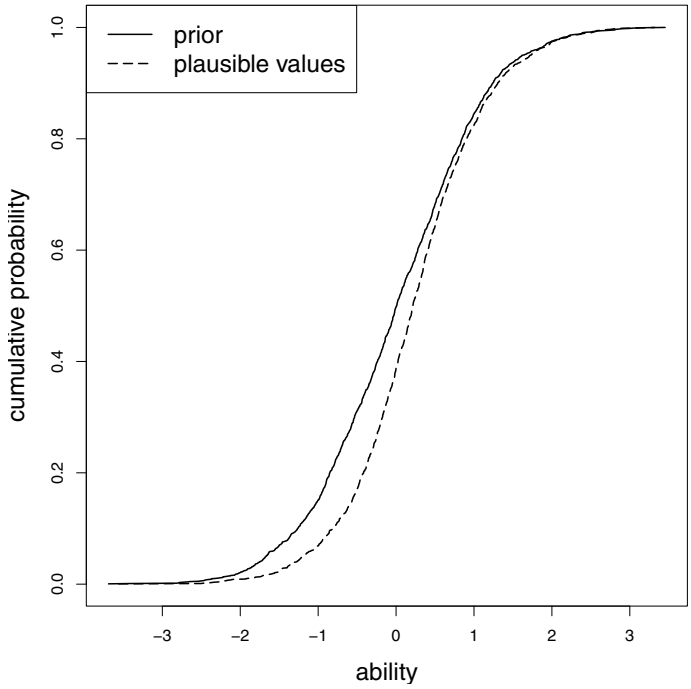


Figure 2.1: Ecdf of $\tilde{g}$ and N draws from the prior $g(\theta) = \phi(\theta)$ in PISA.

We agree with Kreiner and Christensen that the population model can be misspecified and, if so, the distribution of the plausible values can differ from the true distribution of ability. We do not agree, however, to the suggestion that there is no reason to be interested in the plausible value distribution if the population model is misspecified. The plausible value distribution provides a consistent estimate of the true ability distribution which is at least as "plausible" as the population model which figures as a prior. Furthermore, we can evaluate the fit of the population model by testing the hypothesis $H_0 : \tilde{g} = g$ against $H_1 : \tilde{g} \neq g$. If H_0 is rejected, there is no reason to be interested in g: $\tilde{g}$ is our best guess of what the true distribution of ability would look like.

Empirical example

We return to our PISA example to illustrate that we can test the hypothesis $H_0 : \tilde{g} \neq g$ against $H_1 : \tilde{g} \neq g$ using real-data with a relatively small number of observations, and that the power to reject this test is increasing with n. To illustrate the influence of the number of items n, we randomly assigned each student two items out of the 26 items that were available. Using the previously established IRT model and a standard normal prior distribution, we generated a single plausible value for each of the N persons in the sample. The ecdf of N draws from the prior distribution g, and the ecdf of generated plausible values are shown in Figure 2.2. It is clear that even with two items, the marginal distribution of plausible values

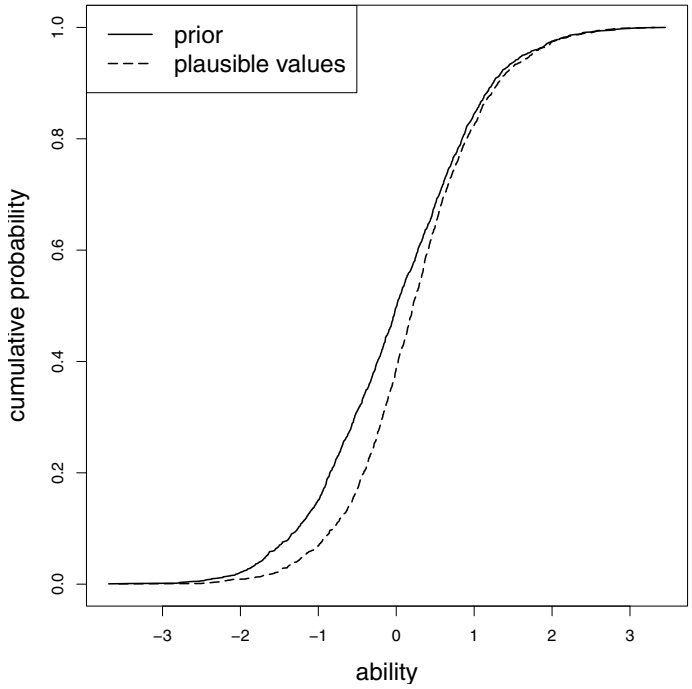


Figure 2.2: Ecdf of $\tilde{g}$ using $n = 2$ items, and N draws from the prior $g(\theta) = \phi(\theta)$ in PISA.

differs from the specified prior distribution and H_0 will be rejected using any test. Furthermore, from looking at Figure 2.1, where plausible values were generated using all 26 items, and Figure 2.2, using only two items, we see that the plausible values diverge from the prior distribution as n increases.

2.4.3 What if we missed a covariate?

In educational surveys, the chosen prior distribution is often conditionally Gaussian, conditioning on some covariates assessed during the survey. There is another misconception about plausible values, which has to do with analyses of plausible values using covariates that were not conditioned on in the generating model, i.e., covariates missing in g. For instance, Wu (2005) writes:

> If regression analyses are carried out with plausible values as dependent variables, and background variables as independent variables, then the "correct" regression coefficients will be "recovered", *if the model that produced the plausible values included the background variables.* If the model that produced the plausible values did not include the background variables as regressors, then the regression coefficients produced will be an under-estimate of the true regression coefficients. *[Page 125]*

This is, of course, partially true but depends on the answer to three questions; first, how large is the effect of the covariate on Θ, second, how many items are

in the study, and third, what prior distribution was used? When n is sufficiently large, the "correct" regression coefficient will be "recovered." Sufficiently large, here, relates to the effect size of the covariate on the distribution of Θ; that is, for small effects relatively few items are needed, and for large effects relatively many items are needed. Note that how large n needs to be in order to recover the correct coefficients using plausible values also relates to the prior distribution that is used. For instance, if the effect of a binary predictor is ignored in the analyses, then even for small n the correct coefficient will be recovered when the prior distribution is a mixture of two normal distributions.

If the prior distribution is a regression model in which a covariate is missing, then the exclusion of this covariate may lead to bias in parameter estimates for effects that are part of the model[2], or one might not observe that the missing covariate makes the unknown f skewed. This means that we run the risk of performing an incorrect inference about the unknown f using the prior distribution. It follows from our results that the marginal distribution of plausible values will always be a better estimate of f than the population model is in this situation, even if we do not recover the correct regression coefficient of the missing covariate.

Empirical example

To illustrate that we may recover the correct regression coefficients in practice, we return to the PISA example and look at the distribution of boys and girls in Canada who took booklet 6 using PISA's final student weights. We consider two prior distributions; a simple $\mathcal{N}(\mu, \sigma^2)$ prior distribution and a more complex $\mathcal{N}(\widehat{\mathbf{\Lambda}}\boldsymbol{\beta}, \sigma^2)$ prior distribution, where $\widehat{\mathbf{\Lambda}}$ constitute the principal component scores estimated on student covariates assessed in the PISA student questionnaire. In the latter, we used 50 principal components explaining roughly 60% of the variance in the student questionnaire. Note that in the latter, gender was included as a predictor (i.e. it was a covariate in the PCA), while in the former it was not included as a predictor. The previously established OPLM model was used, and we estimated hyper-parameters using the Gibbs sampler (Geman & Geman, 1984) with non-informative hyper-priors (Gelman, Carlin, Stern, & Rubin, 2004).

In Figure 2.3 we show the plausible value distributions of boys and girls weighted by the PISA student weights, using $\mathcal{N}(\widehat{\mathbf{\Lambda}}\boldsymbol{\beta}, \sigma^2)$ and using $\mathcal{N}(\mu, \sigma^2)$. From Figure 2.3 it is clear that the (weighted) distribution of plausible values under both priors are nearly indistinguishable, apart from some sampling error in the procedure.

Secondary analyses of plausible values usually involve linear regression type analyses, in which researchers try to identify the variables that have an effect on the latent variable distribution. From Figure 2.3 we see that the average ability of boys and the average ability of girls differ, and that girls perform better than boys. The weighted average ability for the boys was estimated at 0.180 and that of girls at 0.242. However, note that the distributions also differ in their standard deviations. The weighted standard deviation of ability for the boys was estimated

[2]The simplest example would be a prior where the mean is assumed to be equal to zero and one estimates the variance. If the true mean is not equal to zero, the variance estimate will be biased.

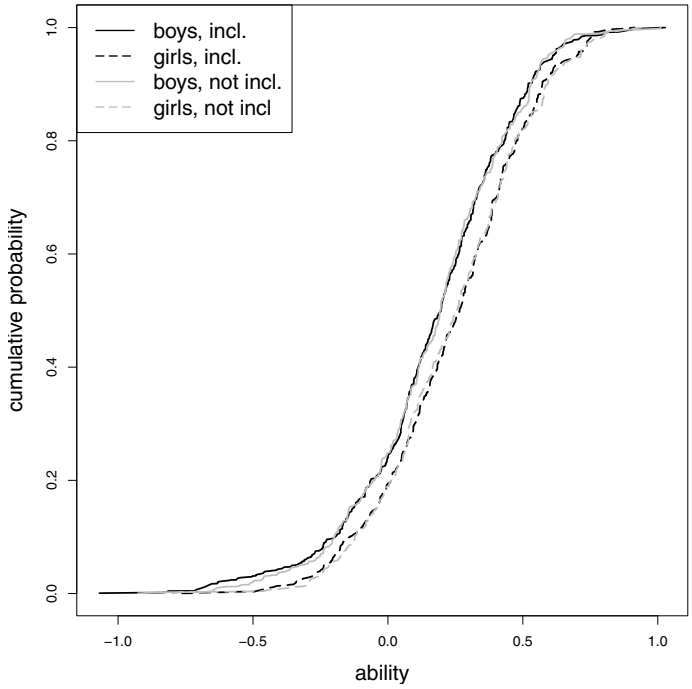


Figure 2.3: Plausible value distribution of boys and girls with and without including gender as covariate.

at 0.304 and that of the girls at 0.282. This is one particular example of why one should look at plausible values, since these particular differences in variances are not found in regression modelling when they are not explicitly modelled.

2.4.4 How to design analyses of educational surveys?

Lemma 1 states that EKL divergence is the difference between prior and marginal divergence. Hence, we can decrease EKL divergence by decreasing prior divergence and/or increasing marginal divergence. Decreasing prior divergence does not require that we have full knowledge about f. To wit, we can study the distribution of the plausible values found in previous cycles of a survey and use this to choose a prior that resembles f. When little or nothing is known about f, as, for instance, at the start of a survey, we may opt for a flexible prior; that is, one that easily adapts to different shapes. In each case, convergence is improved if we estimate the parameters of the prior so that it adapts itself to the data as much as possible. The Monotonicity theorem implies that adding items improves convergence but in practice there is a limit to the number of items that can be administered. The effect of adding items to increase marginal divergence was already illustrated in Figures 2.1 and 2.2, and is especially useful when we know very little about f.

Empirical example

We return to our PISA example to illustrate the effect of more flexible prior distributions on prior divergence, and consider the following increasingly more flexible prior distributions: a standard normal prior, a normal prior with a mean and a variance and the PCA regression prior. To illustrate the influence of n, we also manipulated the number of items in the analyses using the same procedure that we used before. To test the hypothesis $H_0 : \tilde{g} = g$ against $H_1 : \tilde{g} \neq g$ for the different prior distributions (given the number of items), we use the two-sample Kolmogorov-Smirnov (KS) test.

The previously established OPLM model was used, and we estimated hyper-parameters using the Gibbs sampler (Geman & Geman, 1984) with non-informative hyper-priors (Gelman et al., 2004). After convergence, we ran an additional $1,000$ iterations of the Gibbs sampler. In each of these additional $1,000$ iterations, we performed the following procedure; generate a plausible value for each person in the sample, generate a sample of size N from the prior, and compute the KS test statistic comparing both samples. In this manner we have $1,000$ replications for the test statistic and we show its average in Table 2.1 for the various conditions. Note that the results in Table 2.1 are significant at an α level of 0.05 whenever the corresponding test statistic is larger than 0.046.

The values in Table 2.1 are consistent with the result that prior divergence decreases as more flexible prior distributions are used. Note that neither the values using the $\mathcal{N}(\mu, \sigma^2)$ and the $\mathcal{N}(\widehat{\mathbf{\Lambda}}\boldsymbol{\beta}, \sigma^2)$ prior are significant, showing that there is not much to learn from $\mathcal{N}(\widehat{\mathbf{\Lambda}}\boldsymbol{\beta}, \sigma^2)$ about the marginal distribution that is not already known from $\mathcal{N}(\mu, \sigma^2)$.

Table 2.1: Average value of KS test statistic using PISA data to compare $\tilde{g}$ with the prior distributions used to generate $\tilde{g}$.

	$g(\theta) =$		
n	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{N}(\widehat{\mathbf{\Lambda}}\boldsymbol{\beta}, \sigma^2)$
26	0.161	0.034	0.026
20	0.256	0.032	0.026
14	0.277	0.032	0.026
8	0.290	0.030	0.026
2	0.292	0.029	0.028

Our main concern is whether or not the plausible value distribution converges to the true ability distribution. Since we do not know the true ability distribution, we compare our results with the best guess that we have, i.e. the distribution of plausible values obtained by using $n = 26$ items and the PCA regression prior. We repeated the procedure to obtain Table 2.1, but instead of comparing the generated plausible values with draws from the prior, we compare the generated plausible values with the plausible values generated using $n = 26$ items and the PCA regression prior. The results are shown in Table 2.2, and are consistent with the result that the plausible value distributions converge toward a single (true) distribution as n increases and the prior becomes more flexible.

Table 2.2: Average value of KS test statistic using PISA data to compare $\tilde{g}$ using different prior distributions with the best guess.

	$g(\theta)$		
n	$\mathcal{N}(0,1)$	$\mathcal{N}(\mu,\sigma^2)$	$\mathcal{N}(\widehat{\mathbf{\Lambda}}\boldsymbol{\beta},\sigma^2)$
26	0.052	0.021	- - -
20	0.064	0.022	0.022
14	0.088	0.024	0.024
8	0.133	0.030	0.027
2	0.235	0.065	0.109

Table 2.2 also allows us to illustrate a limitation of the use of flexible prior distributions that has to do with the counter-intuitive results for $n = 2$ and the $\mathcal{N}(\mu,\sigma^2)$ and $\mathcal{N}(\widehat{\mathbf{\Lambda}}\boldsymbol{\beta},\sigma^2)$ prior distributions. We would expect that the results for these two priors are in favour of the $\mathcal{N}(\widehat{\mathbf{\Lambda}}\boldsymbol{\beta},\sigma^2)$ prior, since we compare it with the marginal distribution of plausible values using $\mathcal{N}(\widehat{\mathbf{\Lambda}}\boldsymbol{\beta},\sigma^2)$, but with a different number of items. Instead, we find that the results are in favour of the $\mathcal{N}(\mu,\sigma^2)$ prior. The reason for this result has to do with the poor estimation of hyper-parameters when there is little information available, i.e., when both N and n are small. Note that, the first prior has just two parameters, μ and σ, but the second prior has 52 parameters, $\boldsymbol{\beta} = \{\beta_0, \beta_1, ..., \beta_{50}\}$ and σ^2. Thus, we expect that in the latter parameters are estimated more poorly than in the former, with the parameters' posterior standard deviations (or standard errors) being much larger. Due to this, there are large variations in the generated plausible value distributions from iteration to iteration, and we consequently observe larger deviations. This explains the counter-intuitive results, and shows that there is a limit to the amount of parameters that we can estimate, and thus the amount of flexibility that we can achieve in practice.

2.5 Discussion

In this chapter we have proved that plausible values are a non-parametric and consistent estimator of the distribution of ability in the population, and that convergence is *monotone* in an embedding in which the number of items tends to infinity. In plain words, this implies that we can use plausible values to learn about the true distribution of ability in the population. We have used this result to clear up some of the misconceptions about plausible values, and also to show how they can be used in the analyses of educational surveys. Thus far, plausible values have been used in educational surveys merely to simplify secondary analyses. Our result suggests that the distribution of plausible values should play the leading role, and that the population model is merely a vehicle to produce plausible values.

The population model is properly seen as a prior and the consistency of the plausible value distribution as an estimator of the true distribution is essentially due to the common result that the data overrule the prior when the number of observations increases. We have demonstrated that convergence of the plausible value distribution to the true distribution of ability can be improved if we estimate the parameters $\boldsymbol{\lambda}$ of the prior distribution, but it is unclear whether it makes sense to interpret the estimates $\widehat{\boldsymbol{\lambda}}$, especially when the prior distribution is misspecified. Technically, as N tends to infinity, $\widehat{\boldsymbol{\lambda}}$ are the parameter values that minimize prior divergence under the prior w.r.t. the true ability distribution (White, 1982). However, when the prior distribution is misspecified and prior divergence is not zero, the result of White (1982) does not tell us how wrong our conclusions are when inference is based on $\widehat{\boldsymbol{\lambda}}$.

In closing, we mention a limitation of our results. Our results imply that *if* Θ can be consistently estimated (i.e. its posterior is consistent), *then* the marginal distribution of plausible values converges to the unknown f. For models where the "if" part is resolved, our results apply.

Chapter 3

Composition Algorithms for Conditional Distributions

3.1 Introduction

In applied statistics we frequently have to sample from a conditional distribution. For example, in Bayesian inference where we need a sample from a posterior distribution, in Gibbs sampling (Geman & Geman, 1984) where samples from full conditional distributions are required, or in time series analyses when we wish to predict conditional upon the current state. This chapter is about two recently published algorithms designed for this problem: A rejection algorithm that was mentioned by Rubin (1984) and was recently applied in the *European Survey on Language Competences* (ESLC; Maris, 2012), and the *Single-Variable Exchange* (SVE) algorithm developed by Murray, Ghahramani, and MacKay (2012).

Both algorithms are based on the observation that a sample from a conditional distribution can be obtained from samples drawn from the joint distribution. The practical significance of this observation lies in the fact that sampling from the joint distribution is often easier because it can be done in two ways. To wit, the joint density of X and Y can be factored in two ways:

$$f(x|y)f(y) = f(y|x)f(x),$$

and to obtain a sample from the joint distribution we can use the *method of composition* (Tanner, 1993) and sample from $f(y)$ and then from $f(x|y)$, or sample from $f(x)$ and then from $f(y|x)$. Thus, if it is difficult to sample from $f(x|y)$, we can try to sample from $f(y|x)$, or vice versa. For instance, if we encounter a posterior distribution that is highly intractable, we can sample from it by generating data. Thus, the algorithms are extremely useful when it is difficult to sample from the posterior but easy to generate data as is the case for *Item Response Theory (IRT)* models. As both algorithms use composition to sample from the joint distribution we refer to them as *composition algorithms*. The algorithms differ in the way

Adapted from: Marsman, M., Maris, G., Bechger, T. & Glas, C. (2014). Composition Algorithms for Conditional Distributions. Submitted for publication.

they select observations from the joint distribution to obtain a sample from the conditional distribution of interest.

In this chapter, we use the composition algorithms to sample from conditional distributions of the following form:

$$f_r(\theta|\mathbf{x}_r) \propto f(\mathbf{x}_r|\theta)f_r(\theta) \tag{3.1}$$

where Θ is a random-effect that varies across replications $r = 1, \ldots, n$. Our main objective is to demonstrate how the composition algorithms can be tailored for the situation where n is very large. Over the last decade, large values of n have become increasingly more common as more and more data are being produced. This implies that there is a growing need to analyse large data sets and our algorithms are specifically designed for this purpose; mainly because their efficiency increases with n. The algorithms are not developed for situations where n is small.

The algorithms are useful in many contexts. In this chapter, we focus on applications in educational measurement where $\mathbf{X}$ is a vector of discrete item responses[2], Θ is a latent ability, $P(X|\theta)$ an IRT model with fixed item parameters, and we use the composition algorithms to sample from the posterior distribution of ability for each of n persons; either for its one right or as part of a Gibbs sampler. Compared to alternative approaches, the main advantage of the composition algorithms is that they become more efficient when the number of persons increases, as explained in Section 3 below.

The composition algorithms only require that we can generate data which is trivial for common IRT models. A nice feature is that we only need to know $f(\theta)$ and $P(X|\theta)$ up to a constant. This opens the door to new applications which would be difficult to handle with existing algorithms. We will illustrate this with an example involving a random-effects Gamma model for response times. The normalizing constant (i.e., the Gamma function) is not available in closed-form and often difficult to approximate.

To set the stage, we will first introduce the two composition algorithms as they stand. After having introduced the composition algorithms, we explain how they can be made more efficient, and illustrate their use with simulated and real-data applications. The chapter ends with a discussion.

3.2 Sampling from a conditional distribution

3.2.1 The Rejection Algorithm

The rejection algorithm (see Algorithm 1) works as follows. To sample from a conditional distribution $f(\theta|\mathbf{x})$, we repeatedly sample $\{\theta^*, \mathbf{x}^*\}$ from the joint distribution of θ and $\mathbf{x}$ until we produce a sample for which $\mathbf{x}^* = \mathbf{x}$. This generates an i.i.d. sample from the conditional distribution $f(\theta|\mathbf{x})$. The algorithm requires two things: First, it must be possible to sample from $f(\theta)$ and $P(\mathbf{x}|\theta)$; that is, we should be able to generate data under the model. Second, the random variable $\mathbf{X}$ must be discrete with a finite range so that there is a non-zero probability to generate a value $\mathbf{x}^*$ equal to the observed value $\mathbf{x}$.

[2]The responses are allowed to be continuous in the SVE algorithm, and we use this to sample from posteriors of the form $f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)f(\theta)$ in the examples section.

Algorithm 1 A rejection algorithm for $f(\theta|\mathbf{x})$

1: **repeat**
2: Generate $\theta^* \sim f(\theta)$
3: Generate $\mathbf{x}^* \sim P(\mathbf{x}|\theta^*)$
4: **until** $\mathbf{x}^* = \mathbf{x}$
5: Set $\theta = \theta^*$

The number of trials needed increases with the number of values $\mathbf{X}$ can assume and the rejection algorithm is only useful when this number is small. In the special case when $P(\mathbf{x}|\theta)$ belongs to the exponential family, the posterior depends on the data only via the sufficient statistic $t(\mathbf{x})$ (Dawid, 1979). Since $\mathbf{X}$ is a discrete random variable, $t(\mathbf{X})$ is also a discrete random variable, and this means that we may replace $\mathbf{x}^* = \mathbf{x}$ with $t(\mathbf{x}^*) = t(\mathbf{x})$ in line 4 of Algorithm 1. This version of the rejection algorithm was the one used in the ESLC (Maris, 2012) and in the present chapter. Note that, the more realizations of $\mathbf{X}$ lead to the same value on the sufficient statistic, the more efficient the algorithm becomes. The ESLC shows that the algorithm is efficient enough to be used in large-scale educational surveys using the *Partial Credit Model* (PCM; Masters, 1982). No doubt, the same holds for other exponential family IRT models, such as the *Rasch model* (Rasch, 1960), the *One Parameter Logistic Model* (OPLM; Verhelst & Glas, 1995), and special cases of the *Generalized Partial Credit Model* (GPCM; Muraki, 1992) and *Nominal Response Model* (NRM; Bock, 1972) where the category parameters are integer.

3.2.2 The Single Variable Exchange Algorithm

The rejection algorithm rejects all samples for which $\mathbf{x}^*$ does not exactly match $\mathbf{x}$, and thus requires the random variable $\mathbf{X}$ to be *discrete*, preferably assuming a small number of values. To allow $\mathbf{X}$ to be continuous, we adapt the rejection step such that we accept or reject samples with a probability other than zero or one. That is, we consider the generated θ^* as a sample from the *proposal distribution* $f(\theta|\mathbf{x}^*)$ and accept this value as a realization from the *target distribution* $f(\theta|\mathbf{x})$ with a probability $f(\theta^*|\mathbf{x})/Mf(\theta^*|\mathbf{x}^*)$, where $M > 0$ is an appropriate bound on $f(\theta^*|\mathbf{x})/f(\theta^*|\mathbf{x}^*)$ for all possible values of $\mathbf{x}$ and $\mathbf{x}^*$. In general, however, it is difficult to find M and we therefore consider a Metropolis algorithm. That is, we choose the probability to accept such that the accepted values are a sample from a Markov chain whose stationary distribution is $f(\theta|\mathbf{x})$. The price to pay is that we now produce a *dependent and identically distributed (d.i.d.)* sample.

