A Cautionary Note on Data Augmentation Algorithms

Katerina Papadimitropoulou

MSc in Mathematics

Specialization: Statistical Science for the Life and Behavioural Sciences

Thesis Advisors

CITO:
Dr. Timo Bechger
Maarten Marsman

Leiden University:
Prof. Dr. Willem Heiser
Prof. Dr. Richard Gill

Abstract

For the applied statistician, data augmentation is a powerful tool for solving optimization problems. In this thesis, I address a problem in some data augmented Gibbs samplers. I show that although introducing latent variables renders a sampling problem tractable, this comes at the price of raising the autocorrelation of the Markov chain, as the number of parameters increases, in this case the number of items in a test. By means of an example, I show that data augmentation is a powerful yet inefficient tool in cases of increased number of items, since the autocorrelation (and hence the rate of the convergence) of the addressed augmented Gibbs sampler is proved to be dependent on the number of item parameters. We wish to show that although most data-augmented samplers are well behaved, in this example the algorithm becomes really slow and faces the possibility of grinding to a halt.
KEY WORDS: Gibbs sampler, Posterior sampling, Autocorrelation

1 Acknowledgments

I would like to express my deep gratitude to Maarten Marsman and Timo Becher, my thesis supervisors from Cito, Dutch Institute for Educational Measurement, for their patient guidance and their encouragement throughout my internship and thesis project. I would like to thank them for the interesting problem they provided me and their ideas on how to address it. My grateful thanks are also extented to Gunter Maris for the discussions we had on the data augmentation and data augmentation-transformation algorithm and especially since the material of section 5.1 of the thesis is based on yet unpublished work of all three of them. I would like to express my very great appreciation to Richard Gill, my thesis supervisor from the Mathematical Institute of Leiden University for the generous amount of time he spent on my thesis, providing helpful comments and suggestions during the preparation of the final version of the thesis. Special thanks should be given to Willem Heiser, my supervisor from Leiden University for his constructive feedback and comments throughout the writing of the thesis. I would also like to thank Peter Grunwald for finding some time to read my thesis and share constructive suggestions with me that were incorporated in the final version of the thesis.

Contents

2 Introduction

Data augmentation (DA) is a statistical tool for constructing sampling and optimization algorithms by introducing unobserved or latent variables. Ever since the seminal paper of Tanner and Wong (1987), in which the authors introduced the term *data augmentation algorithm*, an increasing number of difficult statistical problems has been addressed and has yielded results that could not be obtained so far. Since then, the applications of the DA algorithm have been numerous and diverse. Tanner and Wong used DA schemes for posterior calculation, Swendsen and Wang (1987) to sample from the Ising model, Albert (1992) to estimate Item Response Theory (IRT) models and Albert and Chib (1993) for probit regression, to name a few.

In this thesis we study the asymptotic behavior of data augmented Gibbs samplers. Our objective is to show that the autocorrelation between successive samples may depend on the number of items in a test. To this aim, we discuss an example where the Markov chain will eventually stop mixing as more items are being observed. Note that most data-augmented samplers are well behaved and normally become slower when the amount of latent data increases. In the content of this thesis, we will address an example where the algorithm might grind to a halt and therefore fail in exploring the posterior support.

The thesis is organized as follows: In Chapter 2, an introduction and background information of the problem are given. We briefly describe the Gibbs sampler and key concepts of item response theory. In Chapter 3, the data augmentation and data augmentation-transformation algorithms are explained thoroughly. In Chapter 4, we study the asymptotic behavior of the algorithm for the Normal Ogive and the Rasch model. Chapter 5, describes the relation of lag-1 autocorrelation to the sample size in terms of increase of item parameters of a test. The paper ends with a discussion. The R-code (R Core Team, 2013) used to obtain the figures in this thesis is presented in Appendix C.

3 Introduction and Background Concepts of the Problem

3.1 Introduction and Model Specification

We consider the situation of a random sample of N persons; each responds to items. The items are constructed to measure some sort of ability and are scored as right or wrong. Thus, for each person $p = 1, ..., N$ and item $i = 1, ..., n$ we observe one realization x_{pi} of a Bernoulli response variable X_{pi}, where $x_{pi} = 1$ if the response of person p to item i was correct, and $x_{pi} = 0$ if the answer was incorrect, respectively.

It is assumed that each person in the sample is characterized by a unidimensional ability denoted θ, sampled independently from a population model $f(\theta|\lambda)$ as a function of a parameter λ. An **Item Response Theory model** is used for the conditional distribution, $P(\mathbf{X}_p = \mathbf{x}_p|\theta, \boldsymbol{\delta})$, of the response vector $\mathbf{X}_p = \{X_{p1}, X_{p2}, \ldots, X_{pN}\}$ of a person p as a function of ability θ, and item parameters $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_n\}$. Conditional on ability, the responses are assumed to be independent, and together the IRT model and population model induce the following statistical model:

$$P(\mathbf{X}_p = \mathbf{x}_p|\underline{\delta}, \lambda) = \int_{\mathbb{R}} \prod_i P(X_{pi} = x_{pi}|\delta_i, \theta) f(\theta|\lambda) d\theta, \tag{1}$$

called a marginal IRT model.