To ensure that the Markov chain generated by the Metropolis algorithm has the desired stationary distribution, the following detailed balance condition must hold (Tierney, 1994):

$$\pi(\theta' \to \theta^*)\frac{P(\mathbf{x}|\theta')f(\theta')}{P(\mathbf{x})}\frac{P(\mathbf{x}^*|\theta^*)f(\theta^*)}{P(\mathbf{x}^*)} = \pi(\theta^* \to \theta')\frac{P(\mathbf{x}|\theta^*)f(\theta^*)}{P(\mathbf{x})}\frac{P(\mathbf{x}^*|\theta')f(\theta')}{P(\mathbf{x}^*)},$$

where θ' is the current parameter setting, and $\pi(\theta' \to \theta^*)$ the probability to make a transition of θ' to θ^*. It is easily checked that the detailed balance condition

holds when $\pi(\theta' \to \theta^*) = \min\{1, \alpha(\theta' \to \theta^*)\}$, with

$$\alpha(\theta' \to \theta^*) = \frac{P(\mathbf{x}|\theta^*)f(\theta^*)P(\mathbf{x}^*|\theta')f(\theta')}{P(\mathbf{x}|\theta')f(\theta')P(\mathbf{x}^*|\theta^*)f(\theta^*)} = \frac{P(\mathbf{x}|\theta^*)P(\mathbf{x}^*|\theta')}{P(\mathbf{x}|\theta')P(\mathbf{x}^*|\theta^*)}, \tag{3.2}$$

and the probability to accept θ^* depends on the relative likelihood to observe $\mathbf{x}^*$ and $\mathbf{x}$ given the parameter settings θ' or θ^*, respectively. Using this probability in the Metropolis algorithm, we arrive at the Single-Variable Exchange (SVE) algorithm developed by Murray et al. (2012) (see Algorithm 2).

Algorithm 2 The Single-Variable Exchange algorithm.

1: Draw $\theta^* \sim f(\theta)$
2: Draw $\mathbf{x}^* \sim P(\mathbf{x}|\theta^*)$
3: Draw $u \sim \mathcal{U}(0, 1)$
4: **if** $(u < \pi(\theta' \to \theta^*))$ **then**
5: $\quad \theta' = \theta^*$
6: **end if**

To use the SVE algorithm we must be able to compute $\alpha(\theta' \to \theta^*)$ and the SVE algorithm was designed to make this task as simple as possible. To see this, we write

$$P(\mathbf{x}|\theta) = \frac{h(\mathbf{x};\theta)}{Z(\theta)},$$

where $Z(\theta) = \sum_{\mathbf{x}} h(\mathbf{x};\theta)$ is a normalizing constant, or partition function, which is often difficult or even impossible to compute[3]. Since $\alpha(\theta' \to \theta^*)$ in (3.2) is the product of likelihood ratios, it follows that

$$\alpha(\theta' \to \theta^*) = \frac{\frac{h(\mathbf{x};\theta^*)}{Z(\theta^*)}\frac{h(\mathbf{x}^*;\theta')}{Z(\theta')}}{\frac{h(\mathbf{x};\theta')}{Z(\theta')}\frac{h(\mathbf{x}^*;\theta^*)}{Z(\theta^*)}} = \frac{h(\mathbf{x};\theta^*)h(\mathbf{x}^*;\theta')}{h(\mathbf{x};\theta')h(\mathbf{x}^*;\theta^*)}.$$

Thus, there is no need to compute $Z(\theta)$ (or $P(\mathbf{x})$).

As an illustration, Table 3.1 gives $\ln(\alpha(\theta' \to \theta^*))$ for a selection of IRT models. Note that for many of the models in Table 3.1, $\ln(\alpha(\theta' \to \theta^*))$ is of the form:

$$(\theta^* - \theta')(t(\mathbf{x}) - t(\mathbf{x}^*)).$$

That is, the acceptance probability depends on the product of the difference in parameter settings and the difference between the statistics of the generated and observed data. It also shows that, as the range of $t(\mathbf{X})$ increases, $\alpha(\theta' \to \theta^*)$ tends to become lower, on average.

[3]When both $Z(\theta)$ and $P(\mathbf{x})$ are difficult or even impossible to compute, the posterior distribution is called doubly-intractable. Murray et al. (2012) specifically developed the SVE algorithm for these doubly-intractable distributions.

Table 3.1: $\ln(\alpha(\theta' \to \theta^*))$ for a selection of IRT Models.

IRT Model	$\ln(\alpha(\theta' \to \theta^*))$	$t()$
Rasch	$(\theta^* - \theta')(t(\mathbf{x}) - t(\mathbf{x}^*)$	$\sum_i x_i$
2PL	$(\theta^* - \theta')(t(\mathbf{x}, \mathbf{a}) - t(\mathbf{x}^*, \mathbf{a}))$	$\sum_i a_i x_i$
3PL	$\sum_i (x_i - x_i^*) \ln\left(\frac{c_i + \exp(a_i(\theta^* - b_i))}{c_i + \exp(a_i(\theta' - b_i))}\right)$	
1PNO	$\sum_i (x_i - x_i^*) \ln\left(\frac{\Phi(\theta^* - b_i)(1 - \Phi(\theta' - b_i))}{\Phi(\theta' - b_i)(1 - \Phi(\theta^* - b_i))}\right)$	
2PNO	$\sum_i (x_i - x_i^*) \ln\left(\frac{\Phi(a_i\theta^* - b_i)(1 - \Phi(a_i\theta' - b_i))}{\Phi(a_i\theta' - b_i)(1 - \Phi(a_i\theta^* - b_i))}\right)$	
3PNO	$\sum_i (x_i - x_i^*) \left[\ln\left(\frac{c_i + (1-c_i)\Phi(a_i\theta^* - b_i)}{c_i + (1-c_i)\Phi(a_i\theta' - b_i)}\right) + \ln\left(\frac{1 - \Phi(a_i\theta' - b_i)}{1 - \Phi(a_i\theta^* - b_i)}\right)\right]$	
PCM	$(\theta^* - \theta')(t(\mathbf{x}) - t(\mathbf{x}^*)$	$\sum_i \sum_j x_{ij}$
GPCM	$(\theta^* - \theta')(t(\mathbf{x}, \mathbf{a}) - t(\mathbf{x}^*, \mathbf{a}))$	$\sum_i a_i \sum_j x_{ij}$
NRM	$(\theta^* - \theta')(t(\mathbf{x}, \mathbf{a}) - t(\mathbf{x}^*, \mathbf{a}))$	$\sum_i \sum_j a_{ij} x_{ij}$
MD2PL	$(\boldsymbol{\theta}^* - \boldsymbol{\theta}')^T(\mathbf{t}(\mathbf{x}, \underline{\mathbf{a}}) - \mathbf{t}(\mathbf{x}^*, \underline{\mathbf{a}}))$	$\sum_i x_i \boldsymbol{a}_i$

The abbreviations 2PL and 3PL stand for the Two- and Three-Parameter Logistic models, 1PNO, 2PNO and 3PNO for the One-, Two-, and Three-Parameter Normal Ogive models and MD2PL for the Multidimensional Two-Parameter Logistic model. We used $\Phi(x)$ as shorthand for $\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)dy$.

3.2.3 Limitations

In educational measurement we often have to sample from the posterior ability distribution of each of n persons, where n is large. To sample from each of the n posteriors, the composition algorithms would require about n times the amount of work required to sample from a single posterior, see below. Thus, the algorithms do not become more efficient when n increases, and are inefficient when n is large. The algorithms are also inefficient for applications with many items. If the number of possible response patterns (or sufficient statistics) increases, the rejection algorithm will need increasingly more trials and the SVE algorithm will tend to have lower acceptance probabilities so that the correlation between successive draws will tend to be higher.

We illustrate this with a small simulation study, the results of which are shown in Figures 3.1 and 3.2. We simulate data with n persons answering to each of k dichotomous items, with n varying between 100 and 10,000, and $k \in \{10, 20, 30\}$. We assume a standard normal distribution for ability Θ. For the rejection algorithm, the IRT model is the Rasch model. For the SVE algorithm, we use the *Two-Parameter Logistic* (2PL) model. The item parameters are fixed, with difficulty parameters sampled from a standard normal distribution, and discrimination parameters sampled uniformly between 1 and 3. For each combination of n and k we generated 100 data sets. With the item parameters fixed, our goal is to sample for each of the n persons an ability from the posterior distribution given his or her observed response pattern.

Results for the rejection algorithm are in Figure 3.1, which shows the average number of trials that are required to sample from each of the n posteriors as a function of n and k. It is clear that the average number of trials required quickly stabilizes around the number of possible realizations of $t(\mathbf{X})$, which is $k+1$ in this simulation[4]. Thus, we need approximately $(k+1) \times n$ iterations to produce a value from each of the n posteriors, and this number grows linear in both n and k.

Results for the SVE algorithm are in Figure 3.2 which shows the average proportion of values accepted in the 100th iteration of the algorithm as a function of n and k. The acceptance probabilities are seen to be low, and decreasing with an increase of the number of items. Thus, for both algorithms it follows that as n and k grow, we need more iterations to obtain a certain amount of independent replicates from each of the n posteriors. We conclude that the algorithms, as they stand, are unsuited for applications with large n (and k).

3.3 Large-Scale Composition Sampling

The rejection and SVE algorithm sample from one posterior at the time. Consequently, sampling from n posteriors requires n times the amount of work needed to sample from a single posterior. If the algorithms are to be prepared for applications with an increasing number of posteriors, the amount of work per posterior has to decrease with n. To see how, observe that both algorithms generate samples that are not used efficiently, i.e., samples that are either rejected or accepted with a low probability. Thus, to improve the efficiency of the algorithms for increasing n, we need to make more efficient use of the generated samples. To this aim, we consider the SVE algorithm as an instance of what Tierney (1994, 1998) refers to as a *mixture of transition kernels*. This way of looking at the SVE algorithm suggests two approaches to improve its efficiency. One of these will be seen to apply to the rejection algorithm as well.

3.3.1 A Mixture Representation of the SVE algorithm

In every realization of the SVE algorithm, we sample one of the possible response patterns (denoted $\mathbf{x}^*$), together with a random value for ability (denoted θ^*). The sampled ability value is a sample from the posterior distribution $f(\theta|\mathbf{x}^*)$ which is the proposal distribution in the SVE algorithm. The probability that we use $f(\theta|\mathbf{x}^*)$ as proposal distribution in the SVE algorithm is equal to $P(\mathbf{x}^*)$, which follows from the factorization:

$$P(\mathbf{x}^*|\theta^*)f(\theta^*) = f(\theta^*|\mathbf{x}^*)P(\mathbf{x}^*).$$

[4]The number of trials $W = w$ required to generate a realization $t(\mathbf{x})$ follows a geometric distribution with parameter $P(t(\mathbf{x}))$; the (marginal) probability to generate $t(\mathbf{x})$ under the model. From this we see that $\mathbb{E}(W|t(\mathbf{x}))$ equals $P(t(\mathbf{x}))^{-1}$ and

$$\mathbb{E}(W) = \sum_{t(\mathbf{x})} \mathbb{E}(W|t(\mathbf{x}))P(t(\mathbf{x})),$$

where the sum is taken over all possible realizations. It follows that $\mathbb{E}(W)$ equals the number of possible realizations of $t(\mathbf{X})$.

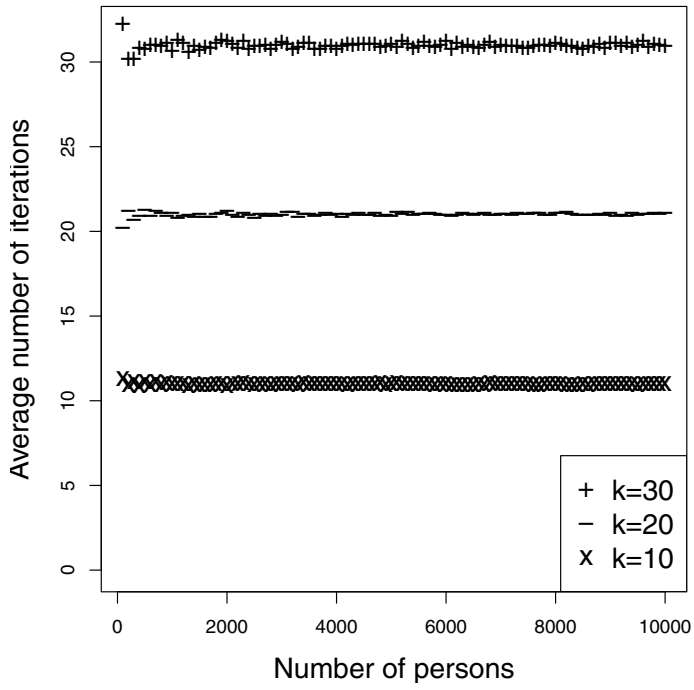


Figure 3.1: Average number of trials required for rejection.

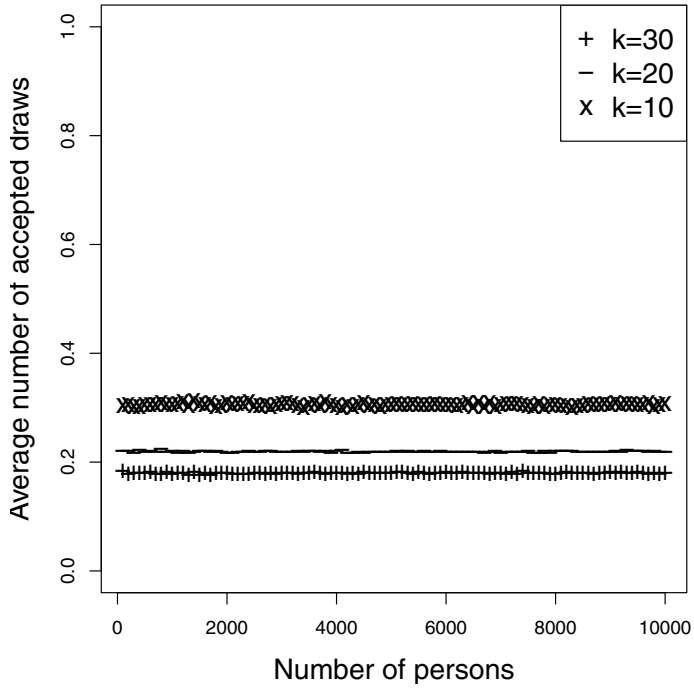


Figure 3.2: Average proportion of accepted values with SVE.

That is, every simulated response pattern corresponds to a unique proposal distribution, and hence, to a unique transition kernel $f(\theta^*|\theta, \mathbf{x}^*)$. Each of these transition kernels has the target posterior distribution as its invariant distribution; that is,

$$f(\theta^*|\mathbf{x}) = \int_{\mathbb{R}} f(\theta^*|\theta, \mathbf{x}^*) f(\theta|\mathbf{x}) d\theta.$$

As shown by Tierney (1994), the same is true for their mixture, that is,

$$f(\theta^*|\mathbf{x}) = \int_{-\infty}^{\infty} \sum_{\mathbf{x}^*} f(\theta^*|\theta, \mathbf{x}^*) P(\mathbf{x}^*) f(\theta|\mathbf{x}) d\theta,$$

where the sum is taken over all possible response patterns, and we now see that the $P(\mathbf{x}^*)$ are the mixture weights.

To make matters concrete, consider the posterior distribution for a Rasch model with k items, and a standard normal prior for ability θ. Because the Rasch model is an exponential family model with the test score $t(\mathbf{x})$ as sufficient statistic for ability, we know that posteriors for the different ways to obtain the same test score are all the same (Dawid, 1979). That is, the mixture weights are nothing but the distribution of test scores. Moreover, the posterior distributions $f(\theta|t(\mathbf{x}))$ are stochastically ordered by the test score, which makes the acceptance probability lower, the larger the difference between the value of $t(\mathbf{x})$ conditioned on in the target and $t(\mathbf{x}^*)$ conditioned on in the proposal distribution, see Table 3.1. Figure 3.3 shows the mixture probabilities $P(t(\mathbf{x}))$ for a test of 20 items. We see in Figure 3.3 that the SVE algorithm will tend to generate many transition kernels for which the acceptance probability is not very high.

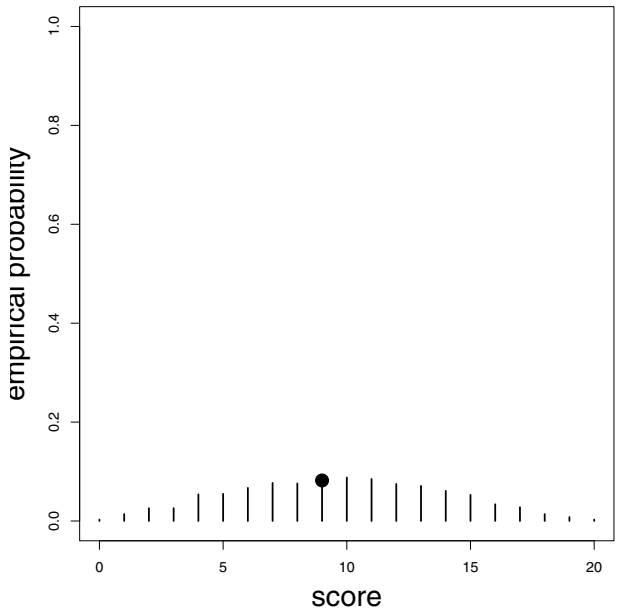


Figure 3.3: Empirical distribution over transition kernels for the SVE algorithm.

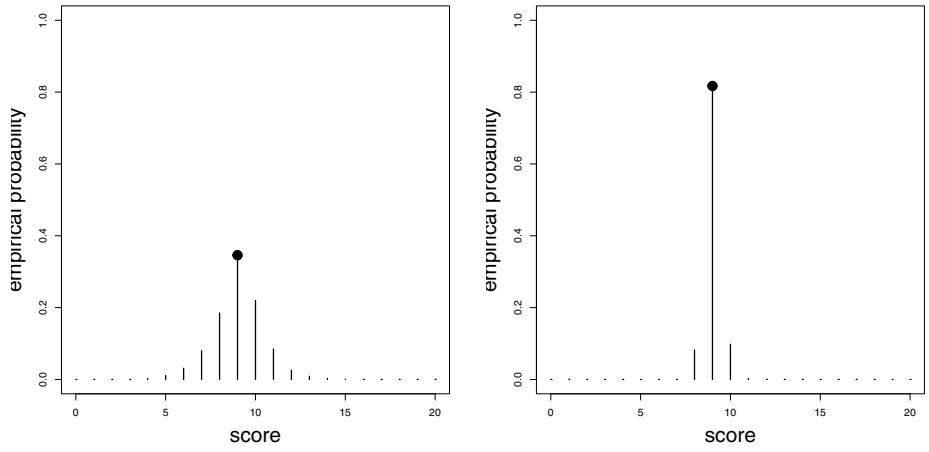
3.3.2 Oversampling

Since the SVE algorithm tends to frequently generate transition kernels for which the acceptance probability is low, we consider changing the mixture probabilities, in such a way that more probability mass is concentrated on transition kernels with high acceptance probability.

Suppose that instead of simulating a single proposal value θ^*, with a corresponding single response pattern $\mathbf{x}^*$, we simulate a number of i.i.d. proposal values, each with its own response pattern. From these, we choose the one for which the test score is closest to the test score conditioned on in the target distribution, and hence the acceptance rate tends to be the highest.

In Figure 3.4 we illustrate the effectiveness of this oversampling approach in sampling from a posterior $f(\theta|t(\mathbf{x}) = 9)$. Clearly, even with 5 samples we already improve the probability to generate directly from the target from close to 0.1 to close to 0.4. With 20 samples, this probability even exceeds 0.8. Moreover, if the proposal is not identical to the target, it is increasingly more likely to be close to the target as the number of samples increases.

Since oversampling can easily be implemented in a parallel implementation, this approach need not lead to a large increase in computer time. This makes the approach computationally attractive.



(a) 5 samples. (b) 20 samples.

Figure 3.4: Probability distribution over transition kernels after modulating the mixture probabilities.

3.3.3 Matching

Consider the situation where there are many proposal distributions (i.e. n large), and hence many target posterior distributions, each one independent from the others. The SVE algorithm can once again be considered as a mixture of transition

kernels for the whole collection of n posteriors:

$$f(\boldsymbol{\theta}^*|\underline{\mathbf{x}}) = \int_{\mathbb{R}^n} \prod_p f(\theta_p^*|\theta_p, \mathbf{x}_p^*) P(\underline{\mathbf{x}}^*) f(\boldsymbol{\theta}|\underline{\mathbf{x}}) d\boldsymbol{\theta},$$

where $\underline{\mathbf{x}}$ denotes the matrix $\underline{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Observe that the transition kernel for person p only depends on $\underline{\mathbf{x}}^*$ via the p-th response pattern. Suppose that for a matrix $\underline{\mathbf{x}}^*$, we permute the person indices p, in some fixed way (denoted perm(p)). Then, the transition kernel for person p depends on $\underline{\mathbf{x}}^*$ via one of the response patterns in $\underline{\mathbf{x}}^*$, and every response pattern is used exactly once:

$$f(\boldsymbol{\theta}^*|\underline{\mathbf{x}}) = \int_{\mathbb{R}^n} \prod_p f(\theta^*_{\text{perm}(p)}|\theta_p, \mathbf{x}^*_{\text{perm}(p)}) P(\underline{\mathbf{x}}^*) f(\boldsymbol{\theta}|\underline{\mathbf{x}}) d\boldsymbol{\theta}.$$

Clearly, not all proposal distributions lead to the same acceptance probability, and thus, not all permutations lead to the same overall acceptance rate. Hence, some permutations work better than others. Notice that all permutations lead to a valid transition kernel with the posterior distribution as its invariant distribution, as long as our permutation strategy does not depend on $\boldsymbol{\Theta}'$ or $\boldsymbol{\Theta}^*$. Finding, for every matrix $\underline{\mathbf{x}}^*$ and every observed matrix $\underline{\mathbf{x}}$ the best permutation, will in general be an NP-complete problem. However, the better the permutation, the more efficient the algorithm.

In Algorithm 3, we consider the general situation where each person may receive its own prior distribution, and we denote the prior of a person p with $f_p(\theta)$. We generate a proposal using the prior $q = 1, \ldots, n$ (q now indexes the proposals), and we reorder the index vector $\mathbf{Q} = [q_p]$ of the proposals by using a *permutation function* perm(). When we use $\theta_q^* \sim f_q(\theta|\mathbf{x}_q^*)$ as a proposal for a posterior $f_p(\theta|\mathbf{x}_p)$ (p need not equal q), then we accept θ_q^* with probability $\pi(\theta_p' \to \theta_q^*) = \min\{1, \alpha(\theta_p' \to \theta_q^*)\}$, and

$$\alpha(\theta_p' \to \theta_q^*) = \frac{f_p(\theta^*|\mathbf{x}_p) f_q(\theta'|\mathbf{x}^*)}{f_p(\theta'|\mathbf{x}_p) f_q(\theta^*|\mathbf{x}^*)} = \frac{h(\mathbf{x}_p;\theta^*) h(\mathbf{x}^*;\theta')}{h(\mathbf{x}_p;\theta') h(\mathbf{x}^*;\theta^*)} \times \frac{f_p(\theta^*) f_q(\theta')}{f_p(\theta') f_q(\theta^*)},$$

a product of likelihood ratios times a product of prior ratios, where the normalizing constants $P(\mathbf{x})$ and $Z(\theta)$ cancel as before (as do the normalizing constants of the prior distributions).

Simple permutation functions are often readily available. For instance, the test score is usually correlated with Θ, and gives a simple procedure to permute the indices of proposals and targets. When the IRT model is a member of the exponential family, the sufficient statistic $t(\mathbf{x})$ contains all information about Θ from the data, and gives another simple procedure for permutation. More general solutions would be the use of Maximum Likelihood or Bayes' Modal estimates, when they are not too expensive to compute. We give some examples of permutation strategies in our applications below.

3.3.4 Recycling in the rejection algorithm

The main idea underlying matching is that a proposal need not be associated to one particular posterior. We can use the same idea for the rejection algorithm

Algorithm 3 Single-Variable Exchange algorithm with matching.

Require: Index vector $\mathbf{Q} = [q_p] = p$, for $p = 1, 2, ..., n$
Require: A permutation function perm()
for $q = 1$ to n **do**
  Generate $\theta_q^* \sim f_q(\theta)$
  Generate $\mathbf{x}_q^* \sim P(\mathbf{X}|\theta_q^*)$
end for
Match proposals to targets by rearranging $\mathbf{Q}$ based on perm().
for $p = 1$ to n **do**
  Set $q = q_p$
  Draw $u \sim \mathcal{U}(0, 1)$
  if $(u < \pi(\theta_p' \rightarrow \theta_q^*))$ **then**
    Set $\theta_p' = \theta_q^*$
  end if
end for

for the situation with n posteriors using a common prior $f(\theta)$. The idea behind recycling is that if we sample $\{\theta^*, \mathbf{x}^*\}$, θ^* can be assigned to *any* observation p where $t(\mathbf{x}_p) = t(\mathbf{x}^*)$ (or $\mathbf{x}_p = \mathbf{x}^*$). In general, we need to sample from $n = \sum_{j=1}^{J} n_j$ posteriors $f(\theta|t(\mathbf{x}) = t_j)$, where t_j is one of the J unique values the statistic $t(\mathbf{X})$ can take, n_j is the number of observations of response patterns $\mathbf{x}_p$ for which $t(\mathbf{x}_p) = t_j$, and it is arbitrary how the values of $t(\mathbf{X})$ are indexed. As seen in Algorithm 4, we sample from the joint distribution of Θ and $\mathbf{X}$ until we have n_j values for each j. In Algorithm 4 we store generated values in a vector $\mathbf{R}$ and the index corresponding to the generated statistic in a vector $\mathbf{S}$. If necessary, we can use $\mathbf{S}$ to assign the drawn parameters to observations.

Algorithm 4 A rejection algorithm with recycling.

Require: n_j for $j = 1, 2, ..., J$.
Require: A counter c and vectors $\mathbf{R} = [r_i]$ and $\mathbf{S} = [s_i]$, $i = 1, 2, ..., n$.
$c = 0$.
repeat
  Generate $\theta^* \sim f(\theta)$
  Generate $\mathbf{x}^* \sim P(\mathbf{X}|\theta^*)$
  Determine j, such that $t(\mathbf{x}^*) = t_j$
  if $n_j \geq 1$ **then**
    $n_j = n_j - 1$
    $c = c + 1$
    $[r_c] = \theta^*$
    $[s_c] = j$
  end if
until $n_j = 0$ for $j = 1, ..., J$

In the context of IRT, the situation with n posteriors using a common prior describes the situation of n persons sampled from the same population. In practice, however, we often encounter situations where the persons are sampled from

different groups, e.g., boys and girls. In this situation, posteriors are of the form

$$f(\theta|\mathbf{x}_p) \propto P(\mathbf{x}_p|\theta)f_m(\theta),$$

i.e., persons are grouped into marginals m, where $f_m(\theta)$ denotes the prior distribution in marginal m, and recycling applies to each marginal separately. It is clear that in this situation, the algorithm becomes efficient only when there are many persons in each marginal. When the prior distributions are person specific, and each person has its own marginal distribution, *recycling* reduces to the standard rejection algorithm.

3.3.5 Has the efficiency of the algorithms improved?

We considered *recycling* and *matching* as ways to improve the rejection and SVE algorithm when samples are required from many posteriors. To illustrate that this works, we compare the efficiency of the rejection algorithm with and without recycling and the SVE algorithm with and without matching under the conditions of our previous simulation.

Results for the rejection algorithm with *recycling* are in Figure 3.5, which shows the average number of trials required to sample from the n posteriors as a function of n and k. If we compare the results in Figure 3.5 with the results in Figure 3.1, we see that recycling requires relatively few iterations per posterior. Note that the required number of iterations decreases as n increases and increases when k increases. It is clear from Figure 3.5 that as both n and k increase, recycling makes the rejection algorithm more efficient when n increases faster than k. For fixed k, Figure 3.5 confirms that as n becomes large the number of iterations per posterior tends to 1.

To illustrate that the *matching* procedure improves the efficiency of the SVE algorithm, we consider the following simple strategy. We order target distributions using the statistic $t(\mathbf{x}_p, \mathbf{a}) = \sum_{i=1}^{k} x_{pi}a_i$ (see Table 3.1), such that the values of the statistic are ordered from small to large, and we do the same for the proposal distributions using the $t(\mathbf{x}^*, \mathbf{a})$. This simple permutation strategy ensures that if the Markov chain is stationary, the first proposal is likely to be a good proposal for the first target (since the difference between $t(\mathbf{x}, \mathbf{a})$ and $t(\mathbf{x}^*, \mathbf{a})$ is likely to be small), and the same holds for the second, the third, and so on. Results for the SVE algorithm using this procedure are given in Figure 3.6, which shows the average acceptance rate in the 100th iteration of the algorithm as a function of n and k. If we compare the results in Figure 3.6 with the results in Figure 3.2, we see that matching results in much higher acceptance rates. Note that, similar to the results for the recycling, the proportion of accepted values increase as n increases and decrease as k increases, and matching makes the SVE algorithm more efficient when n increases faster than k. For fixed k, Figure 3.6 confirms that as n becomes large the average acceptance rate tends to 1.

We conclude that recycling and matching make sampling from a large number of posteriors entirely feasible. Most appealing is that the efficiency improves as a function of n. As n tends to infinity this means that we need to generate the data only once to obtain a draw from each of n posteriors, and both algorithms generate i.i.d. from each of the n posteriors. For moderate n we can already see

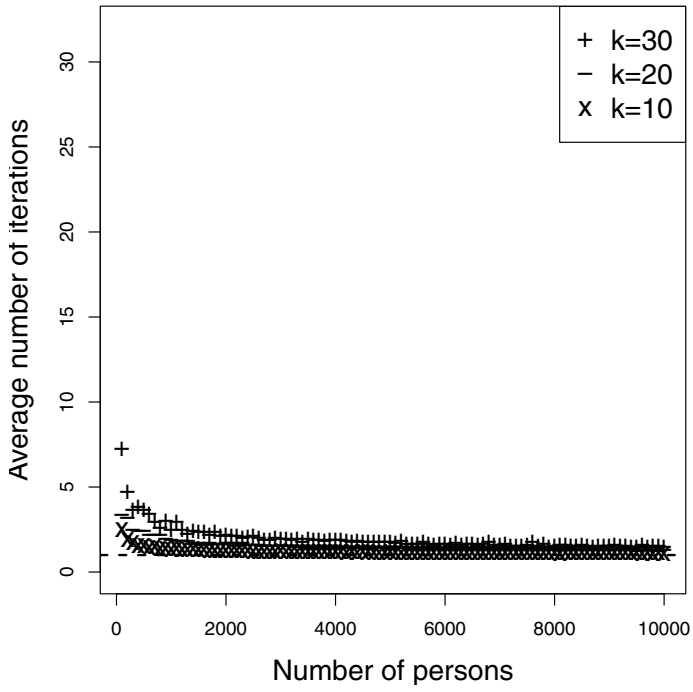


Figure 3.5: Average number of trials required for recycling.

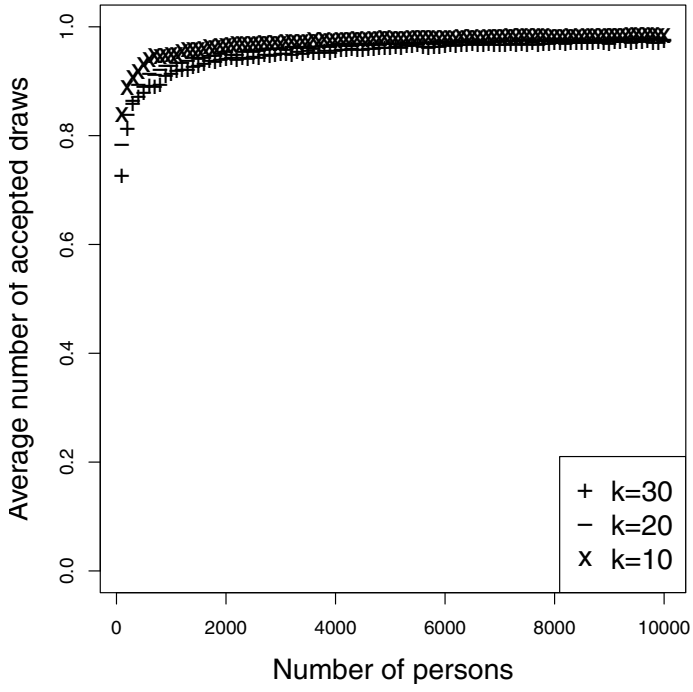


Figure 3.6: Average proportion of accepted values with matching.

that the number of trials needed for the rejection algorithm approaches one, and that the acceptance rates of the SVE algorithm approaches one. This shows that, even for moderate n both algorithms require little more than one generated data set, and that the SVE algorithm is close to sampling i.i.d.

To illustrate that matching makes the autocorrelation in the SVE algorithm a decreasing function of n, we perform a small simulation. We run $5,000$ Markov chains for 500 iterations each. We use the $5,000$ Markov chains to estimate the autocorrelation by correlating the $5,000$ draws in some iteration i and iteration $i+1$, $i+2$, ... Figure 3.7 shows the autocorrelation spectra for the SVE algorithm with matching. In Figure 3.7 we see that the autocorrelations are a decreasing function of n, meaning that as n becomes sufficiently large, we sample approximately i.i.d.

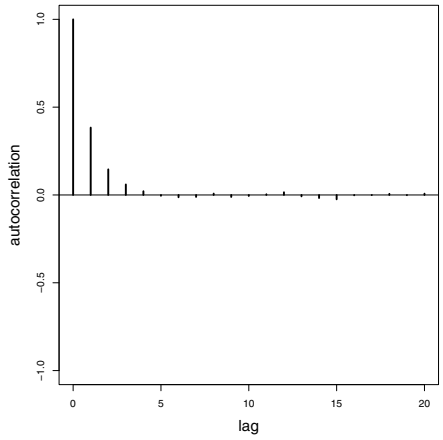
3.3.6 Comparison with existing algorithms

When it is difficult to sample from $f(\theta|\mathbf{x})$ directly, it is sometimes easier to sample from a more complex (augmented) posterior distribution $f(\theta, \mathbf{y}|\mathbf{x})$ using the Gibbs sampler. In the context of educational measurement, this approach has been advocated by Albert (1992) for Normal Ogive models. Due to the use of conditioning in the Gibbs sampler, the *data augmentation* procedure of Albert (1992) introduces a constant amount of autocorrelation to the Markov chain (Liu, Wong, & Kong, 1994). As a result, the number of iterations that are required to obtain a fixed amount of independent replicates from each of the n posteriors is linear in n. In this sense our algorithms scale better, since the amount of autocorrelation reduces as a function of n.

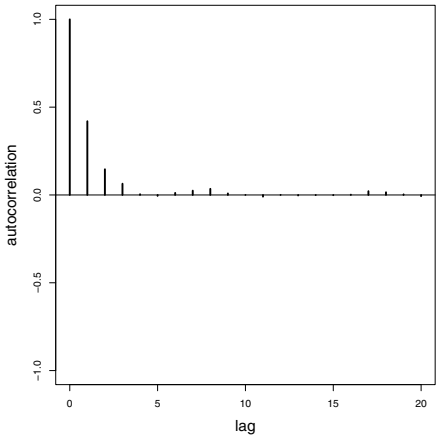
A more general approach to sampling from $f(\theta|\mathbf{x})$ is to sample a proposal value θ^* from a conditional distribution $f(\theta^*|\theta')$ and use the Metropolis-Hastings algorithm to either move to the proposed value θ^* or stay at the current state θ'. This approach has been advocated by Patz and Junker (1999b), who suggest to use $f(\theta^*|\theta') = \mathcal{N}(\theta', \sigma^2)$ as proposal distribution (i.e. a *random walk*). Setting the value of σ^2 in the proposal distribution requires some effort from the user (Rosenthal, 2011): when σ^2 is too large most samples are rejected, but when σ^2 is too small only small steps are taken and the chain does not mix properly. To overcome this problem altogether, one could use an unconditional proposal distribution $g(\theta)$ (i.e. an *independence chain*). This is the approach we took in this chapter. Whenever the proposal distribution $g(\theta)$ closely resembles the target distribution, the Metropolis-Hastings algorithm is very efficient. In general, it can be difficult to find good proposal distributions, but the matching procedure automatically finds proposal distributions $g(\theta|\mathbf{x}^*)$ that closely resemble the target $f(\theta|\mathbf{x})$, and as n increases, this procedure becomes more likely to generate good proposal distributions.

3.4 Simulated and Real-data examples

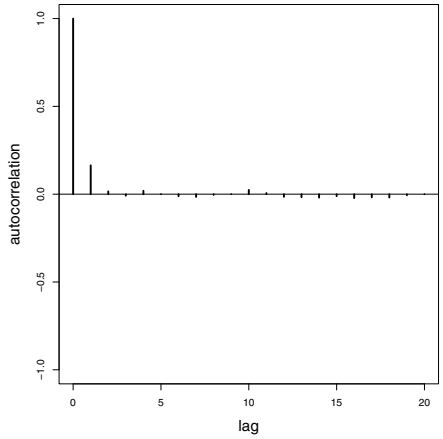
In this section, we discuss three examples illustrating the practical use of the SVE algorithm for Bayesian estimation using the Gibbs sampler. The Gibbs sampler is an abstract divide-and-conquer algorithm that generates a dependent sample from a multivariate posterior distribution. In each iteration, the algorithm generates a



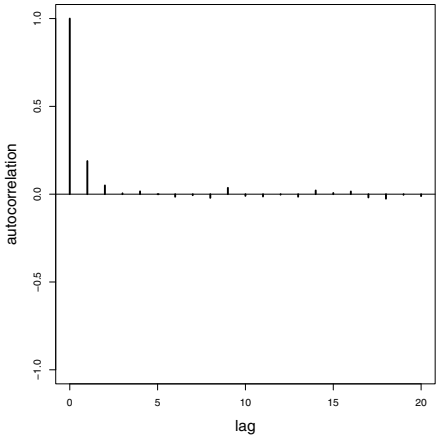
(a) $n = 100$ persons, $\theta = 0$



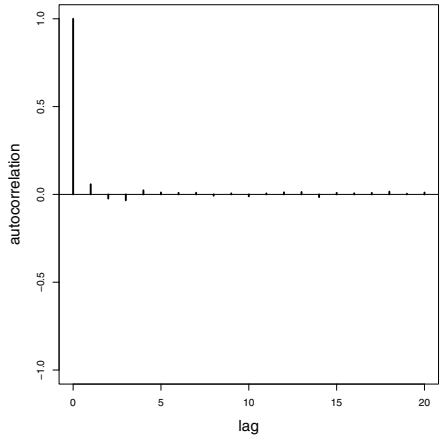
(b) $n = 100$ persons, $\theta = 0.5$



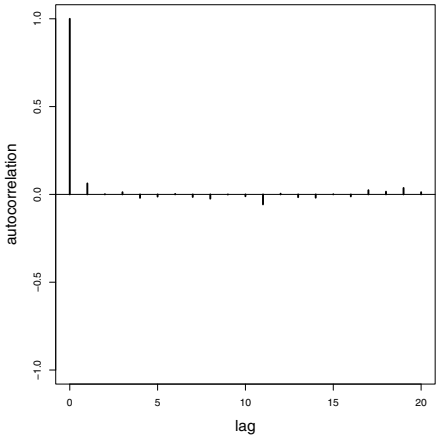
(c) $n = 1,000$ persons, $\theta = 0$



(d) $n = 1,000$ persons, $\theta = 0.5$



(e) $n = 10,000$ persons, $\theta = 0$



(f) $n = 10,000$ persons, $\theta = 0.5$

Figure 3.7: Estimated autocorrelation spectra using $k = 30$ items.

sample from the distribution of each variable in turn, conditional on the current values of the other variables. These are called the *full-conditional distributions*. It can be shown that the sequence of samples constitutes a Markov chain, and the stationary distribution of that Markov chain is the joint posterior distribution of interest.

In each of our examples, there will be one or more full-conditional distributions that are not easily sampled from and we use the SVE algorithms developed in this chapter to sample from these full-conditional distributions. Note that all analysis have been performed using GNU-R-3.0.1 (R Core Team, 2014) on an ordinary computer. More specifically, we have used a Dell OptiPlex 980 PC with an Intel Core 5 CPU and clock speed 3.20 Ghz and 4Gb of memory running on Windows 7 Enterprise(32 bit) with a single core.

3.4.1 Gamma regression

The random-effects gamma model is a model for responses times proposed by Fox (2013) as an alternative to the log-normal model that is commonly used (van der Linden, 2007; Klein Entink, Fox, & van der Linden, 2009). The model is difficult to estimate, because the normalizing constant of the gamma distribution (i.e. the gamma function $\Gamma(\cdot)$) is not available in closed-form and can produce overflow errors in its computation. We develop a Gibbs sampler for this model to illustrate how the SVE algorithm can be used to avoid the calculation of the gamma function.

Let X_{pi} denote the time needed by person p to respond to item i; $p = 1, \ldots, n$, and $i = 1, \ldots, k$. The X_{pi} are assumed to be independent, gamma distributed random variables with

$$f(\underline{\mathbf{x}}|\boldsymbol{\lambda}, \boldsymbol{\eta}) = \prod_{p=1}^{n} \prod_{i=1}^{k} \frac{\lambda_{pi}^{\eta_{pi}}}{\Gamma(\eta_{pi})} x_{pi}^{\eta_{pi}-1} \exp\left\{-x_{pi}\lambda_{pi}\right\}.$$

Fox (2013) used $\lambda_{pi} = \nu/2\theta_p$ and $\eta_{pi} = \nu/2$. We will use a slight alteration in this simulation, with $\lambda_{pi} = \nu/\theta_p\delta_i$ and $\eta_{pi} = \nu$, such that $E[X_{pi}] = \theta_p\delta_i$, and $Var(X_{pi}) = E[X_{pi}]^2/\nu$. The person parameter $\theta_p > 0$ represents the speed of person p, the item parameter $\delta_i > 0$ the time intensity of item i, and ν a common rate parameter. We further assume that $\theta_p \sim \ln\mathcal{N}(\mu_\theta, \sigma_\theta^2)$, and $\delta_i \sim \ln\mathcal{N}(\mu_\delta, \sigma_\delta^2)$, where $\ln\mathcal{N}(\mu, \sigma^2)$ denotes the log-normal distribution with mean μ and variance σ^2. The location and scale parameters of the person and item parameters are unknown and are to be estimated. To complete the specification of the model we use the following priors: $\nu \sim \Gamma(a, b)$, $f(\mu_\theta, \sigma_\theta^2) \propto \sigma_\theta^{-2}$ and $f(\mu_\delta, \sigma_\delta^2) \propto \sigma_\delta^{-2}$.

Given the person and item parameters, the location and scale parameters are easily sampled from their full-conditional distributions (Gelman et al., 2004):

$$f(\mu_\theta|\boldsymbol{\theta}, \sigma_\theta^2) \propto \mathcal{N}\left(\frac{1}{n}\sum_{p=1}^{n}\ln(\theta_p), \sigma_\theta^2/n\right)$$

$$f(\sigma_\theta^2|\boldsymbol{\theta}) \propto \text{Inv-}\chi^2\left(n-1, \frac{1}{n-1}\sum_{p=1}^{n}\left(\ln(\theta_p) - \frac{1}{n}\sum_{p=1}^{n}\ln(\theta_p)\right)^2\right)$$

$$f(\mu_\delta|\boldsymbol{\delta},\sigma_\delta^2) \propto \mathcal{N}\left(\frac{1}{k}\sum_{i=1}^{k}\ln(\delta_i), \sigma_\delta^2/k\right)$$
$$f(\sigma_\delta^2|\boldsymbol{\delta}) \propto \text{Inv-}\chi^2\left(k-1, \frac{1}{k-1}\sum_{i=1}^{k}\left(\ln(\delta_i) - \frac{1}{k}\sum_{i=1}^{k}\ln(\delta_i)\right)^2\right).$$

The full-conditional distribution of ν, the person and the item parameters, however, are not easily sampled from, and for these we will use the SVE algorithms developed in this chapter.

To sample from the full-conditional distribution of ν, we generate ν^* from the prior $f(\nu|a,b)$ and generate a data matrix $\underline{\mathbf{x}}^*$ from $f(\underline{\mathbf{x}}|\boldsymbol{\theta},\boldsymbol{\delta},\nu^*)$. The probability $\pi(\nu' \rightarrow \nu^*)$ to make a transition from ν^* to ν' using this set-up is then equal to $\min\{1, \alpha(\nu' \rightarrow \nu^*)\}$, with

$$\ln\alpha(\nu' \rightarrow \nu^*) = (\nu^* - \nu')(t(\underline{\mathbf{x}},\boldsymbol{\theta},\boldsymbol{\delta}) - t(\underline{\mathbf{x}}^*,\boldsymbol{\theta},\boldsymbol{\delta})),$$

where

$$t(\underline{\mathbf{x}},\boldsymbol{\theta},\boldsymbol{\delta}) = \sum_{p=1}^{n}\sum_{i=1}^{k}\left(\ln(x_{pi}) - \frac{x_{pi}}{\theta_p\delta_i}\right).$$

Note that we do not need to evaluate the $\Gamma()$ function at ν' or ν^*, making $\ln\alpha$ a relatively simple function to compute.

We have seen earlier that in this set-up the SVE algorithm is likely to generate transition kernels for which the acceptance probability is low. We therefore use the *oversampling* procedure. That is, we generate a number of i.i.d. proposal values ν^*, each with its own data matrix $\underline{\mathbf{x}}^*$. From these, we choose the one for which the statistic $t(\underline{\mathbf{x}}^*,\boldsymbol{\theta},\boldsymbol{\delta})$ is closest to $t(\underline{\mathbf{x}},\boldsymbol{\theta},\boldsymbol{\delta})$. We use 100 proposals in this example. The GNU-R code that we used for this full-conditional is given in Appendix B.

To sample from the full-conditional distributions of the person and the item parameters we use the matching procedure. Since we use the same matching procedure for the person and the item parameters, we only describe the procedure for the person parameters. We generate θ_q^*, $q = 1,\ldots,n$, from $f(\theta|\mu_\theta,\sigma_\theta^2)$ and use it to generate a vector of response times $\mathbf{x}_q^*$ from $f(\mathbf{x}|\theta_q^*,\boldsymbol{\delta},\nu)$. Say that we use $f(\theta|\mathbf{x}_q^*,\nu,\mu_\theta,\sigma_\theta)$ as proposal for a target p (p need not equal q), the probability $\pi(\theta_p' \rightarrow \theta_q^*)$ to make a transition from θ_p' to θ_q^* is then equal to $\min\left\{1, \alpha(\theta_p' \rightarrow \theta_q^*)\right\}$, with

$$\ln\alpha(\theta_p' \rightarrow \theta_q^*) = \nu\left(\frac{1}{\theta_q^*} - \frac{1}{\theta_p'}\right)\left(t(\mathbf{x}_q^*,\boldsymbol{\delta}) - t(\mathbf{x}_p,\boldsymbol{\delta})\right), \tag{3.3}$$

where

$$t(\mathbf{x}_p,\boldsymbol{\delta}) = \sum_{i=1}^{k}\frac{x_{pi}}{\delta_i}.$$

Note again, that we do not need the evaluate the $\Gamma()$ function in $\ln\alpha$, and the acceptance probabilities are simple to compute.

From (3.3) we see that it is opportune to use $t(\mathbf{x}_p,\boldsymbol{\delta})$ to permute proposals and targets. To this aim, we compute $t(\mathbf{x},\boldsymbol{\delta})$ for each person in the sample, and for each proposal. Then, we order the targets using the $t(\mathbf{x}_p,\boldsymbol{\delta})$, such that the

corresponding statistics are ordered from small to large, and do the same for the proposals using the $t(\mathbf{x}_q^*, \boldsymbol{\delta})$. This simple permutation strategy ensures that if the Markov chain is stationary, the first proposal is likely to be a good proposal for the first target (since the difference between $t(\mathbf{x}, \boldsymbol{\delta})$ and $t(\mathbf{x}^*, \boldsymbol{\delta})$ is likely to be small), and the same holds for the second, the third, and so on. The GNU-R code that we used for this full-conditional is given in Appendix C.

To see how it works, we simulated data for $n = 10,000$ persons on a test consisting of $k = 40$ items. We set the mean and variance of the person and the item parameters equal to 10 and 1, respectively, from which we can solve for the location and scale parameters in the log-normal model. Using these location and scale parameters, we sample the person and item parameters from the log-normal model. The parameter ν was set equal to 40.

We ran the Gibbs sampler for $2,000$ iterations, which took approximately 2.5 hours (about $4,7$ seconds per iteration). The main computational cost of this Gibbs sampler resides in sampling the entire $n \times k$ data matrix $j + 2$ times in each iteration, of which j times for sampling from the full-conditional of ν. Since the cost per iteration is the same in each iteration, we see that we need approximately 0.1 seconds to sample the person and the item parameters in each iteration, and approximately 4.6 seconds to sample ν. This means that we can reduce the computational time by reducing j. Note, however, that this would also reduce the acceptance rate in sampling ν.

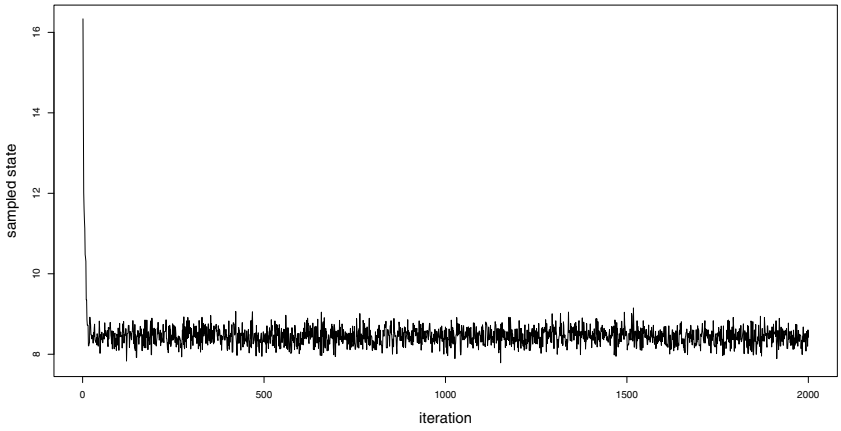
The results are in Figures 3.8 and 3.9. As expected, our use of the SVE algorithm does not lead to high acceptance rates for the item parameters; the average acceptance rate was 0.05. The main reason is that we only generate 40 proposals to assign to 40 targets, with a large variation on the conditioning statistic $t(\mathbf{x}, \boldsymbol{\theta})$ due to the large number of observations. In the next example we show that the oversampling procedure can be used to remedy this. We did obtain high efficiency for the person parameters, with an average acceptance rate of 0.96. In Figure 3.8 we show the trace plot for a person and an item parameter. It is clear that both converge quickly to the stationary distribution. In Figure 3.9 we show scatter plots of the true person and item parameters against the parameter states in iteration $2,000$, which illustrates that we are able to recover the parameters of the generating model. Finally, the proportion of accepted values for the ν parameter equalled 0.30, which is certainly reasonable for such a complex full-conditional distribution. In Figure 3.8(c) we show the trace plot of ν, from which we see that once the person and item parameters converge, ν also quickly converges to its stationary distribution.

3.4.2 The Amsterdam Chess Test data

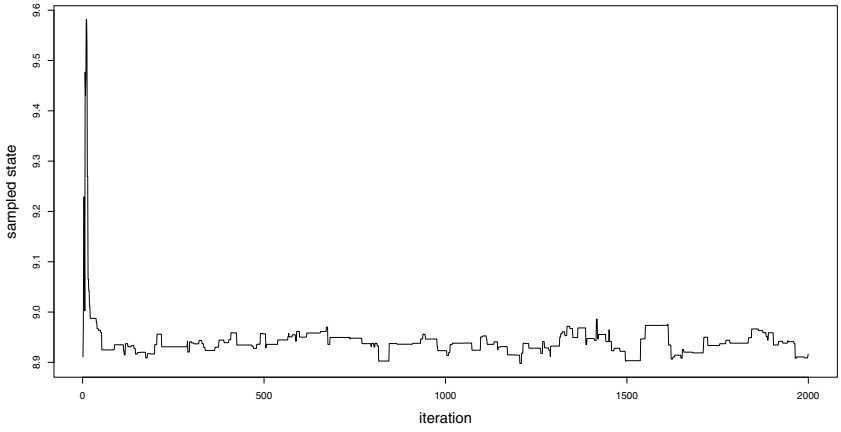
The Signed Residual Time (SRT) model is an exponential family IRT model for item response accuracy and response times, and is derived by Maris and van der Maas (2012) from the following scoring rule,

$$(2X_{pi} - 1)(d - S_{pi}),$$

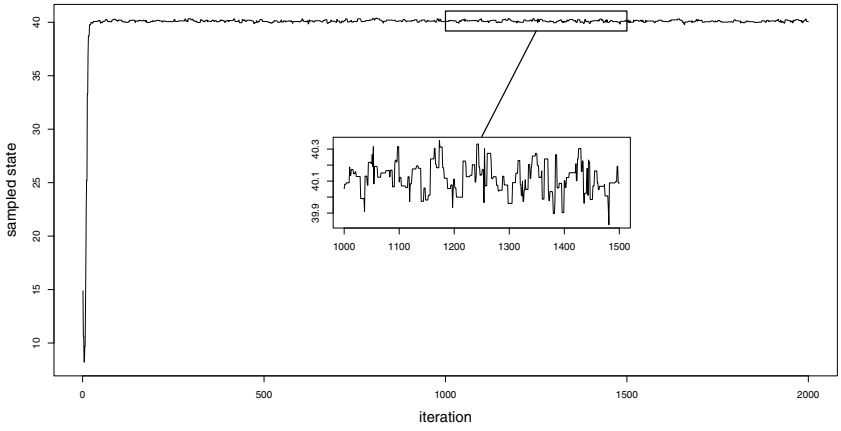
for an item response X_{pi}, which equals one if the response is correct and zero if incorrect, after S_{pi} time units when the time limit for responding is d. This scoring



(a) Trace plot of a person parameter.

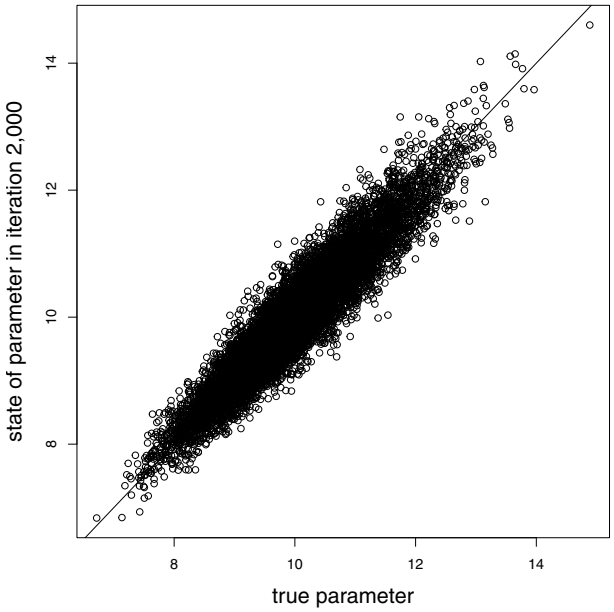


(b) Trace plot of an item parameter.

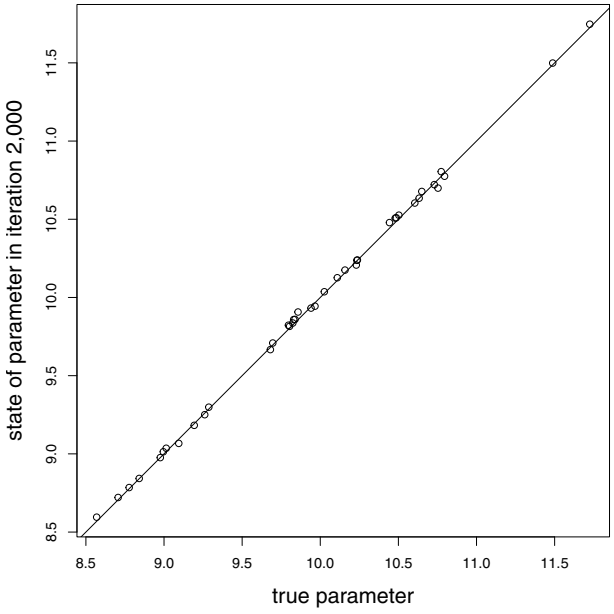


(c) Trace plot of ν.

Figure 3.8: Trace plot of ν, a person and an item parameter in the Gamma mixture example.



(a) Scatterplot of the person parameters.



(b) Scatterplot of the item parameters.

Figure 3.9: Scatterplot of the true person (item) parameters at the states of the person (item) parameters in iteration 2,000 of the Gibbs sampler for the Gamma mixture example.

rule assigns the residual time as the score for a correct response, and minus the residual time for an incorrect response. Thus, subjects need to be both fast and accurate to obtain a high score and, thereby, a high estimated ability. The SRT model is:

$$f(X_{pi} = x_{pi}, S_{pi} = s_{pi}|\theta_p, \delta_i, d) = (\theta_p - \delta_i)\frac{\exp\left[(2x_{pi} - 1)(d - s_{pi})(\theta_p - \delta_i)\right]}{\exp\left[d(\theta_p - \delta_i)\right] - \exp\left[-d(\theta_p - \delta_i)\right]},$$

for $0 \leq s \leq d$. The statistics

$$\begin{aligned} t(\mathbf{x}_p, \mathbf{s}_p) &= \sum_{i=1}^{k}(2x_{pi} - 1)(d - s_{pi}) \\ t(\mathbf{x}_i, \mathbf{s}_i) &= -\sum_{p=1}^{n}(2x_{pi} - 1)(d - s_{pi}) \end{aligned} \tag{3.4}$$

are sufficient for the ability θ_p of a person p and the difficulty δ_i of an item i, respectively. We assume that $\theta_p \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$, and $\delta_i \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2)$, and to complete specification of the model used the following priors: $f(\mu_\theta, \sigma_\theta^2) \propto \sigma_\theta^{-2}$ and $f(\mu_\delta, \sigma_\delta^2) \propto \sigma_\delta^{-2}$.

Given the person and item parameters, the location and scale parameters are easily sampled from their full-conditional distributions (Gelman et al., 2004):

$$\begin{aligned} f(\mu_\theta|\boldsymbol{\theta}, \sigma_\theta^2) &\propto \mathcal{N}\left(\frac{1}{n}\sum_{p=1}^{n}\theta_p, \sigma_\theta^2/n\right) \\ f(\sigma_\theta^2|\boldsymbol{\theta}) &\propto \text{Inv-}\chi^2\left(n - 1, \frac{1}{n-1}\sum_{p=1}^{n}\left(\theta_p - \frac{1}{n}\sum_{p=1}^{n}\theta_p\right)^2\right) \\ f(\mu_\delta|\boldsymbol{\delta}, \sigma_\delta^2) &\propto \mathcal{N}\left(\frac{1}{k}\sum_{i=1}^{k}\delta_i, \sigma_\delta^2/k\right) \\ f(\sigma_\delta^2|\boldsymbol{\delta}) &\propto \text{Inv-}\chi^2\left(k - 1, \frac{1}{k-1}\sum_{i=1}^{k}\left(\delta_i - \frac{1}{k}\sum_{i=1}^{k}\delta_i\right)^2\right). \end{aligned}$$

The full-conditional distributions of the person and item parameters are not easily sampled from and we will use an SVE algorithm to sample from these full-conditional distributions. We use the same procedure to sample from the full-conditional distributions of the person and the item parameters, and we will only describe the procedure for the person parameters.

We generate θ_q^*, $q = 1, \ldots, n$, from $f(\theta|\mu_\theta, \sigma_\theta^2)$ and use it to generate a vector of item responses $\mathbf{x}_q^*$ and response times $\mathbf{s}_q$ from $f(\mathbf{x}, \mathbf{s}|\theta_q^*, \boldsymbol{\delta})$ (see Appendix D). Say that we use $f(\theta|\mathbf{x}_q^*, \mathbf{s}_q^*, \mu_\theta, \sigma_\theta)$ as proposal for a target p (p need not equal q), the probability $\pi(\theta_p' \rightarrow \theta_q^*)$ to make a transition from θ_p' to θ_q^* is then equal to $\min\left\{1, \alpha(\theta_p' \rightarrow \theta_q^*)\right\}$, with

$$\ln \alpha(\theta'_p \to \theta^*_q) = \left(\theta^*_q - \theta'_p\right)\left(t(\mathbf{x}^*_q, \mathbf{s}^*_q) - t(\mathbf{x}_p, \mathbf{s}_p)\right),$$

and $t(\mathbf{x}_p, \mathbf{s}_p)$ defined in (3.4).

Although the sufficient statistics (3.4) can be used to permute the indices of targets and proposals, we only have a few person and item parameters in this example. To obtain some efficiency of the SVE algorithm in this application, we use a variant of the *oversampling* strategy. In each iteration, we generate a number of i.i.d. proposals, and for each target distribution choose the proposal for which the statistic $t(\mathbf{x}^*, \mathbf{s}^*)$ is closest to the observed statistic $t(\mathbf{x}, \mathbf{s})$, while ensuring that each proposal is used only once.

Van der Maas and Wagenmakers (2005) describe data from the Amsterdam Chess Test (ACT), collected during the 1998 open Dutch championship in Dieren, the Netherlands. The data we consider consists of the accuracy and response times of $n = 259$ subjects on $k = 80$ choose-a-move items administered with a time limit of 30 seconds.

We started the mean and variance of the person and item parameters at zero and one, respectively. Using these values we sampled the person and item parameters from the prior. In each iteration, we generated $2 \times n = 498$ proposals for the persons and $5 \times k = 400$ proposals for the items. We ran the Gibbs sampler for 10,000 iterations, which took approximately 12 minutes (about 0.07 seconds per iteration). The average acceptance rate was 0.98 for the persons and 0.93 for the items.

An important advantage of this illustrative application is that for chess expertise an established external criterion is available in the form of the Elo ratings of chess players, which has high predictive power for game results. For those 225 participants for whom a reliable Elo rating was available, we correlated the *expected a posteriori* (EAP) estimates with their Elo ratings. The results are given in Figure 3.10. The correlation between EAP estimates and Elo ratings is equal to 0.822.

3.4.3 The 2012 Eindtoets data

In educational measurement, population models are commonly used to describe structure in the distribution of the latent abilities. For example, in equating two exams one can characterize the two exam groups by using a normal distribution with a group specific mean and variance, in the analyses of tests consisting of different scales a multivariate normal distribution can be used to characterize the latent correlations, and in educational surveys a normal regression model can be used to study the effects of covariates on the ability distribution. Whenever the abilities are observed, inference is relatively straightforward in each of these situations. Our focus in this section is to show how the SVE algorithm can be used to sample from the full-conditional distribution of the latent abilities, allowing the analyses of structural IRT models using the Gibbs sampler, even for large datasets.

We use response data of $n = 158,637$ Dutch end of primary school pupils on the 2012 Cito Eindtoets to illustrate our approach using a multidimensional IRT model. In specific, we used data from the non-verb spelling (10 items), verb spelling (10 items), reading comprehension (30 items), basic arithmetic (14 items),

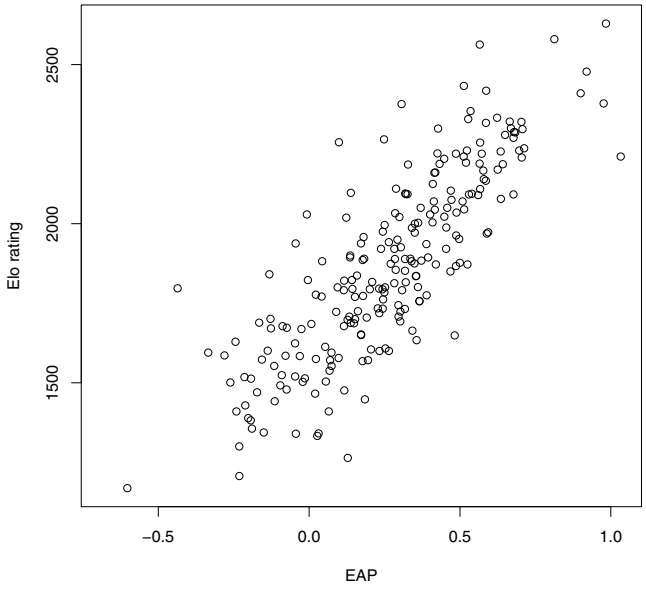


Figure 3.10: Scatter plot of EAP versus Elo rating in the ACT example.

fractions (20 items) and geometry (15 items) scales. That is, we have six unidimensional IRT models (a between multidimensional IRT model) and use a multivariate normal distribution to infer about the latent correlations between the six scales. To keep our focus on sampling the latent abilities, we assume that an IRT model is given (i.e. the parameters characterizing the items in the IRT model are known). For simplicity, we use the Rasch model for each of the scales in our example, and fix the item parameters at the conditional maximum likelihood (CML) estimates.

We use a multivariate normal distribution with an unknown $D \times 1$ vector of means $\boldsymbol{\mu}$ and $D \times D$ covariance matrix $\boldsymbol{\Sigma}$ to describe the latent correlations between the $D = 6$ dimensions. To complete the model, we use the multivariate Jeffreys prior for the mean vector and the covariance matrix:

$$f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{D+1}{2}}.$$

The Gibbs sampler is used to sample from the joint posterior distribution $f(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\underline{\mathbf{x}})$. For this model, the full-conditional distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are easily sampled from (Gelman et al., 2004):

$$\begin{aligned} f(\boldsymbol{\mu}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) &\propto \mathcal{N}_D(\bar{\boldsymbol{\theta}}, \boldsymbol{\Sigma}/n) \\ f(\boldsymbol{\Sigma}|\boldsymbol{\theta}) &\propto \text{Inverse-Wishart}_{n-1}(\mathbf{S}^{-1}) \end{aligned}$$

where $\bar{\boldsymbol{\theta}} = \frac{1}{n}\sum_{p=1}^{n} \boldsymbol{\theta}_p$ is the mean ability vector and $\mathbf{S} = \sum_{p=1}^{n}(\boldsymbol{\theta}_p - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_p - \bar{\boldsymbol{\theta}})^T$ the sums of squares matrix around the mean ability vector. The full-conditional

distributions $f(\boldsymbol{\theta}_p|\underline{\mathbf{x}}_p, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are intractable, however, and for this we use the SVE algorithm.

Instead of sampling from $f(\boldsymbol{\theta}_p|\mathbf{x}_p, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ directly, we sample pupil abilities in a dimension d given the $D-1$ other dimensions, for $d = 1, \ldots, D$. The full-conditional distribution for the ability of a pupil p in a dimension d is proportional to

$$f(\theta_{pd}|\mathbf{x}_{pd}, \boldsymbol{\theta}^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \prod_{i=1}^{k_d} \frac{\exp\{x_{pid}(\theta_{pd} - \delta_{id})\}}{1 + \exp\{\theta_{pd} - \delta_{id}\}} \exp\left\{-\frac{(\theta_{pd} - \lambda_{pd})^2}{2\eta_d^2}\right\},$$

where δ_{id} is the difficulty of the i-th out of k_d items in dimension d, $\boldsymbol{\theta}_p^{(d)}$ is the ability vector of pupil p excluding entry d, and λ_{pd} and η_d^2 are the conditional mean and variance of θ_{pd} given $\boldsymbol{\theta}_p^{(d)}$ in the population model, respectively, given by:

$$\lambda_{pd} = \mu_d + \boldsymbol{\sigma}_d^{(d)} \left(\boldsymbol{\Sigma}^{(d,d)}\right)^{-1} \left(\boldsymbol{\theta}_p^{(d)} - \boldsymbol{\mu}^{(d)}\right)$$
$$\eta_d^2 = \sigma_{dd} - \boldsymbol{\sigma}_d^{(d)} \left(\boldsymbol{\Sigma}^{(d,d)}\right)^{-1} (\boldsymbol{\sigma}_d^{(d)})^T,$$

where $\boldsymbol{\sigma}_d^{(d)}$ contains the off-diagonal elements of the d-th row d in $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{\sigma}_2^{(2)} = [\sigma_{21}, \sigma_{23}, \ldots, \sigma_{26}]$.

We sample from the full-conditionals $f(\theta_{pd}|\mathbf{x}_{pd}, \boldsymbol{\theta}^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, as follows. First, we compute λ_{pd} for $p = 1, \ldots, n$ (note that these depend on the abilities from the remaining $D-1$ dimensions). Then, we sample θ_{qd}^* from $\mathcal{N}(\lambda_{qd}, \eta_d^2)$ and use these to generate an item response vector $\mathbf{x}_{qd}^*$ from $P(\mathbf{X}_d|\theta_{qd}^*, \boldsymbol{\delta}_d)$, for $q = 1, \ldots, n$. Say that we use $f(\theta_{qd}|\mathbf{x}_{qd}^*, \boldsymbol{\theta}_q^{(d)}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ as proposal for a target p (p need not equal q), then the probability $\pi(\theta'_{pd} \rightarrow \theta_{qd}^*)$ to make a transition of θ'_{pd} to θ_{qd}^* is equal to $\min\{1, \alpha(\theta'_{pd} \rightarrow \theta_{qd}^*)\}$, with

$$\ln \alpha(\theta'_{pd} \rightarrow \theta_{qd}^*) = (\theta'_{pd} - \theta_{qd}^*)(t(\mathbf{x}_{qd}^*, \lambda_{qd}, \eta_d) - t(\mathbf{x}_{pd}, \lambda_{pd}, \eta_d)),$$

where,

$$t(\mathbf{x}_{pd}, \lambda_{pd}, \eta_d) = \sum_{i=1}^{k_d} x_{pid} + \lambda_{pd}/\eta_d^2.$$

Note that $t(\mathbf{x}_{pd}, \lambda_{pd}, \eta_d)$ combines information from the likelihood with information from the population model.

To match proposals to targets, it is opportune to use $t(\mathbf{x}_{pd}, \lambda_{pd}, \eta_d)$, since if $t(\mathbf{x}_{qd}^*, \lambda_{qd}, \eta_d)$ is close to $t(\mathbf{x}_{pd}, \lambda_{pd}, \eta_d)$ the acceptance probability tends to be high. In matching the n proposals to the n targets, we start with computing $t(\mathbf{x}_{pd}, \lambda_{pd}, \eta_d)$ for each target, and computing $t(\mathbf{x}_{qd}^*, \lambda_{qd}, \eta_d)$ for each proposal. Then, we order the targets using the $t(\mathbf{x}_{pd}, \lambda_{pd}, \eta_d)$, such that the corresponding statistics are ordered from small to large, and do the same for the proposals using the $t(\mathbf{x}_{qd}^*, \lambda_{qd}, \eta_d)$. If the Markov chain is stationary, the first proposal is likely to be a good proposal for the first target (since the difference between $t(\mathbf{x}, \lambda, \eta)$ and $t(\mathbf{x}^*, \lambda, \eta)$ will be small), and the same holds for the second, the third, and so on.

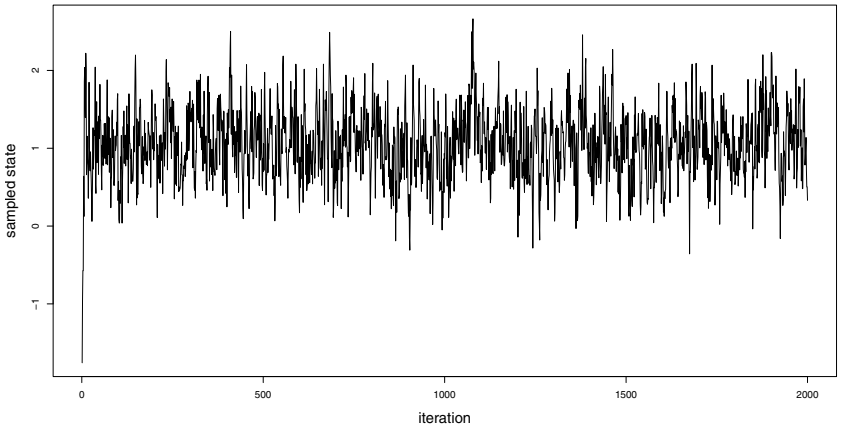
We start our analyses by setting $\boldsymbol{\mu}$ equal to $\mathbf{0}$ and $\boldsymbol{\Sigma}$ equal to the $D \times D$ identity matrix. To get reasonable starting values for the latent ability vectors, we performed a single run of the SVE algorithm where we accepted all proposals. We ran the Gibbs sampler for $2,000$ iterations, which took approximately 80 minutes (about 2.5 seconds per iteration). The acceptance rates of the SVE algorithm were high in this example, averaging to 0.98, 1.00, 0.97, 0.99, 0.99 and 1.00 for dimensions one to six, respectively. This means that we sample approximately i.i.d. from the full-conditional distributions of the abilities, and thus, using the SVE algorithm in this example does not introduce additional auto-correlation to the Markov chain.

Despite the observation that we sample the abilities approximately i.i.d. in this example, the amount of auto correlation in the chain is high. To illustrate, we show the trace plot for three parameters; an ability, a mean and a variance in Figure 3.11. Note the wave-like patterns that emerge, which indicate a strong relation between subsequent states in the Markov chain (i.e. high amount of autocorrelation). The reason for this high amount of autocorrelation is due to the high correlations that we obtain between some of the dimensions (see Table 3.2 below), and the fact that we sampled from each dimension conditional upon the others. The high correlations between dimensions then provide a strong relation between draws in subsequent iterations, inducing a high amount of autocorrelation.

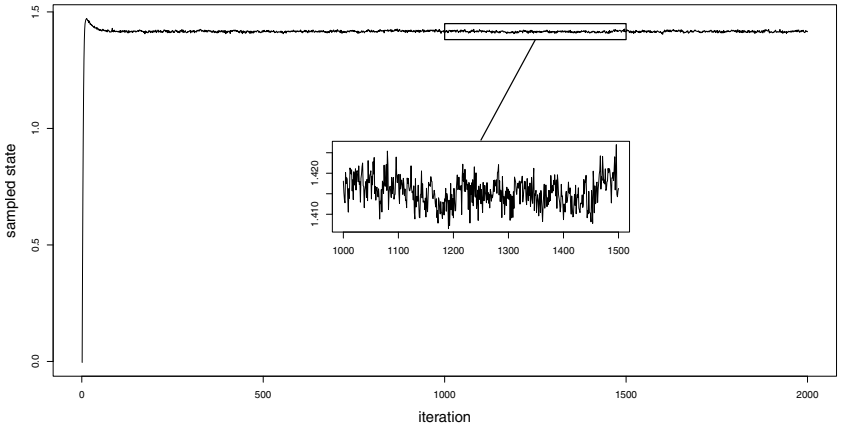
The estimated correlation matrix is shown in Table 3.2. From Table 3.2 it is seen that the two spelling scales are closely related, as are the three mathematics scales. The remaining correlations are only moderately large, yet they are all positively correlated. The correlations in Table 3.2 suggest that there are three distinct dimensions in this problem; spelling, reading comprehension and mathematics.

Table 3.2: Estimated correlations between scales in the 2012 Cito Eindtoets.

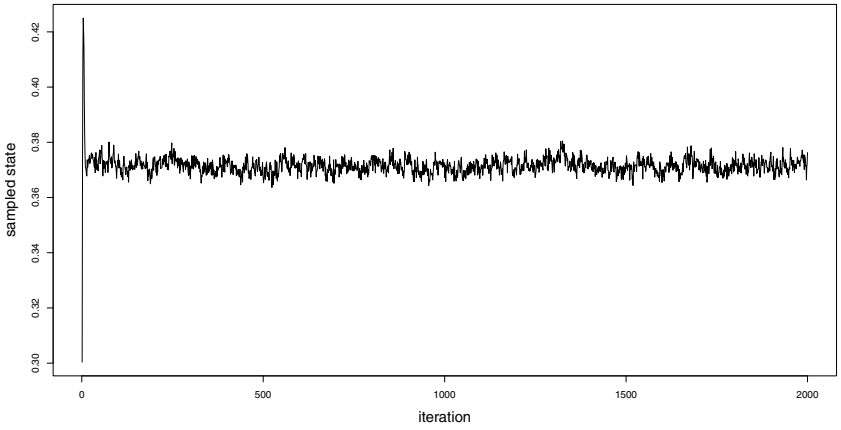
dimension	*correlations*					
non-verb spelling	1.00					
verb spelling	0.93	1.00				
reading comprehension	0.64	0.71	1.00			
basic arithmetic	0.60	0.61	0.71	1.00		
fractions	0.63	0.63	0.71	0.99	1.00	
geometry	0.61	0.62	0.69	0.97	0.98	1.00



(a) The ability of person $p = 59, 137$ in dimension one



(b) The mean of dimension six



(c) The variance of dimension three

Figure 3.11: Trace plots of an ability, a mean and a variance in the Eindtoets example.

3.5 Discussion

In this chapter, we have described two recently published algorithms that can be used to sample from conditional distributions, and have shown how the two algorithms are related. Of the two algorithms, the SVE algorithm had not been discussed in the psychometric literature, thus far. The main contribution of this chapter was to show how the efficiency of the algorithms can be improved in situations where one needs draws from not one but many conditional distributions.

The mixture of transition kernels idea from Tierney (1994) to analyse the SVE algorithm, which led us to consider manipulations of this mixture representation to improve the efficiency of the SVE algorithm. We considered two simple manipulations; manipulating the distribution of the mixture probabilities (oversampling) and manipulating the rule with which proposals are assigned to targets given a sampled data matrix (matching). The mixture representation also allows for other manipulations that we did not address in this chapter, such as manipulating the proposal distributions. This idea will be used in Chapter 4 to sample the item parameters from a multidimensional IRT model.

Finally, we note that we used GNU-R to perform the analyses, which was entirely feasible, even for the large applications. Computational time can be decreased by implementing (parts of) the code in a compiled language (e.g. Fortran, C, Delphi). Furthermore, most computer systems run on multiple cores and computational time could be decreased further by making use of the additional cores in implementations. For instance, proposals can be generated in batches, with each batch running on a single core.

Chapter 4

Bayesian Inference for Low-Rank Ising Network Models

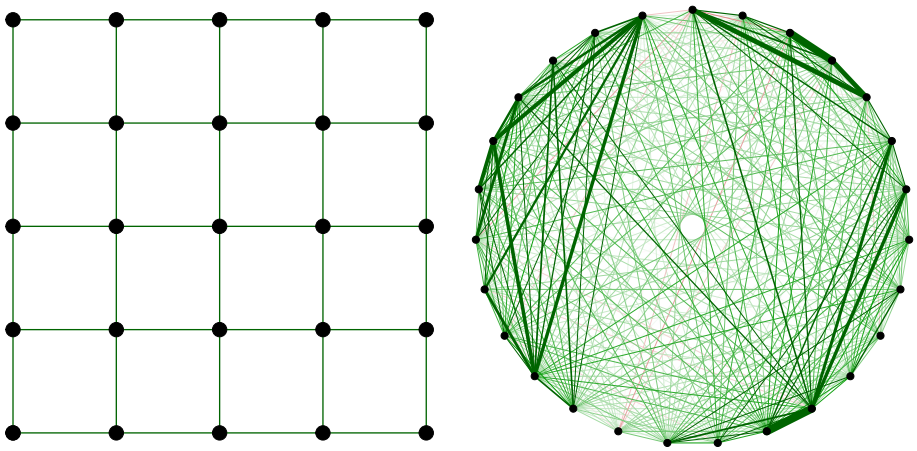
4.1 Introduction

Modelling the joint distribution of binary variables is of importance in many fields of science: Ranging from the study of phase transitions in statistical mechanics (Lee & Yang, 1952) and the study of spatial statistics in biology (Besag, 1974), to the study of comorbidity of mental disorder symptoms in psychiatry (Cramer et al., 2010). The Ising model (Ising, 1925) is an appropriate model for such distributions, as it captures all main effects and pairwise interactions (Jaynes, 1957).

Applications of the Ising model come in two distinct flavours; represented schematically in Figure 4.1. On the left hand side we have an application in which variables interact only with their nearest neighbours. This is the kind of application for which physicists originally developed the Ising model. On the right hand side we have an application where every variable is (positively) correlated with nearly all other variables. This application is typical of the social sciences (Jensen, 1998; Deary, 2000; van der Maas et al., 2006).

In physics, the structure of the Ising network is derived from first principles, hence known, and used to study macroscopic phenomena arising from the microscopic interactions in such networks. The situation in the social sciences is typically different, as the connectivity matrix encoding the network structure is usually unknown and needs to be estimated from independent realizations of the network state. Estimating the network structure is difficult, however, because the likelihood is intractable and the number of parameters is very large. The nearest neighbour network has relatively few non-zero interactions, and the use of pseudo-likelihood estimation (Besag, 1975) combined with regularization constraints on

Adapted from: Marsman, M., Maris, G., Bechger, T. & Glas, C. (2014). Bayesian Inference for Low-Rank Ising Network Models. Submitted for publication.



(a) Nearest neighbour network. (b) Dense network.

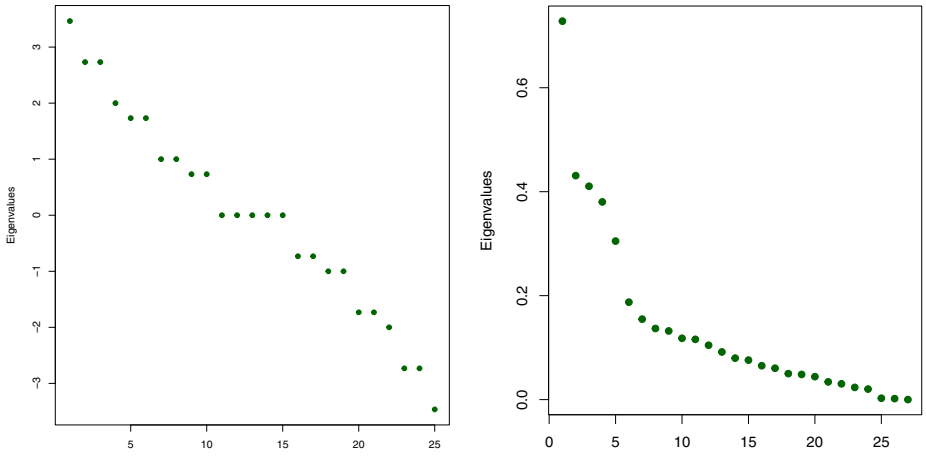
Figure 4.1: Two distinct flavours in applications of the Ising model.

the interaction effects makes approximate inference entirely feasible (van Borkulo et al., in press). In contrast, the dense network has all interactions non-zero, which makes even approximate inference difficult due to the large amount of unknown parameters.

With the current big data revolution, the number of variables on which an increasing amount of data becomes available grows fast, holding the promise of rapid progress in uncovering the relations between these variables. To date this promise is left unfulfilled since the number of realizations on the network state simply cannot keep pace with the number of variables and unknown effects, prohibiting direct estimation of the full connectivity matrix. How are we then to study the phenomena arising in such networks?

A key distinction between social science and physical science applications of the Ising model is in the rank of the connectivity matrix. In Figure 4.2 we see the eigenvalue spectrum for the two networks in Figure 4.1. Whereas a near linear spectrum is found in the nearest neighbour network, the eigenvalues in the dense network rapidly decrease in magnitude. This observation suggests that the connectivity matrix of a dense network is well approximated by a low-rank matrix.

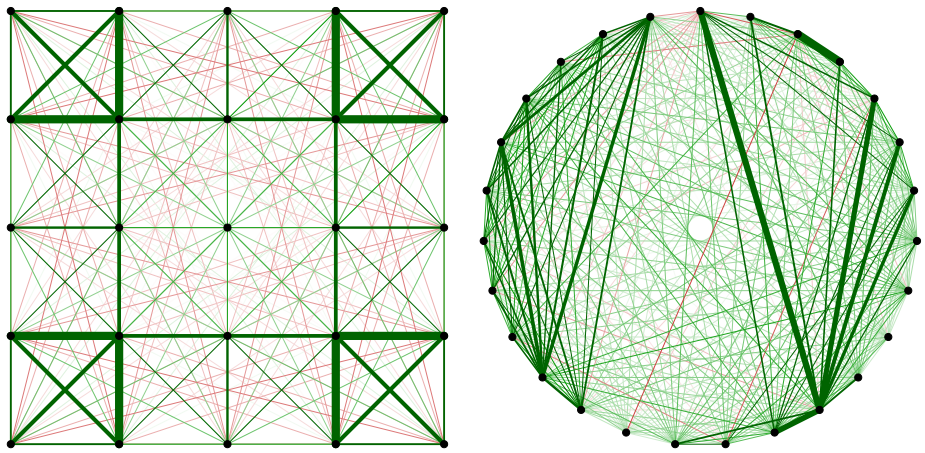
A vital property of low rank approximations follows from a theorem published by Eckart and Young (1936) in the very first volume of the journal Psychometrika. The Eckart-Young Theorem shows that in a least-squares sense, the best approximation of rank r to a matrix consists of the eigenvalue decomposition in which all but the largest r eigenvalues are equated to zero. This theorem allows us to find a low-rank approximation to the full connectivity matrix, with the key property that the first r eigenvalues and their corresponding eigenvectors can be recovered even if the remaining eigenvalues and eigenvectors are ignored. To demonstrate the power of this result, Figure 4.3 shows the rank four approximation to the two networks in Figure 4.1. As expected, the nearest neighbour network is not recovered in the rank four approximation, yet the dense network clearly is. This shows



(a) Nearest neighbour network. (b) Dense network.

Figure 4.2: Key distinction in eigenvalue spectra.

that this parsimonious low-rank approximation is useful to uncover the structure of the connectivity matrices of dense networks.



(a) Nearest neighbour network. (b) Dense network.

Figure 4.3: Rank four approximation to the networks in Figure 4.1.

An other vital result is a latent variable representation of the Ising model developed by Kac (1968), further developed by Emch and Knops (1970) and independently discovered in many places in the statistical literature (Olkin & Tate, 1961; Besag, 1974; Lauritzen & Wermuth, 1989; McCullagh, 1994; Anderson & Yu, 2007; Anderson, Li, & Vermunt, 2007). Specifically, every eigenvector for a connectivity matrix gives rise to a latent variable, such that all variables are independent conditionally on the full set of latent variables:

$$\exists \boldsymbol{\Theta} : \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\Theta}.$$

That is, there exist latent variables (Θ) that explain all the pairwise interactions in a statistical sense. The distribution of the variables conditionally on the latent variables is known as a multidimensional Item Response Theory (IRT) model in the field of psychometrics (Reckase, 2009), see Section 4.4.1. The insight of Kac, Emch and Knops is schematically represented in Figure 4.4 using four latent variables. Ignoring some of the latent variables by equating the smallest eigenvalues to zero, amounts to ignoring residual pairwise interactions, but leaves the recovered eigenvalues and corresponding eigenvectors unaffected.

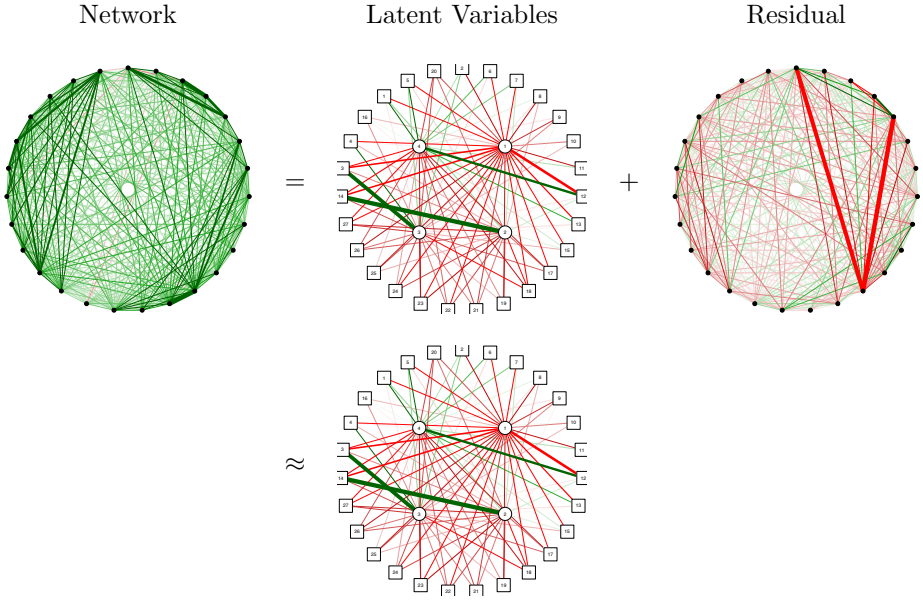


Figure 4.4: Rank four latent variable approximation to the dense network.

4.2 Results

4.2.1 Full-data-information estimation

The Ising model is mathematically elegant, yet notoriously difficult to compute. The main problem is the normalizing constant Z in equation (4.3), called the partition function, which involves a sum over all 2^n possible states of an n variable network. As the partition function depends on all the model parameters, likelihood based statistical inference is impossible, except for very small or severely constrained Ising models.

The computational problem becomes more tractable when we use the latent variable representation of the Ising model. The conditional distribution of the observed variables conditionally on the latent variables does not depend on the partition function and is available in an easily computed closed form. The partition function only figures in the distribution of the latent variables themselves. The posterior distribution of the Ising model parameters ($\mathbf{A}$) and the latent variables ($\boldsymbol{\theta}$) given the data ($\boldsymbol{\sigma}$) is the following:

$$f(\mathbf{A}, \boldsymbol{\theta}|\boldsymbol{\sigma}) \propto p(\boldsymbol{\sigma}|\boldsymbol{\theta}, \mathbf{A}) f(\boldsymbol{\theta}|\mathbf{A}) f(\mathbf{A}). \tag{4.1}$$

The whole computational complexity of this posterior distribution resides in the distribution of the latent variables, which depends on the model parameters and, in particular, on the partition function. Considering a Gibbs sampler (Geman & Geman, 1984) for simulating from this posterior distribution, we find that the full conditional distribution $f(\boldsymbol{\theta}|\mathbf{A}, \boldsymbol{\sigma})$ of the latent variables is highly tractable, and does not involve the partition function, whereas the full conditional distribution $f(\mathbf{A}|\boldsymbol{\theta}, \boldsymbol{\sigma})$ of the Ising model parameters is highly intractable because it involves the partition function.

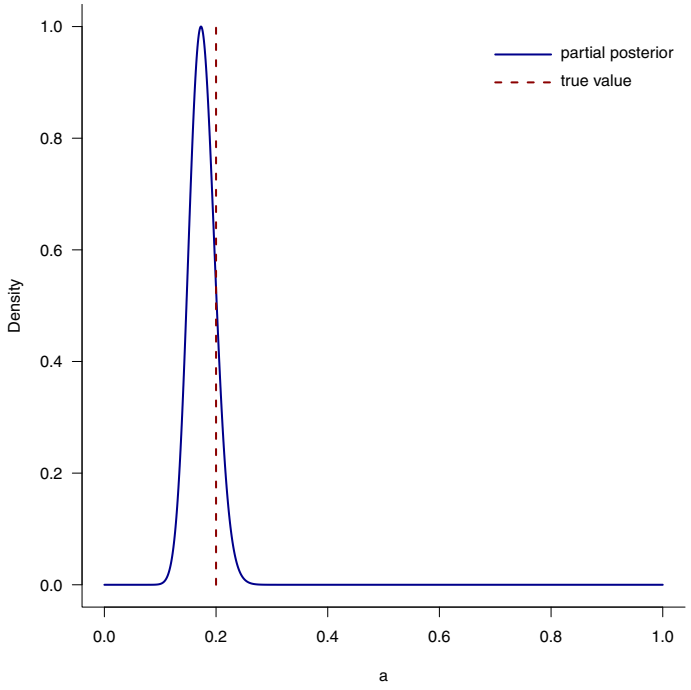
When the latent variable distribution $f(\boldsymbol{\theta}|\mathbf{A})$ in the Ising model is replaced by a prior distribution $g(\boldsymbol{\theta})$ that does not dependent on the model parameters, we have a regular multidimensional IRT problem (Reckase, 2009):

$$g(\mathbf{A}, \boldsymbol{\theta}|\boldsymbol{\sigma}) \propto p(\boldsymbol{\sigma}|\boldsymbol{\theta}, \mathbf{A}) g(\boldsymbol{\theta}) f(\mathbf{A}), \tag{4.2}$$

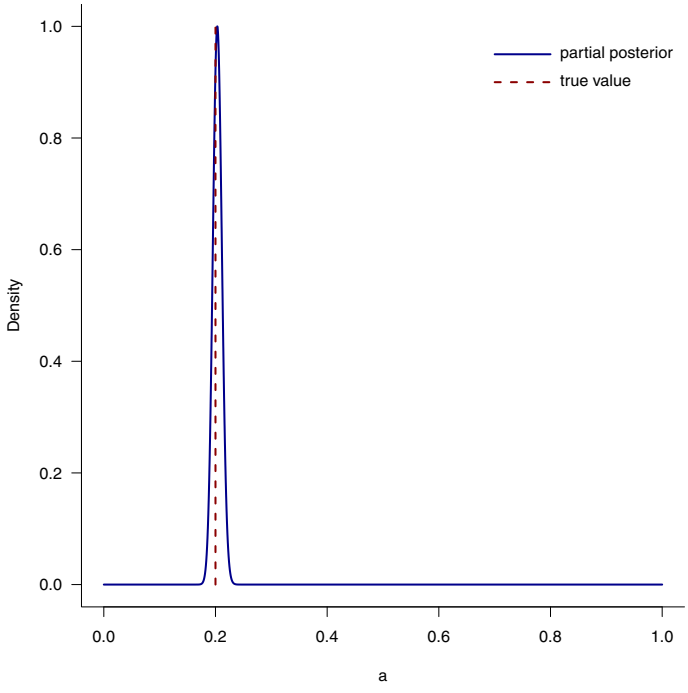
for which the full-conditionals $g(\boldsymbol{\theta}|\mathbf{A}, \boldsymbol{\sigma})$ and $g(\mathbf{A}|\boldsymbol{\theta}, \boldsymbol{\sigma})$ are easily sampled from.

As we illustrate here, and show in the Methods section, we can disregard the distribution of the latent variables (i.e. $f(\boldsymbol{\theta}|\mathbf{A}, \boldsymbol{\sigma})$), when we simulate from the conditional distributions of the Ising model's parameters. In this way all conditional distributions become tractable, and at the same time all direct information on them provided by the observed variables is retained. The only information we ignore is that which is encoded in the prior distribution for the latent variables. That is, we combine the full-conditional of the latent variables from equation (4.1) with the full-conditional of the model parameters from equation (4.1). We call this approach full-data-information estimation.

We illustrate this using the simplest non-trivial case, which is a fully connected network with all pairwise interactions equal, known as the Curie-Weiss model. This simplified model only involves one parameter, the unknown interaction strength a. In Figure 4.5 we see that the approximate posterior distribution nicely covers the true parameter value, and becomes more concentrated around this value as the sample size increases.



(a) $N = 100$ realizations.



(b) $N = 1,000$ realizations.

Figure 4.5: Partial conditional distribution of a.

4.2.2 Data example

It is the combination of the Eckart and Young theorem and the latent variable representation of Kac with full-data-information estimation that allows us to estimate low-rank Ising networks as we see them in the social sciences. To illustrate the approach we consider a large educational measurement application. The Cito Eindtoets (`www.cito.com`) is a test consisting of 200 questions, related to 12 theoretically distinct primary school subjects in arithmetic, language, and general study skills. The test is administered yearly to some 150,000 children at the end of Dutch primary education. We estimate a rank three approximation to the connectivity matrix.

Figure 4.6 displays both the rank three approximation, as the individual rank one components as a heatmap. As argued above, even though the true connectivity matrix might be of a much higher rank than three, the three estimated components correspond to the three eigenvectors of the true connectivity matrix with the highest eigenvalues. The first component corresponds to a network in which all nodes are connected to one another, and (almost) all interactions are positive. The second and third component are such that particular sets of questions get higher positive interactions amongst themselves, whereas the interactions between questions from different sets is negative. The second component is a contrast between the different language subjects writing ("W"), spelling ("S"), and reading comprehension and vocabulary ("RV") and the subjects of mathematics ("M") and study skills ("SK"), whereas the third component is a contrast between the spelling subject and the other language subjects combined with the study skills subject. Note that, the positive interactions in one component can cancel against negative interactions in another component. For instance, in the first and second component mathematics and study skills have a positive interaction, whereas in the third component they have a negative interaction.

Figure 4.6: Rank three approximation to the Cito Eindtoets data. Negative interactions are in red, positive interactions in blue, and small or absent interactions in gray.

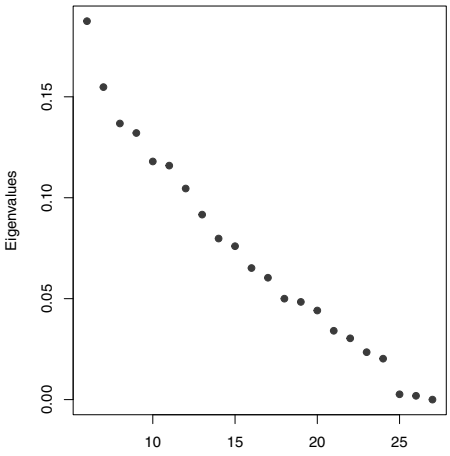


Figure 4.7: Spectrum of the last 22 eigenvalues in the dense network.

4.3 Discussion

Typical eigenvalue spectra found in the social sciences have a sharp drop in magnitude for the first few eigenvalues, after which the eigenvalues slowly decay, as depicted in Figure 4.2(b). These plots resemble a mountain cliff with broken rock fragments at the base, and are therefore called *scree plots*, where scree refers to the set of slowly decaying eigenvalues after the elbow. Scree plots are used to determine the relative importance of the eigenvalues, in which values after the elbow are often assumed ignorable, due to, for instance, sampling error. However, note that the eigenvalues after the elbow in Figure 4.2(b) have a near linear spectrum, see Figure 4.7, which resembles the spectrum found for the nearest neighbour network (see Figure 4.2(a)). This suggests that the typical social science application calls for a mix of: a) a dense network approximation, as discussed in this chapter, to use for the first few eigenvectors, and b) the nearest neighbour approximation of van Borkulo et al. (in press), to explain the residual structure.

We have shown how the Ising model could be estimated using full-data-information, in which we ignore prior structure on the parameters that resides in the latent variable model. This approximate estimation technique opens the door for estimation of other models, such as the Potts model (Potts, 1952) or a mix of models for discrete and continuous random variables.

The latent variable representation of Kac (1968) opens the way to new areas of research. For instance, the IRT model is closed under marginalization, which shows that the IRT problem in equation (4.2) can be used as a starting point to approximate incomplete networks.

4.4 Methods

4.4.1 From Ising to Item Response Theory

The Ising model is characterized by the following distribution:

$$p(\boldsymbol{\sigma}|\mathbf{A},\mathbf{b}) = \frac{1}{Z}\exp\left\{\boldsymbol{\sigma}^T\mathbf{A}\boldsymbol{\sigma} + \boldsymbol{\sigma}^T\mathbf{b}\right\}, \tag{4.3}$$

where the partition function Z serves to make the distribution sum to one and is a function of the main effects $\mathbf{b}$ and the connectivity matrix $\mathbf{A}$ containing the pairwise interactions. It is readily observed that all parameters are identifiable from the data, except for entries on the diagonal of the connectivity matrix.

Choosing diagonal values for the connectivity matrix such that all eigenvalues are non-negative, we obtain

$$\mathbf{A} + c\mathbf{I} = \mathbf{Q}(\boldsymbol{\Lambda} + c\mathbf{I})\mathbf{Q}^T = \mathbf{E}^T\mathbf{E}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of the original matrix $\mathbf{A}$. In this expression we conserve the off-diagonal entries in the connectivity matrix and at the same time ensure that the scaled connectivity matrix $\mathbf{E}^T\mathbf{E}$ is positive (semi-)definite. This allows us to use the well known Gaussian identity to represent the Ising model equivalently in the following form:

$$p(\boldsymbol{\sigma}|\mathbf{E},\mathbf{b}) = \int_{\mathbb{R}^n} \frac{1}{Z\sqrt{\pi}^n}\exp\left\{\boldsymbol{\sigma}^T\mathbf{b} + 2\boldsymbol{\sigma}^T\mathbf{E}\boldsymbol{\theta} - \boldsymbol{\theta}^T\boldsymbol{\theta}\right\} d\boldsymbol{\theta}.$$

In this expression the quadratic form is linearised, allowing for an explicit factorization:

$$p(\boldsymbol{\sigma}|\mathbf{E},\mathbf{b}) = \int_{\mathbb{R}^n} \prod_i p(\sigma_i|\boldsymbol{\theta},\mathbf{e}_i,b_i) f(\boldsymbol{\theta}|\mathbf{E},\mathbf{b}) d\boldsymbol{\theta}.$$

We can now recognize a multidimensional IRT model:

$$p(\sigma_i|\boldsymbol{\theta},\mathbf{e}_i,b_i) = \frac{\exp\{\sigma_i[b_i + 2\mathbf{e}_i\boldsymbol{\theta}]\}}{\exp\{+[b_i + 2\mathbf{e}_i\boldsymbol{\theta}]\} + \exp\{-[b_i + 2\mathbf{e}_i\boldsymbol{\theta}]\}}$$

with a particular distribution for the latent variables

$$f(\boldsymbol{\theta}|\mathbf{E},\mathbf{b}) = \frac{\prod_i \left(\exp\{+[b_i + 2\mathbf{e}_i\boldsymbol{\theta}]\} + \exp\{-[b_i + 2\mathbf{e}_i\boldsymbol{\theta}]\}\right)}{Z\sqrt{\pi^n}} \exp\left\{-\boldsymbol{\theta}^T\boldsymbol{\theta}\right\},$$

where $\mathbf{e}_i$ is the i-th row-vector of $\mathbf{E}$. Note that, in this representation, the partition function only figures as a normalizing constant of the latent variable distribution.

4.4.2 Full-data-information estimation

Upon choosing a proper prior distribution for the model parameters $\mathbf{E}$ and $\mathbf{b}$, we obtain a posterior distribution for both the model parameters and the latent variables. It is not simple to simulate directly from this joint posterior distribution, and we use the Gibbs sampler (Geman & Geman, 1984). The full-conditional

distribution of the latent variables corresponding to observation p, $p = 1, \ldots, N$, is a simple multivariate normal distribution:

$$f(\boldsymbol{\theta}_p|\mathbf{E}, \mathbf{b}, \boldsymbol{\sigma}_p) \propto \exp\left\{2\boldsymbol{\sigma}_p^T \mathbf{E}\boldsymbol{\theta}_p - \boldsymbol{\theta}_p^T \boldsymbol{\theta}_p\right\}.$$

For the model parameters we find complicated full-conditional distributions, as these explicitly depend on the partition function. Since the partition function only figures as the normalizing constant of the latent variable distribution, our proposal is to simulate the model parameters from the *partial* full-conditional distribution:

$$f_{\text{partial}}(\mathbf{E}, \mathbf{b}|\boldsymbol{\theta}, \boldsymbol{\sigma}) \propto \prod_p \prod_i p(\sigma_{pi}|\boldsymbol{\theta}_p, \mathbf{e}_i, b_i) f(\mathbf{E}, \mathbf{b}).$$

This partial conditional distribution ignores the information encoded in the marginal distribution of the latent variables but retains all the information about the model parameters that is contained in the data. This is why we refer to our approach as full-data-information estimation.

We alternately sample from the full-conditional distribution of the latent variables and the partial conditionals of the Ising model parameters. This does not amount to a proper Gibbs sampler in the sense that, after discarding the latent variables, the posterior distribution of the model parameters is not the invariant distribution. One way of looking at this scheme is that a second posterior distribution is set up for the latent variables and model parameters, one in which we have a proper prior distribution for the latent variables that does not depend on the model parameters. In this second posterior distribution, the partial full conditional distributions of the model parameters would be the correct full conditional distributions in a Gibbs sampler, and it would be the multivariate normal full conditional distribution for the latent variables that would no longer be correct. When the Bayesian Central Limit Theorem would be the same for both posteriors, our scheme that mixes the full-conditional for the one posterior with another from the other posterior would still converge to the same Central Limit Theorem. In that sense our full-data-information estimation admits of the same large sample properties as does full Bayesian estimation, but without ever having to deal with computing the partition function.

4.4.3 Simulating from the partial conditionals

Simulating from the partial-conditional distribution of the parameters may seem to be a difficult problem in its own right, but this problem has been resolved in many different places (Patz & Junker, 1999b, 1999a; Albert, 1992; Béguin & Glas, 2001; Maris & Maris, 2002). We consider here a new method that is both computationally simple and highly efficient.

Our proposal is to use a Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) for simulating from the partial-conditional distribution of the model parameters, which differs from earlier such approaches (Patz & Junker, 1999b, 1999a) in the particular choice of the proposal distribution. We consider an independence Metropolis-Hastings algorithm (Tierney, 1994) in which the proposal

distribution is easy to simulate from, and approximates its target ever better the more data we have. This combination makes the algorithm ideally suited for large data sets.

Consider a set of random variables $\mathbf{Z}$, such that $\perp\!\!\!\perp_p Z_p$ and $Z_p \sim F_p$. Define a matrix of binary indicator variables (coded as zero/one) with entries: $x_{pj} = (z_p < z_j)$, such that for column j of the matrix we obtain:

$$f_j(z_j|\mathbf{x}_j) \propto \prod_{p \neq j} F_p(z_j)^{x_{pj}} (1 - F_p(z_j))^{1-x_{pj}} f_j(z_j) = f(\mathbf{x}_j, z_j). \tag{4.4}$$

This distribution closely resembles the partial-conditionals. For any parameter W, the partial conditional is of the form:

$$f_{\text{partial}}(w|\mathbf{y}) \propto \prod_{p=1}^{N} F_p(w)^{y_p} (1 - F_p(w))^{1-y_p} f(w),$$

where $y_p = (\sigma_p + 1)/2$, F_p is a (logistic) distribution function and $f(w)$ the prior density of W. Thus, $f_j(z_j|\mathbf{x}_j)$ will be used as a proposal density.

To illustrate how the algorithm works, consider a simple case with $N = 2$ and a (target) partial conditional:

$$f_{\text{partial}}(w|\mathbf{y}) \propto F_1(w)(1 - F_2(w))f(w),$$

where $y_+ = \sum_p y_p = 1$. We now generate the vector $\mathbf{z}$ and choose j (in equation (4.4)) such that $x_+ = \sum_{p \neq j} x_p = y_+$ and find (for instance):

$$f(z_j|\mathbf{x}_j) \propto F_2(z_j)(1 - F(z_j))f_1(z_j).$$

That is, z_j is a draw from a posterior based on N observations and a total score x_+, which differs from the target distribution w.r.t. the distribution of one of the observations and the prior density. In fact, the prior density and the distribution of the first observation have switched places.

The F_p are logistic distributions with a mean μ_p and scale s_p. If we choose a standard logistic prior density, the probability to move from the current state w' to a state $w^* = z_j$ is equal to $\min(1, \alpha)$, with

$$\begin{aligned}\alpha &= \frac{f_{\text{partial}}(w^*|\mathbf{y}) f(w'|\mathbf{x}_j)}{f_{\text{partial}}(w'|\mathbf{y}) f(w^*|\mathbf{x}_j)} \\ &= \exp\{(w^* - w')(1/s_2 - 1/s_1)\} \frac{(1 + \exp\left\{\frac{w^* - \mu_1}{s_1}\right\})(1 + \exp\{w'\})}{(1 + \exp\left\{\frac{w' - \mu_1}{s_1}\right\})(1 + \exp\{w^*\})},\end{aligned}$$

a relatively simple expression in which many things cancel. Even for large N this expression remains simple, regardless of the choice of prior density.

Chapter 5

A Cautionary Note About Data Augmentation Algorithms

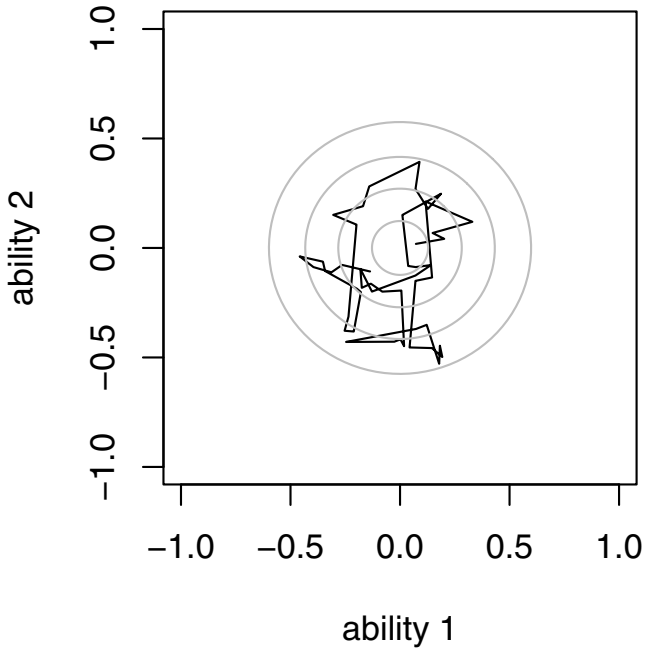
5.1 Introduction

This chapter is about the asymptotic behaviour of Markov Chain Monte Carlo (MCMC) sampling methods using Data-Augmentation (DA); that is, methods where unobserved or latent data are introduced to make sampling easier. More specifically, its purpose is to demonstrate, by means of an example, that the dependence between successive samples (i.e., the autocorrelation) may depend on sample size.

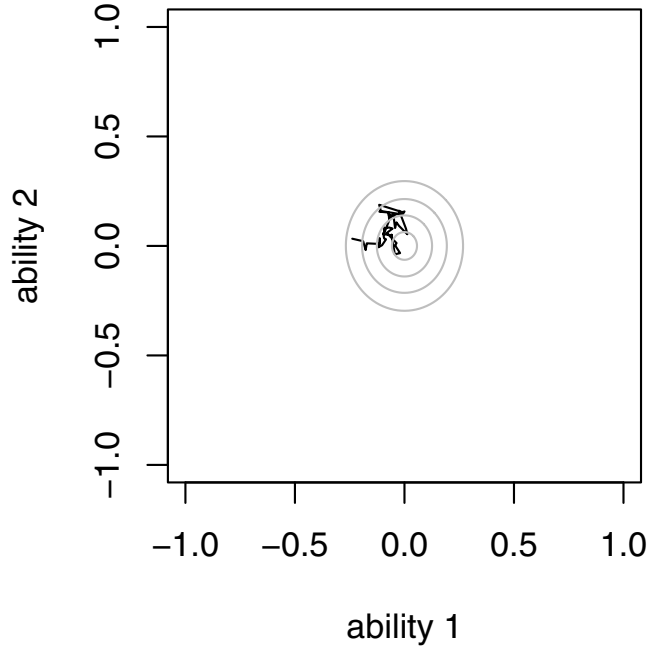
To illustrate the problem consider a commonplace situation where two students independently took a test consisting of several "items" (i.e, questions, tasks, etc.). We employ a model where each student is characterized by an ability and a Gibbs sampler (Geman & Geman, 1984) is used to compute the joint posterior density of the abilities. The result may look like Figure 5.1 where the top panel shows the posterior for a test with 20 items, and the bottom panel for a test that is five times longer as witnessed by a higher concentration around the true ability. Also shown are 50 draws produced by the Gibbs sampler with successive draws from each full conditional distribution connected by line segments to show how the sampler moves through the support of the posterior. A well-behaved Gibbs sampler (e.g., Tierney, 1994) is expected to "walk around" freely producing a (dependent) sample from the posterior.

What inspired this note is the observation that the sampler takes smaller steps in the bottom panel. This means, first of all, that we need more iterations to get to the area where the posterior probability mass is concentrated but this can be remedied with better starting values. It becomes more serious if the steps get smaller *relative to the posterior standard deviation* because this means that it

Adapted from research that was done in collaboration with Gunter Maris and Timo Bechger and Katerina Papadimitropoulou, who was a visiting student at Cito.

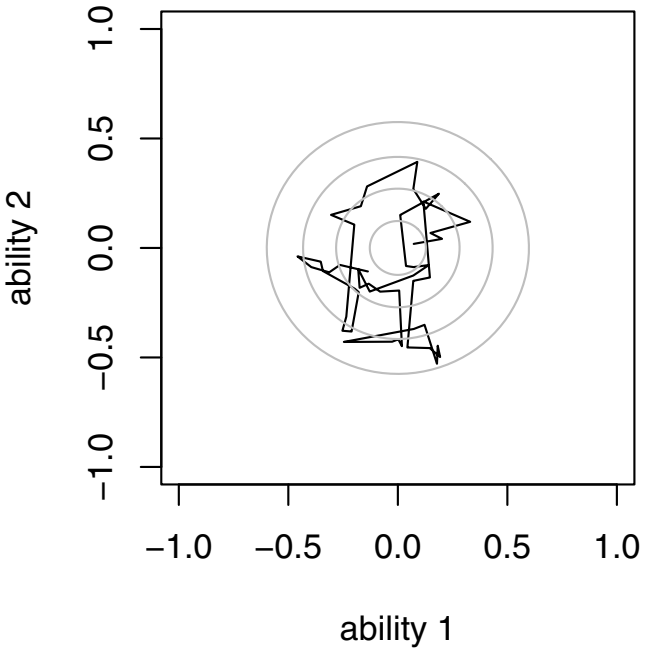


(a) 20 items

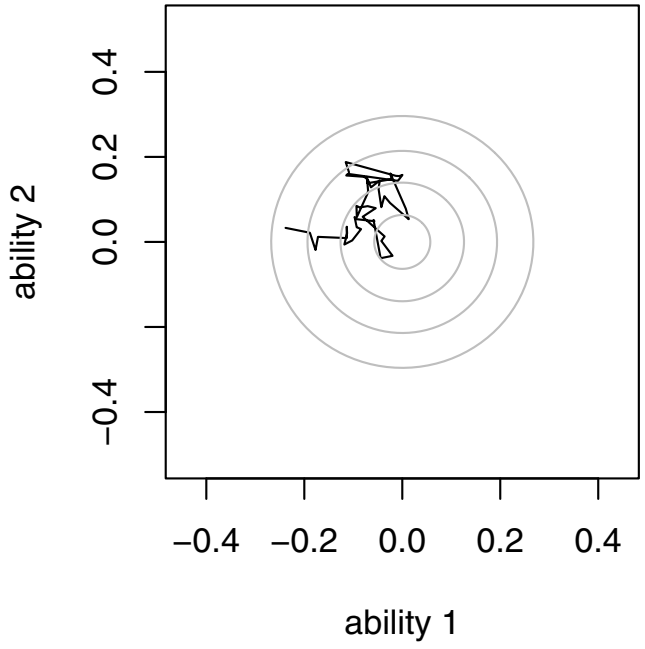


(b) 100 items

Figure 5.1: Contour plots of the joint posterior densities of two abilities and 50 samples produced by the Gibbs sampler connected by line segments.



(a) 20 Items



(b) 100 items

Figure 5.2: Similar to Figure 5.1 but with the posterior standard deviations made equal.

would take longer to explore the posterior support. That this might be the case here is suggested by Figure 5.2 where the abilities in the right panel are re-scaled such that the posterior standard deviation is about the same in both panels.

The example will be presented in more detail in the next section where we give a formal demonstration that the average step size does indeed diminish with the number of observations (i.e., item responses) at a faster rate than the posterior standard deviation and the sampler will eventually stop moving. We will then discuss the relation between step-size and autocorrelation. Our main finding is that the autocorrelation will increase with sample size if the variance of the augmented data posterior goes to zero at a faster rate then the posterior variance; i.e., faster that n^{-1}.

5.2 Example

5.2.1 The Model

Assume that each of n persons answers each of k items. The observations are the item responses $\mathbf{x}_p = (x_{p1}, x_{p2}, \ldots, x_{pk})$, where $x_{pi} = 1$ if the answer of student p on item i was correct and $x_{pi} = 0$ when the answer was incorrect. The responses are assumed to be independent with

$$P(X_{pi} = 1|\theta_p, \delta_i) = \Phi\left(\theta_p - \delta_i\right), \tag{5.1}$$

where Φ is the distribution function of the standard normal distribution, θ_p denotes the ability of person p, and δ_i the difficulty of item i. Thus, the probability of a correct response increases when a person becomes more able relatively to the difficulty of the item. This model is a mixed probit regression model, known as the *normal-ogive model* in educational measurement (Richardson, 1936). To keep the example as simple as possible we assume that the items are equally difficult with $\delta_i = 0$ for all i. We use the standard normal as a prior distribution for ability. Note that we use probit regression for convenience; virtually everything is the same when we consider logistic regression.

5.2.2 A Data-Augmentation Sampler

It is difficult to compute or sample from the posterior directly and Maris and Maris (2002) suggest a DA Gibbs sampler based on the fact that the model can be re-written as:

$$P(X_{pi} = 1|\theta) = P(Z_{ip} \leq \theta_p), \tag{5.2}$$

where the latent response variables Z_{pi} are independent and standard normally distributed. Generating the latent responses is quite simple as the density of the observed responses given the latent responses equals one if the observed response matches the latent response and zero otherwise. Thus, we find that

$$\begin{aligned} f(z_{pi}|x_{pi}, \theta_p) &\propto P(x_{pi}|z_{pi}, \theta_p) f(z_{pi}|\theta_p) \\ &= (z_{pi} \leq \theta_p)^{x_{pi}} (z_{pi} > \theta_p)^{1-x_{pi}} \phi(z_{pi}), \end{aligned}$$

where () denotes the indicator function. This is a truncated normal distribution from which it is easy to sample. Sampling from the *augmented data posterior* of ability turns out to be equally easy as it is readily found that

$$\begin{aligned} f(\theta_p|\mathbf{x}_p, \mathbf{z}_p) &\propto f(\mathbf{x}_p|\mathbf{z}_p, \theta_p) f(\theta_p) \\ &= \left[\prod_i (\theta_p \geq z_{pi})^{x_{pi}} (\theta_p < z_{pi})^{1-x_{pi}} \right] \phi(\theta_p) \qquad (5.3) \\ &= \left(\max_{i:x_{pi}=1} (z_{pi}) \leq \theta_p < \min_{i:x_{pi}=0} (z_{pi}) \right) \phi(\theta_p), \qquad (5.4) \end{aligned}$$

where ϕ denotes the standard normal probability density function. This is a doubly truncated distribution. Note that the truncation interval in (5.4) is the intersection of the half-open intervals in (5.3). Maris and Maris (2002) prove that this intersection is never empty.

5.2.3 Step sizes

The interval in (5.4) literally determines the step sizes in Figures 5.1 and 5.2 and we readily see that it tends to become smaller with sample size. To wit, the lower end-point is a maximum which will tend to increase when taken over a larger set and the higher end-point is a minimum and will tend to decrease. The question is: How fast? To determine the rate by which the step size diminishes we determine the rate by which the interval diminishes.

To this aim, we first transform the latent data so that they are uniformly distributed. Specifically, let $U_{pi} = \Phi(Z_{pi})$ and $\Pi_p = \Phi(\Theta_p)$ so that

$$f(\pi_p|\mathbf{x}_p, \mathbf{u}_p) \propto \left(\max_{i:x_{pi}=1} (u_{pi}) \leq \pi_p < \min_{i:x_{pi}=0} (u_{pi}) \right). \qquad (5.5)$$

The U_{pi} are independent and uniformly distributed and it is a standard result that:

$$\begin{aligned} \frac{\max_{i:x_{pi}=1}(U_{pi})}{\pi_p} &\sim Beta(x_{p+}, 1), \quad \text{and} \\ \frac{\min_{i:x_{pi}=0}(U_{pi})}{1-\pi_p} &\sim Beta(1, k - x_{p+}) \end{aligned}$$

where $x_{p+} = \sum_i x_{pi}$. It follows that:

$$\mathcal{E}\left[\min_{i:x_{pi}=0} (U_{pi}) - \max_{i:x_{pi}=1} (U_{pi}) \right] = \pi_p + (1-\pi_p) \frac{1}{k - x_{p+} + 1} - \pi_p \frac{x_{p+}}{x_{p+} + 1}$$

Thus, we find that the maximum expected step size in an iteration of the sampler is of order k^{-1} whereas, asymptotically, the posterior standard deviation is of order $(\sqrt{k})^{-1}$ (e.g., Chang & Stout, 1993).[2] This confirms that the Markov chain gets "stickier" as the sample size increases, until it ultimately grinds to a halt.

[2] Using basic results from extreme-value theory (e.g., Arnold, Balakrishnan, & Nagaraja, 2008, Chpt. 8), it can be shown that $\min_{i:x_{pi}=0}(U_{pi}) - \max_{i:x_{pi}=1}(U_{pi})$ converges to a Gamma(2,1) distribution.

5.3 Autocorrelation

Looking at step-sizes is an intuitive but indirect way to study autocorrelation. Assuming that the chain has converged, it follows from the *covariance decomposition formula* that the (lag-1) autocorrelation equals:

$$\rho(\theta^{(t)}, \theta^{(t+1)}|x) = \frac{Var(E\{\theta|Z,x\})}{Var(\theta|x)} = 1 - \frac{E\{Var(\theta|Z,x)\}}{Var(\theta|x)}. \tag{5.6}$$

This insightful expression, known as the (Bayesian) fraction of missing information, was derived by Liu (1994) and shows that *the autocorrelation depends on the ratio of the augmented data posterior variance and the posterior variance.* It shows that the autocorrelation will go up when more *layers of latent data* are added; i.e., when we condition on more and the variance of the data-augmented posterior goes down. More interesting for our present purpose is that it also shows that autocorrelation may depend on sample size. For most samplers that we know, both posteriors converge at the same rate and the autocorrelation is constant with respect to sample size. Our example shows that this need not be the case. Although we did not prove this directly, it follows from Equation 5.6 that, in the example, the variance of the augmented data posterior goes to zero at a faster rate than the posterior distribution.

As an aside, we note that the asymptotic behaviour of the autocorrelation may depend on parametrization. For instance, with $Y_{pi} = \theta_p - Z_{pi}$, the normal-ogive model can be written as:

$$P(X_{pi} = 1|\theta_p) = P(Y_{pi} \geq 0). \tag{5.7}$$

With a standard normal prior distribution for ability[3], the augmented data posterior is now a normal distribution with variance $(k+1)^{-1}$ instead of a truncated normal distribution (Albert, 1992; Roy & Hobert, 2007). Thus, we can change the parametrization of the model, and obtain a "healthy" Gibbs sampler where the autocorrelation is constant and doesn't increase with sample size. It is, of course, well-known that parametrization matters. The example illustrates nicely the extent to which this is true.

[3]Note that this method works when the latent responses are normally distributed but not when they follow a logistic distribution as in the Rasch model (Rasch, 1960).

5.4 Conclusion

It has been demonstrated by means of an example that the autocorrelation in DA samplers need not be constant with respect to the number of observations. Hence, although data augmentation is a powerful tool, it is not guaranteed that it is efficient for large data sets. It is possible that the constructed Markov chain may grind to a halt as the sample size increases. Note that the sampler used in the example is a type of *slice sampler* (Damien, Wakefield, & Walker, 1999; Neal, 2003), as shown in Appendix H. Neal (2003) briefly mentions the pathological asymptotic behaviour of the slice sampler and this chapter can be read as an illustration of his remarks. Note further that it is easy to show that the slice sampler would show the same pathological behaviour for any model of the form $P(X_{pi} = 1|x) = F(x)$, where F is the distribution function of the latent responses; logistic, normal or otherwise.

The asymptotic behaviour of the autocorrelation is determined by the asymptotic behaviour of the augmented data posterior. While the posterior converges to a normal distribution according to the usual central limit law, the asymptotic behaviour of the augmented data posterior may be that of extreme values. This has led to disaster in our example, but there is also good news. It implies that it is possible, in principle, to construct samplers that become more efficient when the number of observations increases. Actually constructing such samplers requires creativity and luck: Data augmentation continuous to be an art (van Dyk & Meng, 2001).

Appendices

Appendix A

Chapter 2: PISA analysis details

We used data from the 2006 PISA cycle, specifically data from $n = 28$ items intended to assess reading ability in booklet 6 made by $N = 1,768$ Canadian students. The data of 30 students were omitted due to missing responses, and we fitted a *One Parameter Logistic Model* (OPLM; Verhelst & Glas, 1995) on data from the remaining $N = 1,738$ students.

The item difficulties were estimated using conditional ML (CML) and the item discriminations were estimated using marginal ML (MML) using the OPLM program (Verhelst, Glas, & Verstralen, 1995). We used cross-validation for estimation of the (discrete) item discriminations; First, the discriminations were estimated based on data from a random selection of $1,200$ students. At this stage, we deleted two items that did not fit the scale (item 6 and item 8). The remaining $n = 26$ items scaled reasonably well in this sample, $R_{1C} = 133.067$, $df = 90$, $p = 0.0022$ (for a description of the R_{1C} statistic see Verhelst et al., 1995). Second, the parameters were validated on data from the remaining 538 students, and scaled well, $R_{1C} = 118.686$, $df = 90$, $p = 0.0231$.

The estimated item parameters are shown in Table A.1. When an item has more than two categories, there are more than two item parameters per item; each additional category receives its own "difficulty" parameter. We indicate the categories for these items within brackets after the item number.

Table A.1: Parameters of the estimated IRT model for the PISA example.

item number (category)	discrimination	difficulty
1	3	-0.603
2	3	-0.035
3	5	-0.089
4	5	-0.215
5	4	-0.428
6	-	-
7(1)	3	0.252
7(2)	3	-0.466
8	-	-
9	4	0.173
10	5	-0.390
11	4	-0.334
12	3	0.481
13(1)	4	0.316
13(2)	4	0.871
14	5	-0.041
15(1)	3	-0.092
15(2)	3	0.442
16(1)	3	0.299
16(2)	3	-0.093
17	4	-0.061
18	5	-0.163
19	5	-0.442
20	5	0.269
21	5	0.047
22	5	0.135
23	5	-0.149
24	3	0.065
25	2	0.084
26(1)	3	-0.140
26(1)	3	0.400
27	6	0.072
28	6	-0.167

Appendix B

Chapter 3: Oversampling in the Gamma example

The Gnu-R (R Core Team, 2014) code that was used in the Gamma example to sample from the full-conditional distribution of ν is given below.

```
#Compute t(x):
    tx = rep(0,n)
    for(i in 1:k) tx = tx - X[,i]/(theta*delta[i]) + log(X[,i])
    tx = sum(tx)
#Generate j = 100 proposals:
    anu = rgamma(j,shape = shape.nu, rate = rate.nu)
#Generate statistics t(x*):
    atx = rep(0,j)
    for(i in 1:k)
    {
        for(m in 1:j)
        {
            tmp = rgamma(n,shape = anu[m],
              rate = anu[m]/(theta*delta[i]))
            atx[m] = atx[m] + sum(log(tmp) - tmp/(theta*delta[i]))
        }
    }
#Select proposal:
    m = which(abs(tx-atx) == min(abs(tx-atx)))[1]
    anu = anu[m]
    atx = atx[m]
#Metropolis-Hastings step:
    ln.alpha = (anu-nu)*( tx - atx )
    if(log(runif(1)) < ln.alpha) nu = anu
```

Appendix C

Chapter 3: Matching in the Gamma example

The Gnu-R (R Core Team, 2014) code that was used in the Gamma example to sample from the full-conditional distribution of the person parameters is given below.

```
#Generate proposals:
    atheta = rlnorm(n,theta.mu,theta.sd) #proposals from prior
#Compute statistics:
    tx = atx = rep(0,n)
    for(i in 1:k)
    {
        #Compute t(x*):
        atx = atx + rgamma(n,shape = nu,
          rate = nu / (atheta*delta[i]))/delta[i]
        #Compute t(x):
        tx = tx + (X[,i]/delta[i])
    }
#Permute proposals:
    O = order(order(tx))
    o = order(atx)
    atheta = atheta[o[O]]
    atx = atx[o[O]]
#Metropolis-Hastings step:
    ln.alpha = nu*(atx-tx)*(1/atheta - 1/theta)
    u = log(runif(n))
    theta[u<ln.alpha] = atheta[u<ln.alpha]
```

Appendix D

Chapter 3: Sampling data from the SRT model

In order to apply the SVE algorithm to sample from the full-conditionals of the person and item parameters, we need to be able to generate data from the model. Since we apply the same procedure for the person as for the item parameters, we only describe the strategy for the person parameters here. We use the factorization $f(\mathbf{X}, \mathbf{S}|\theta, \boldsymbol{\delta}, d) = P(\mathbf{X}|\theta, \boldsymbol{\delta}, d)\ f(\mathbf{S}|\mathbf{X}, \boldsymbol{\delta}, \theta, d)$, and use composition. Maris and van der Maas (2012) showed that $P(X = x|\theta, \delta, d)$ derived from the SRT model is a Rasch model with slope equal to the time limit d, and $f(S_{pi} = s_{pi}|X_{pi} = x_{pi}, \theta_p, \delta_i, d)$ is:

$$f(S_{pi} = s_{pi}|X_{pi} = x_{pi}, \delta_i, \theta_p, d) = \frac{(\theta_p - \delta_i)\exp\left((2x_{pi} - 1)(d - s_{pi})(\theta_p - \delta_i)\right)}{(2x_{pi} - 1)\left[\exp\left((2x_{pi} - 1)d(\theta_p - \delta_i)\right) - 1\right]}.$$

An interesting feature of this distribution is that the following set of equalities holds (let ϕ denote $\theta - \delta$ in the equalities):

$$(S|X = 1, \phi) \underset{st}{=} (d - S|X = 0, \phi) \underset{st}{=} (S|X = 0, -\phi) \underset{st}{=} (d - S|X = 1, -\phi).$$

This indicates that we can introduce a new variable $\hat{S}$:

$$\hat{S} = \begin{cases} S & \text{if } X = 1 \\ d - S & \text{if } X = 0 \end{cases} \sim (S|X = 1, \theta, \delta, d),$$

which Maris and van der Maas (2012) call *pseudo time* and is independent of accuracy ($X \perp\!\!\!\perp \hat{S}|\Theta$). Thus, to generate data from the SRT model, we generate X from a Rasch model with slope d, which is a trivial exercise, and to generate S we generate $\hat{S}$ via inversion and solve for S using

$$S = \begin{cases} \hat{S} & \text{if } X = 1 \\ d - \hat{S} & \text{if } X = 0 \end{cases}.$$

That is, draw $u \sim \mathcal{U}(0, 1)$, and set $\hat{S}_{pi}$ equal to

$$\frac{1}{\delta_i - \theta_p} \ln \left[1 - u(1 - \exp(d(\delta_i - \theta_p)))\right].$$

Appendix E

Chapter 3: Generating plausible values with matching

It is assumed that each pupil is characterized by a unidimensional ability, θ_p, sampled independently from a prior $f(\theta|\mathbf{y}_p, \boldsymbol{\lambda})$ as a function of observed characteristics $\mathbf{y}_p$ and parameters $\boldsymbol{\lambda}$. An IRT model is used for the conditional distribution $P(\mathbf{X}_p = \mathbf{x}_p|\theta_p, \boldsymbol{\delta})$ of a vector of item responses $\mathbf{X}_p = \{X_{p1}, X_{p2}, \ldots, X_{pk}\}$ as a function of ability and item parameters $\boldsymbol{\delta}$. Conditional on ability, the responses are assumed to be independent of $\mathbf{y}_p$ and $\boldsymbol{\lambda}$. In this example, the Rasch (1960) model is used in combination with the prior distribution:

$$f(\theta_p|\mathbf{y}_p, \boldsymbol{\lambda} = \{\boldsymbol{\beta}, \sigma\}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\theta_p - \mathbf{y}_p^T\boldsymbol{\beta})^2}{2\sigma^2}\right\}.$$

To sample from the posterior distributions of the pupil parameters (i.e. plausible values) we use the matching procedure. We generate θ_q^*, $q = 1, \ldots, n$, from $f(\theta|\mathbf{y}_q, \boldsymbol{\beta}, \sigma)$ and use it to generate a vector of response times $\mathbf{x}_q^*$ from $P(\mathbf{x}|\theta_q^*, \boldsymbol{\delta})$. Say that we use $f(\theta|\mathbf{x}_q^*, \boldsymbol{\delta}, \mathbf{y}_q, \boldsymbol{\beta}, \sigma)$ as proposal for a target p (p need not equal q), the probability $\pi(\theta_p' \rightarrow \theta_q^*)$ to make a transition from θ_p' to θ_q^* is then equal to $\min\left\{1, \alpha(\theta_p' \rightarrow \theta_q^*)\right\}$, with

$$\ln \alpha(\theta_p' \rightarrow \theta_q^*) = (\theta_q^* - \theta_p')(t(\mathbf{x}_p, \mathbf{y}_p, \boldsymbol{\beta}, \sigma) - t(\mathbf{x}_q^*, \mathbf{y}_q, \boldsymbol{\beta}, \sigma)),$$

where

$$t(\mathbf{x}_p, \mathbf{y}_p, \boldsymbol{\beta}, \sigma) = \sum_{i=1}^{k} x_{pi} + \frac{\mathbf{y}_p^T\boldsymbol{\beta}}{\sigma^2}.$$

We see that it is opportune to use $t(\mathbf{x}_p, \mathbf{y}_p, \boldsymbol{\beta}, \sigma)$ to permute proposals and targets. To this aim, we compute $t(\mathbf{x}_p, \mathbf{y}_p, \boldsymbol{\beta}, \sigma)$ for each person in the sample, and for each proposal. Then, we order the targets using the $t(\mathbf{x}_p, \mathbf{y}_p, \boldsymbol{\beta}, \sigma)$, such that the corresponding statistics are ordered from small to large, and do the same for the

proposals using the $t(\mathbf{x}_q^*, \mathbf{y}_q, \boldsymbol{\beta}, \sigma)$. This simple permutation strategy ensures that if the Markov chain is stationary, the first proposal is likely to be a good proposal for the first target (since the difference between $t(\mathbf{x}_p, \mathbf{y}_p, \boldsymbol{\beta}, \sigma)$ and $t(\mathbf{x}_q^*, \mathbf{y}_q, \boldsymbol{\beta}, \sigma)$ is likely to be small), and the same holds for the second, the third, and so on.

The aforementioned can be summarized in the following GNU-R (R Core Team, 2014) code, where `y` denotes the matrix of pupil covariates:

```
#Pre-compute vector of means
  mu = y%*%beta
#Generate proposals
  atheta = rnorm(n,mu,sigma)
#Compute statistics:
  ascore = rep(0,n)
  for(i in 1:k)
  {
    ascore = ascore + 1*(rlogis(n) <= (atheta - delta[i]))
  }
  tx = score + mu/(sigma^2)
  atx = ascore + mu/(sigma^2)
#Permute proposals
  O = order(order(tx))
  o = order(atx)
  atheta = atheta[o[O]]
  atx = atx[o[O]]
#Metropolis-Hastings step:
  alpha = (atheta - theta)*(tx - atx)
  u = log(runif(n))
  theta[u<alpha] = atheta[u<alpha]
```

Appendix F

Chapter 4: Nearest neighbour / dense network model from Kac

As described in Chapter 4, applications of the Ising network model come in two distinct flavours; nearest neighbour type networks and dense networks. It was argued that there might be residual structure in the connectivity matrix of dense networks that is not accounted for in the low-rank approximation, and that this residual structure is alike the structure found in a nearest neighbour type network. This suggests that a mix of a penalty based approach and a low-rank based approach, might be useful to uncover the structure underlying such networks. To this aim, a model described by Kac (1968) is of interest, since it bridges the gap between the nearest neighbour and dense networks.

In the model of Kac (1968) it is assumed that variables are represented as nodes that are placed equidistant on a one-dimensional lattice (a line), where the strength of interaction between a variable i and j placed on this lattice equals $\gamma \exp\{\gamma|i-j|\}$. That is, there is a long range interaction, which is exponentially decreasing as the variables become more distant from each other. Note that if γ increases, the relative strength between adjacent nodes becomes larger in comparison to more distant nodes, and the model becomes more and more alike a nearest neighbour type model on the lattice. If $\gamma \to 0$, the relative strength between all nodes on the lattice tend to become equally large, and we have an increasingly more dense network on the lattice. It is in this sense that the model bridges the gap between the two types of networks discussed in Chapter 4.

The Kac (1968) model can also be used to illustrate that in a rank M approximation of a high-dimensional model the first M eigenvectors can be recovered as long as the first M eigenvalues can be clearly distinguished in the spectrum of the connectivity matrix. To this aim, we consider a network with $\gamma = 0.1$ and $\gamma = 1.0$, of which the spectra are shown in Figure F.1. For $\gamma = 0.1$ the first few eigenvalues can be clearly distinguished, whereas for $\gamma = 1.0$ they cannot.

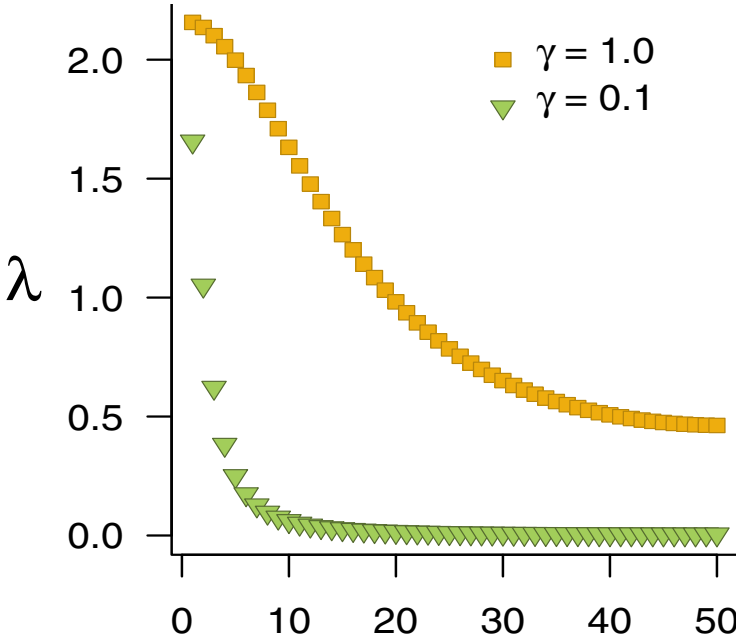


Figure F.1: Spectrum of the connectivity matrix for the $\gamma = 1.0$ and $\gamma = 0.1$ networks.

Data were simulated for $N = 10,000$ observations on a network with $n = 50$ nodes using the Gibbs sampler for $\gamma \in \{0.1, 1.0\}$. A rank three network approximation was estimated on the simulated data, in which the model does not make use of the structure in the true connectivity matrix.

An interesting feature is that the structure in the connectivity matrix is such that it has particular patterns in its eigenvectors; the eigenvectors show clear oscillating patterns with increasing intensity moving up the eigenvalue spectrum. The oscillating pattern can be seen in Figure F.2, where we plotted the true second eigenvector (line) and the state of the second eigenvector (points) in iteration 100 of the Gibbs sampler. Since an eigenvector is identified up to its sign, we fixed the sign of the true eigenvector such that it corresponds to the estimated eigenvector.

For this example it is clear that we can only recover the eigenvector in the $\gamma = 0.1$ network, as expected based on the spectra for both networks. However, it is also evident that the estimated eigenvector for the $\gamma = 1.0$ network exhibits a clear structure, and resembles the true eigenvector as a consequence of the results of Eckart and Young (1936). Upon inspecting the different states of the eigenvector in the $\gamma = 1.0$ network during estimation, the estimated eigenvector changed its shape and resembled different true eigenvectors between iterations. This was to be expected since the eigenvalues cannot be disentangled, and because of this we believe that the estimated eigenvector is a mix of the different true eigenvectors. The results for the other two eigenvectors were roughly the same.

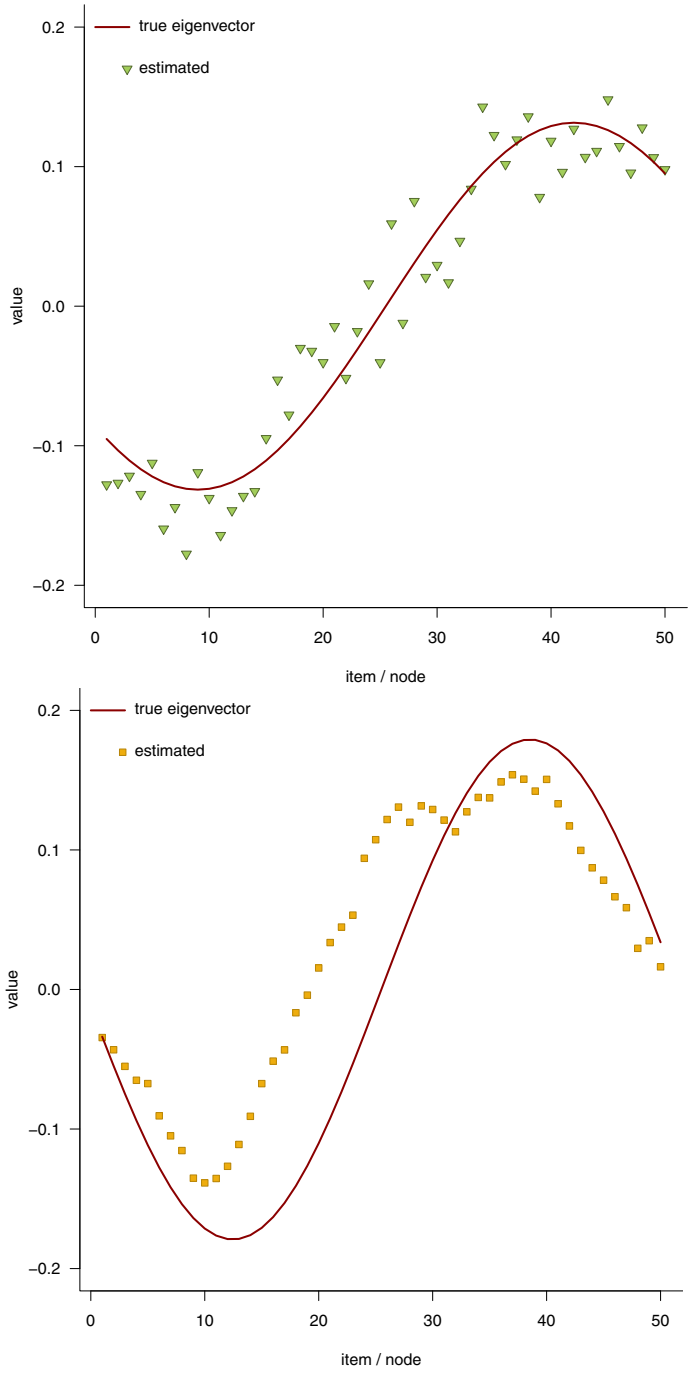


Figure F.2: True eigenvectors (lines) and state of the second eigenvector (points) in iteration 100 of the Gibbs sampler.

This result also illustrates that the low-rank approximation has more and more difficulty to recover the structure in the connectivity matrix (i.e. needs more and more data) as γ becomes larger, and it is here that a penalty based approach is likely to perform better and better, since the model tends to a nearest neighbour type network. On the other hand, if γ tends to zero, the low-rank approximation more easily recovers the underlying structure and it is here where the penalty based approach will have more difficulty. This model might give a direction for the development of mixtures of low-rank and penalty based approaches.

Appendix G

Chapter 4: Estimating the Rasch model

The full-conditional distribution of a difficulty parameter is of the form:

$$f(\delta|\mathbf{x},\boldsymbol{\theta}) \propto \prod_{p=1}^{n} P(x_p|\theta_p,\delta)f(\delta) = \prod_{p=1}^{n}(1-F_p(\delta))^{x_p}F_p(\delta)^{1-x_p}f(\delta),$$

where $F_p(\delta)$ is the cdf of a logistically distributed random variable with expectation θ_p and variance $\frac{\pi^2}{3}$. We sample $Z_p \sim F_p$, $p = 1,\ldots,n$, and Z_{n+1} from $f_{n+1}(\delta) = f(\delta)$, and define $x_p^* = (z_p > z_j)$. This gives the proposal:

$$f(z_j|\mathbf{x}^{*(j)},\boldsymbol{\theta}) \propto \prod_{p\neq j}^{n+1}(1-F_p(z_j))^{x_p^*}F_p(z_j)^{1-x_p^*}f_j(z_j).$$

Note that $x_+ = \sum_{p=1}^{n} x_p$ is the sufficient statistic for δ, and thus it is opportune to choose j such that $x_+^* = \sum_{p=1}^{n+1} x_p^*$ equals x_+ ($x_j^* \equiv 0$).

The proposed value $\delta^* = z_j$ is accepted as a draw from $f(\delta|\mathbf{x},\boldsymbol{\theta})$ with probability $\pi = \min\{1,\alpha\}$, with

$$\begin{aligned}\alpha =& \frac{f(\delta^*|\mathbf{x},\boldsymbol{\theta})f(\delta'|\mathbf{x}^*,\boldsymbol{\theta})}{f(\delta'|\mathbf{x},\boldsymbol{\theta})f(\delta^*|\mathbf{x}^*,\boldsymbol{\theta})}\\ =& \exp\left\{(\delta'-\delta^*)(x_{n+1}^*-x_j)\right\}\frac{(1-F_{n+1}(\delta'))^{x_{n+1}^*}F_{n+1}(\delta')^{1-x_{n+1}^*}}{(1-F_{n+1}(\delta^*))^{x_{n+1}^*}F_{n+1}(\delta^*)^{1-x_{n+1}^*}}\\ &\times\frac{(1-F_j(\delta^*))^{x_j}F_j(\delta^*)^{1-x_j}}{(1-F_j(\delta'))^{x_j}F_j(\delta')^{1-x_j}}\frac{f_{n+1}(\delta^*)f_j(\delta')}{f_{n+1}(\delta')f_j(\delta^*)},\end{aligned}$$

which equals one whenever $j = n+1$.

If the prior distribution of δ is that of a standard logistically distributed random variable, the expression for α is simplified considerably:

$$\alpha = \frac{(1+\exp\{\theta_j - \delta^*\})(1+\exp\{-\delta'\})}{(1+\exp\{\theta_j - \delta'\})(1+\exp\{-\delta^*\})},$$

which does not depend on the data and equals one whenever $j = n+1$.

The standard logistic prior distribution for δ is used in the GNU-R (R Core Team, 2014) code below.

```
#Generate augmented variables
  z = rlogis(n+1,location = c(theta,0),scale=1)
#Determine proposed value
  j = sum(x) + 1
  o = order(z,decreasing = TRUE)
  adelta = z[o[j]]
#Metropolis-Hastings step
  alpha = 0
  if(o[j] <= n)
  {
    alpha = log( 1 + exp( delta - theta[o[j]] ))
    alpha = alpha + log( 1 + exp( adelta ))
    alpha = alpha - log( 1 + exp( adelta - theta[o[j]] ))
    alpha = alpha - log( 1 + exp( delta ))
  }
  if(log(runif(1)) <= alpha) delta = adelta
```

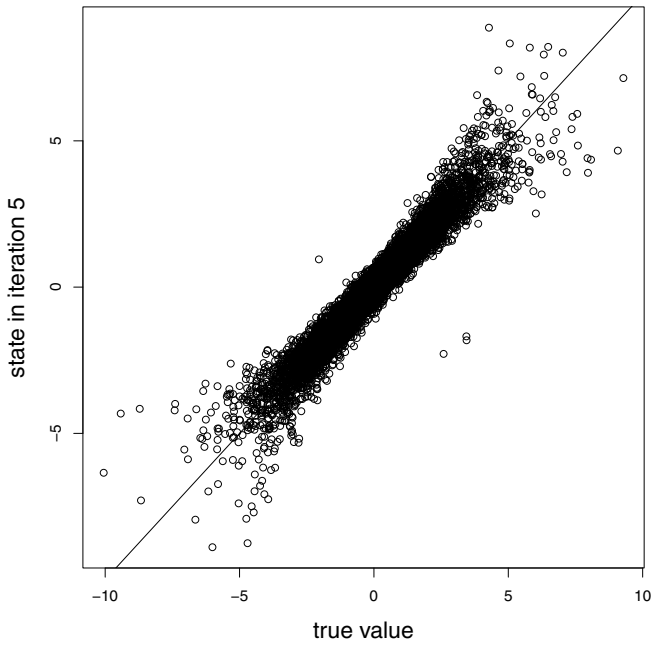
Small simulation

A small simulation serves to indicate the performance of this algorithm. In the GNU-R code below we sample the abilities of $10,000$ pupils from a standard logistic distribution, together with their responses to 40 Rasch items. The item difficulties were sampled uniformly between -1 and 1.

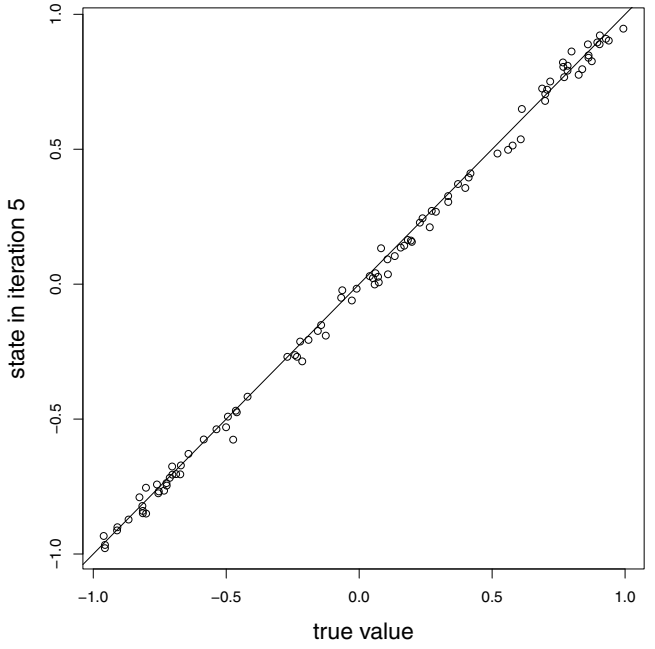
```
n = 10000
k = 40
theta = rlogis(n)
delta = runif(k,-1,1)
x = matrix(0,n,k)
for(i in 1:k) x[,i] = 1*(rlogis(n) <= (theta - delta[i]))
```

We use the standard logistic distribution as prior distribution for the pupil and item parameters, and use the algorithm in the GNU-R code above to sample from the full-conditional distributions of both the pupil and item parameters in a Gibbs sampler. The GNU-R code for this Gibbs sampler is provided below. We ran the Gibbs sampler for $10,000$ iterations, and the average acceptance rate for the pupil and item parameters was found to be 0.9885 and 0.9970, respectively. The Gibbs sampler converges very fast. To illustrate, the true parameter values are plotted against their states in iteration 5 of the Gibbs sampler in Figure G.1.

```
#Compute sufficient statistics and set starting values
Pscore = rowSums(x)
Iscore = colSums(x)
theta.e = rlogis(n)
delta.e = rlogis(k)
#Gibbs sampler
for(it in 1:1000)
{
#Generate pupil parameters
  for(p in 1:n)
  {
    z = rlogis(k+1,location = c(delta.e,0),scale=1)
    j = Pscore[p] + 1
    o = order(z)
    atheta = z[o[j]]
    alpha = 0
    if(o[j] <= k)
    {
      alpha = log( 1 + exp(atheta - delta.e[o[j]] ))
      alpha = alpha + log( 1 + exp( theta.e[p] ))
      alpha = alpha - log( 1 + exp( theta.e[p] - delta.e[o[j]] ))
      alpha = alpha - log( 1 + exp( atheta))
    }
    if(log(runif(1)) <= alpha) theta.e[p] = atheta
  }
#Generate item parameters
  for(i in 1:k)
  {
    z = rlogis(n+1,location = c(theta.e,0),scale=1)
    j = Iscore[i] + 1
    o = order(z,decreasing = TRUE)
    adelta = z[o[j]]
    alpha = 0
    if(o[j] <= n)
    {
      alpha = log( 1 + exp( theta.e[o[j]] - adelta ))
      alpha = alpha + log( 1 + exp( - delta.e[i] ))
      alpha = alpha - log( 1 + exp( theta.e[o[j]] - delta.e[i] ))
      alpha = alpha - log( 1 + exp( - adelta ))
    }
    if(log(runif(1)) <= alpha) delta.e[i] = adelta
  }
#Screen output
  par(mfrow=c(1,2))
  plot(theta,theta.e,main=it)
  points(delta,delta.e,pch=19,col="blue")
}
```



(a) pupil parameters



(b) item parameters

Figure G.1: Scatterplot of true values against parameter states in iteration 5 of the Gibbs sampler.

Appendix H

Chapter 5: Relation between DA-T Gibbs and Slice Sampling

Suppose we wish to generate a sample from a distribution $f(x) \propto a(x)b(x)$, where $a(x)$ is a density. The idea is to introduce an auxiliary random variable Z such that the joint distribution $f(x,z) \propto a(x)\,(0 < z \leq b(x))$. Clearly, the marginal density for X is $f(x)$ and a sample from $f(x,z)$ gives us a sample from $f(x)$. Sampling from the joint distribution can be done using a Gibbs sampler:

1. Sample z_i from $f(z|x_i) = U\,(0, b(x_{i-1}))$
2. Sample x_i from $f(x|z_i) = a(x)\,(z_i < b(x))$

This scheme includes several more recently developed samplers. When $a(x) \propto 1$ (i.e., when $f(x)$ is known upto a normalizing constant) we obtain the slice sampler by Neal (2003). If $b(x) = (l < x \leq u)$, we obtain Damien and Walker's (2001) Gibbs sampler for truncated distributions, with l and u, respectively, the lower and upper truncation points.

Damien et al. (1999) apply the idea to sample from a posterior distribution $f(\theta|\mathbf{x}_p) \propto f(\theta)\prod_i f(x_{pi}|\theta)$. Latent variables $Z_1, \ldots Z_n$ are then introduced, such that

$$f(\theta, \mathbf{z}|\mathbf{x}_p) \propto f(\theta)\prod_i (0 < z_i < f(x_{pi}|\theta)) \tag{H.1}$$

The Gibbs sampler then alternates between

1. Sample z_i from $f(z|\theta, \mathbf{x}_p) = U(0, f(x_{pi}|\theta))$
2. Sample θ from $f(\theta|\mathbf{z}, \mathbf{x}_p) \propto f(\theta) \prod_i (0 < z_i < f(x_{pi}|\theta))$

Let us consider the normal-ogive model again and derive a Gibbs sampler to sample from the posterior of ability for a person with response pattern $\mathbf{x}_p$. As in example 5 in Damien et al. (1999), we introduce latent variables $U_1, \ldots, U_k$ and $V_1, \ldots, V_k$ such that:

$$f(\theta, \mathbf{u}, \mathbf{v}|\mathbf{x}_p) \propto f(\theta) \prod_i (u_i < l_{1i}(\theta_p))(v_i < l_{2i}(\theta))$$

where $l_{1i} = \Phi(\theta)^{x_{pi}}$, and $l_{2i} = (1 - \Phi(\theta))^{1-x_{pi}} = \Phi(-\theta)^{1-x_{pi}}$. The full conditional densities for the latent variables are both uniform (i.e., $f(u_i|\mathbf{v}, \theta, \mathbf{x}_p) = U(0, l_{1i}(\theta))$) and the full-conditional for ability is a truncated normal distribution.

Thus, we obtain the DA-T Gibbs sampler. Specifically, the augmented posterior equals Equation 5.4 but with $z^*_{pi} = \Phi^{-1}(z_{pi})$. This means that the same behaviour that was observed for the DA-T sampler may also be found in DA Gibbs samplers based on the principle developed by Damien et al. (1999).

References

Albert, J. (1992). Bayesian estimation of Normal Ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polytomous response data. *Journal of the American Statistical Association*, *88*(422), 669-679.

Anderson, C., Li, Z., & Vermunt, J. (2007). Estimation of models in a Rasch family of polytomous items and multiple latent variables. *Journal of statistical software*, *20*(6).

Anderson, C., & Yu, H. (2007). Log-multiplicative association models as item response models. *Psychometrika*, *72*, 5-23.

Arnold, B., Balakrishnan, N., & Nagaraja, H. (2008). *A first course in order statistics*. Society for Industrial and Applied Mathematics.

Barabási, A. (2012). The network takeover. *Nature Physics*, *8*, 14-16.

Beaton, A., & Johnson, E. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement*, *29*, 163-175.

Béguin, A., & Glas, C. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541-562.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (methodological)*, *36*, 192-236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, *24*, 179-195.

Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 179-197.

Bock, R., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179-197.

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Brooks, S. (2003). Bayesian computation: a statistical revolution. *Philosophical Transactions of the Royal Society of London A*, *361*, 2681-2697.

Cappe, O., & Robert, C. (2000). Markov chain Monte Carlo: 10 years and still running! *Journal of the American Statistical Association*, *95*, 1282-1286.

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167-174.

Chang, H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37-52.

Cover, T., & Thomas, J. (1991). *Elements of information theory.* Wiley-Interscience.

Cramer, A., Waldorp, L., van der Maas, H., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, *33*, 137-150.

Cressie, N., & Holland, P. (1983). Characterizing the manifest probabilities of latent variable models. *Psychometrika*, *48*, 129-141.

Creutz, M. (1980). Monte Carlo study of quantized SU (2) gauge theory. *Physical Review*, *21*, 2308-2315.

Damien, P., Wakefield, J., & Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Society. Series B*, *61*, 331-344.

Damien, P., & Walker, S. (2001). Sampling truncated Normal, Beta and Gamma densities. *Journal of Computational and Graphical Statistics*, *10*, 206-215.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability.* Springer.

Dawid, A. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B (Methodological)*, *41*, 1-31.

Deary, I. (2000). *Looking down on human intelligence: From psychometrics to the brain.* Oxford, England: Oxford University Press.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211-218.

Emch, G., & Knops, H. (1970). Pure thermodynamical phases as extremal KMS states. *Journal of mathematical physics*, *11*, 3008-3018.

Fox, J. (2013). Multivariate zero-inflated modeling with latent predictors: Modeling feedback behavior. *Computational Statistics and Data analysis*, *68*, 361-374.

Fox, J., & Glas, C. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271-288.

Gelfand, A., & Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelman, A., Carlin, B., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*

(Second ed.). Chapman & Hall/CRC.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97-109.

Hoijtink, H., & Molenaar, I. (1997). A multidimensional item response model: constrained latent class analyses using the Gibbs sampler and posterior predictive checks. *Psychometrika*, *62*, 171-190.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, *31*(1), 253-258.

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

Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620-630.

Jensen, A. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Kac, M. (1968). Mathematical mechanisms of phase transitions. In M. Chretien, E. Gross, & S. Deser (Eds.), *Statistical physics: Phase transitions and superfluidity, vol. 1, Brandeis university summer institute in theoretical physics.* (p. 241-305). New York: Gordon and Breach Science Publishers.

Klein Entink, R., Fox, J., & van der Linden, W. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21-48.

Kreiner, S., & Christensen, K. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*, 210-231.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79-86.

Lauritzen, S., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The annals of statistics*, *17*(1), 31-57.

Lee, T., & Yang, C. (1952). Statistical theory of equations of state and phase transitions II. Lattice gas and Ising model. *Physical review*, *87*(3), 410-419.

Liu, J. (1994). Fraction of missing information and convergence rate of data augmentation. In J. Sall & A. Lehmann (Eds.), *Computationally intensive statistical methods: Proceedings of the 26th symposium interface* (p. 490-497).

Liu, J., Wong, W., & Kong, A. (1994). Covariance structure of the Gibbs sam-

pler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, *81*, 27-40.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maris, G. (2012). Analyses. In N. Jones et al. (Eds.), *First european survey on language competences* (p. 298-331). European Commission.

Maris, G., & Maris, E. (2002). A MCMC-method for models with continuous latent responses. *Psychometrika*, *67*, 335-350.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615-633.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-175.

McCullagh, P. (1994). Exponential mixtures and quadratic exponential families. *Biometrika*, *81*, 721-729.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., & Teller, A. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*, 1087-1092.

Mislevy, R. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177-196.

Mislevy, R. (1993). Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika*, *58*, 79-85.

Mislevy, R., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133-161.

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131-154.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Pschological Measurement*, *16*, 159-176.

Murray, I., Ghahramani, Z., & MacKay, D. (2012, August). MCMC for doubly-intractable distributions. *ArXiv e-prints*.

Neal, R. (2003). Slice sampling. *Annals of Statistics*, *31*, 705-767.

Olkin, I., & Tate, R. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The annals of mathematical statistics*, *32*(2), 448-465.

Patz, R., & Junker, B. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366.

Patz, R., & Junker, B. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and*

Behavioral Statistics, *24*(2), 146-178.

Potts, R. (1952). Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, *48*(1), 106-109.

Propp, J., & Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, *9*, 223-252.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* MESA Press, Chicago.

Reckase, M. (2009). *Multidimensional item response theory.* Springer.

Richardson, M. (1936). The relationship between difficulty and the differential vailidity of a test. *Psychometrika*, *1*, 33-49.

Robert, C., & Cassela, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, *26*, 102 - 115.

Rosenthal, J. (2011). Optimal proposal distributions and adaptive MCMC. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), (p. 93-112). Chapman & Hall.

Roy, V., & Hobert, J. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society, Series B*, *69*, 607-623.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151-1172.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New-York: Wiley.

Rubinstein, R. (1981). *Simulation and the Monte Carlo method.* New York: John Wiley & Sons, Inc.

Swendsen, R., & Wang, J. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, *58*, 86-88.

Tanner, M. (1993). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (second ed.). Springer-Verlag.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528-540.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, *22*(4), 1701-1762.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Annals of applied probability*, *8*(1), 1-9.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics*, *9*, 23-30.

van Dyk, D., & Meng, X. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*, 1-50.

van Borkulo, C., Borsboom, D., Waldorp, L., Epskamp, S., Blanken, T., Boschloo, L., & Schoevers, R. (in press). A new method for constructing networks from binary data. *Scientific Reports*.

van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308.

van der Maas, H., Dolan, C., Grasman, R., Wicherts, J., Huizenga, H., & Raijmakers, M. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842-861.

van der Maas, H., & Wagenmakers, E. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology*, *118*, 29-60.

Verhelst, N., & Glas, C. (1995). The one parameter logistic model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (p. 215-238). New York: Springer Verlag.

Verhelst, N., Glas, C., & Verstralen, H. (1995). OPLM: computer program and manual [Computer software manual]. Arnhem, the Netherlands.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large scale assessments (vol. 2).* IEA-ETS Research Institute.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1-25.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Education Evaluation*, *31*, 114-128.

Zeger, S., & Karim, M. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, *86*, 79-86.

Samenvatting

In onderwijskundige peilingen, zoals het *programme for international student assessment* (PISA)[1], het *national assessment of educational progress* (NAEP)[2] en de *European survey on language competences* (ESLC)[3], worden *plausible values* gebruikt ter facilitatie van secundaire analyses door onderzoekers die het de middelen ontbreekt om de latente regressie modellen te schatten. De theorie achter plausible value imputatie is ontwikkeld door Rubin (1987), en is toegepast op grootschalige peilingen door Mislevy en collega's (Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992; Mislevy, Beaton, et al., 1992; Beaton & Johnson, 1992; Mislevy, 1993; von Davier et al., 2009).

Een plausible value voor een kind p is een trekking uit de posterior verdeling van zijn of haar (doorgaans ééndimensionele) vaardigheid θ_p, gegeven zijn of haar item respons vector $\mathbf{x}_p$ en aanvullende informatie beschreven in een vector met covariaten $\mathbf{y}_p$. De posterior verdeling is gegeven door

$$f(\theta|\mathbf{x}_p, \mathbf{y}_p) \propto P(\mathbf{x}_p|\theta, \boldsymbol{\delta}) f(\theta|\boldsymbol{\lambda}, \mathbf{y}_p) \tag{H.2}$$

waarbij $P(\mathbf{x}_p|\theta, \boldsymbol{\delta})$ een *item respons theorie* (IRT) model is met item parameters $\boldsymbol{\delta}$, en $f(\theta|\boldsymbol{\lambda}, \mathbf{y})$ een populatie model met parameters $\boldsymbol{\lambda}$.

Het schatten van het populatiemodel is het primaire doel van analyses in onderwijskundige peilingen, en is doorgaans van de vorm

$$\theta = \mathbf{y}^T\boldsymbol{\beta} + \epsilon, \text{ met } \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2). \tag{H.3}$$

Voor het schatten van het populatie model is normaliter gespecialiseerde software en expertise nodig, maar het kan ook geschat worden met behulp van de plausible values. Het latente regressie model in (H.3) kan bijvoorbeeld geschat worden door herhaaldelijk plausible values als afhankelijke variabele te gebruiken, en de resultaten te aggregeren over de replicaties. Het gebruik van plausible values in deze context is verwarrend, en het is onduidelijk waarom plausible values nodig zijn om het populatie model te schatten als het populatie model al geschat is om de plausible values te maken. Of, anders gezegd: *Wat kunnen we leren van plausible values buiten dat wat al gekend is uit het populatie model?*

In Hoofdstuk 2 wordt beargumenteerd dat plausible values meer zijn dan slechts een middel om secundaire analyses te faciliteren. Plausible values zouden een meer

[1] `www.oecd.org/pisa/`

[2] `nces.ed.gov/nationsreportcard/`

[3] `www.surveylang.org`

centrale rol moeten spelen in populatie inferentie in peilingsonderzoek, omdat ze informatie bevatten die de informatie in het populatie model domineert. Hierbij wordt het populatie model gezien als een prior verdeling. Er wordt aangetoond dat, gegeven het IRT model en het populatie model, de marginale verdeling van plausible values een betere schatter is van de onbekende vaardigheidsverdeling dan het populatie model, waarbij het populatie model slechts een middel is om plausible values te kunnen genereren.

Een interessante vraag is waarom de naam "plausible values" gebruikt wordt, en niet "trekkingen uit de posterior verdeling van vaardigheid". De werkelijke reden is ons onbekend, maar heeft ongetwijfeld te maken met het gebrek aan acceptatie van Bayesiaanse methoden op het moment van introduceren van plausible values in de psychometrische literatuur. Tegenwoordig worden Bayesiaanse methoden veel toegepast en zijn ze breed geaccepteerd, zelfs in de psychometrie. Een reden voor deze verandering is de zogeheten *MCMC revolutie* (Cappe & Robert, 2000; Brooks, 2003; Robert & Cassela, 2011), waarbij men inzag dat complexe Bayesiaanse modellen geschat konden worden met behulp van Markov chain Monte Carlo (MCMC) methoden. Hoewel de Bayesiaanse statistiek en MCMC methoden al veel langer bestonden, maakte de snel stijgende rekenkracht van computers het gebruik van de rekenintensieve MCMC methoden ineens bruikbaar. MCMC methoden zijn dan ook pas populair geworden in de statistiek eind jaren 80 begin jaren 90 (Gelfand & Smith, 1990; Zeger & Karim, 1991; Casella & George, 1992; Albert & Chib, 1993; Tierney, 1994). Het was met name de Gibbs sampler (Geman & Geman, 1984) die in deze tijd veel aandacht kreeg, omdat via deze methode gesimuleerd kon worden uit complexe multivariate (posterior) verdelingen, en hierdoor modellen beschikbaar maakte die in een (marginal) maximum likelihood raamwerk (Bock & Lieberman, 1970; Bock & Aitken, 1981) niet te schatten zijn. De MCMC revolutie is ook terug te vinden in de psychometrie, waarin het populair werd door het werk van Albert (1992) en Patz and Junker (1999b), en verder toegepast op het schatten van testlet modellen (Bradlow et al., 1999), latente klasse modellen (Hoijtink & Molenaar, 1997), multilevel IRT modellen (Fox & Glas, 2001), random item parameters (Janssen et al., 2000), en multidimensionele IRT modellen (Béguin & Glas, 2001).

De huidige *big-data revolutie* biedt een nieuwe uitdaging voor MCMC methoden, waarbij de methoden toegepast worden op problemen van steeds grotere omvang. Een voorbeeld is het PISA peilingsonderzoek uit 2012, waarbij voor meer dan $500,000$ kinderen plausible values nodig waren, en gecombineerd met de complexiteit van het design van de peiling, het IRT model en het populatie model een indrukwekkende toepassing is. In Hoofdstuk 5 wordt aangetoond dat enkele bestaande methoden niet bruikbaar zijn voor toepassingen van deze omvang, wat betekent dat gespecialiseerde methoden ontwikkeld moeten worden. In Hoofdstuk 3 en 4 worden voor dit doel nieuwe MCMC methoden geïntroduceerd. Allereerst worden in Hoofdstuk 3 twee bestaande methoden aangepast, zodanig dat ze efficiënt worden voor big data toepassingen. De methoden in Hoofdstuk 3 zijn met name geschikt voor toepassingen waarbij er veel replicaties zijn van een random effect, zoals de vaardigheden van kinderen in onderwijskundige peilingen. Een methode die geschikt is voor toepassingen waarbij er weinig replicaties zijn van een random effect, zoals de item parameters in onderwijskundige peilingen, wordt

geïntroduceerd in Hoofdstuk 4. De methode in Hoofdstuk 4 wordt efficiënter als er meer en meer observaties over het random effect zijn, en doorgaans hebben we veel observaties op items in onderwijskundige peilingen.

De stijging in de omvang van toepassingen leidt ook tot het gebruik van complexere statistische modellen. Een recente ontwikkeling op dit gebied is het gebruik van netwerk modellen (Barabási, 2012), welke reeds worden toegepast in de psychometrie (Cramer et al., 2010; van Borkulo et al., in press). Het netwerk model van Ising (1925) is hierbij interessant, daar we de meeste MCMC methoden van vandaag de dag danken aan dit model[4]. In Hoofdstuk 4 wordt aangetoond dat het Ising model kan worden gekarakteriseerd als een marginaal IRT model door middel van Mark Kac's (1968) latente variable representatie. Deze relatie opent de weg voor nieuwe onderzoeksrichtingen, en in Hoofdstuk 4 wordt deze relatie gecombineerd met een *full-data-information* procedure om het Ising model te schatten. Het schatten van het Ising model is een notoir lastig probleem omdat de normalizerende constante van het Ising model niet uit te rekenen is, en omdat het model een groot aantal onbekenden heeft. De full-data-information procedure omzeilt het probleem van het uitrekenen van de normalizerende constante, waardoor het schatten van het Ising model computationeel mogelijk is, zelfs voor grote netwerken. Echter, in een netwerk met n variabelen zijn er $n(n+1)/2$ onbekenden, en zelfs voor relatief kleine netwerken zijn er een groot aantal replicaties van het netwerk nodig om alle onbekenden te schatten. In Hoofdstuk 4 wordt daarom gebruik gemaakt van een lage rang benadering van de matrix van paarsgewijze interacties, waarin alleen de grootste eigenwaarden van de matrix geschat worden en de overige eigenwaarden op nul worden gezet. Een dergelijke benadering is zeer bruikbaar voor sterk verbonden netwerken, zoals de netwerken die we doorgaans vinden in de sociale wetenschappen, waarbij een klein aantal eigenwaarden van de matrix het grootste gedeelte van de onderliggende structuur bevatten.

[4]MCMC methoden zoals Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970), Gibbs sampling (Creutz, 1980; Rubinstein, 1981), Perfect Sampling (Propp & Wilson, 1996), en Data Augmentatie (Swendsen & Wang, 1987), zijn ontwikkeld om data te simuleren uit het Ising model.