We consider two simple but commonly used IRT models:

1. The 1-Parameter Normal Ogive model:

   $P(X_{pi} = 1|\theta_p, \delta_i) = \Phi(\theta_p - \delta_i)$, where Φ is the cumulative distribution function of the standard normal distribution

2. The Rasch model (1960):

   $P(X_{pi} = 1|\theta_p, \delta_i) = \frac{e^{\theta_p - \delta_i}}{1+e^{\theta_p - \delta_i}}$

A simple argument leading to these models is as follows: Assume that there exists a continuous **latent response variable** Z_{pi} such that the person p solves item i if Z_{pi} is larger than a threshold δ_i. That is:

$$P(X_{pi} = 1|\theta_p, \delta_i) = P(Z_{pi} > \delta_i|\theta_p)$$

It is seen that the probability of a correct response depends on the threshold as well as the ability of the respondent. Depending on the distribution of the latent response variable we obtain either of the aforementioned models:

$$Z_{pi}|\theta \sim \begin{cases} \mathrm{N}(\theta_p, 1) & \text{gives the Normal Ogive model} \\ \mathrm{L}(\theta_p, 1) & \text{gives the Rasch model} \end{cases}$$

where N stands for the Normal distribution and L for the Logistic distribution, respectively. Thus, both models can be derived from a similar argument involving very little theory about the response process. Note that the Rasch model belongs to an exponential family, while the Normal Ogive model does not. The Rasch model is an exponential family IRT model and all information about the ability in the response vector X_p is contained in the sufficient statistic $\sum_i x_{pi} = x_{p+}$, the number of correct responses. In the context of this thesis we will focus more on the Rasch model yet some concepts will be explained also for the Normal Ogive in terms of comparison of the data augmentation algorithm for the two models.

3.2 Item Response Theory and the Rasch model

Each IRT model predicts the probability that a certain person will give a certain response to a certain item. People can have different levels of ability and items can differ in terms of difficulty. We are always interested in calculating the probability of a correct response $\mathrm{P}(\theta)$, which is a function of the ability θ.

The Rasch model (1960) is a model for dichotomous responses. It is also called one-parameter logistic (1PL) model and is the simplest IRT model for a dichotomous item since it has only one item parameter. The item response function (i.e. the probability of a correct response given the item parameter δ_i and the individual ability θ) is given on Figure 1. The function shown in the graph is known as the one-parameter logistic function and it has the mathematical property that its values are between 0 and 1 for any argument in $(-\infty, \infty)$. Therefore it is obvious that the probability of a

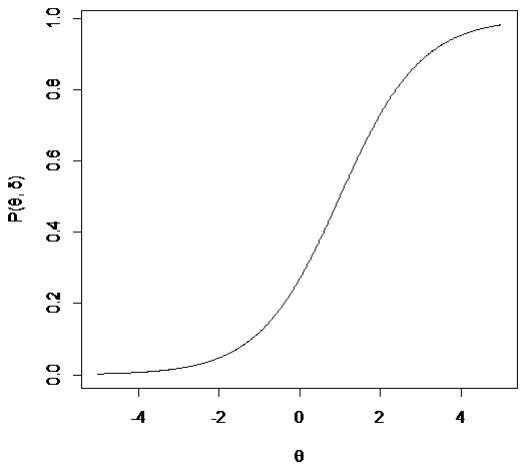


Figure 1: The item response function of the Rasch model

correct response is modeled as a logistic function of the difference between the person and the item parameter. Note that we denoted the horizontal axis with the ability θ, but is also the axis for the difficulty parameter δ. One can find the position of δ on the shared ability/ difficulty axis at the point for which the predicted probability $\mathrm{P}(\theta - \delta)$ equals 0.5. It is apparent from the plot that the higher a person's ability relative to the difficulty of an item, the higher the probability of a correct response on that item.

In the context of IRT analysis, difficulty and ability (trait) level are separate issues but are intrinsically connected. Specifically, as already stated, a difficult item requires a relatively high ability level in order to be answered

correctly compared to an easy item. For example, consider two mathematical items: the item "What is the square root of 10000?" is less likely to be answered correctly than the item "What is the square root of 25?".

3.2.1 Item Characteristic Curves

In an IRT analysis, item characteristics are combined in order to reflect characteristics of the test as a whole. In this context, item characteristics such as difficulty and discrimination can be used to evaluate the items and to maximize the overall quality of a test.

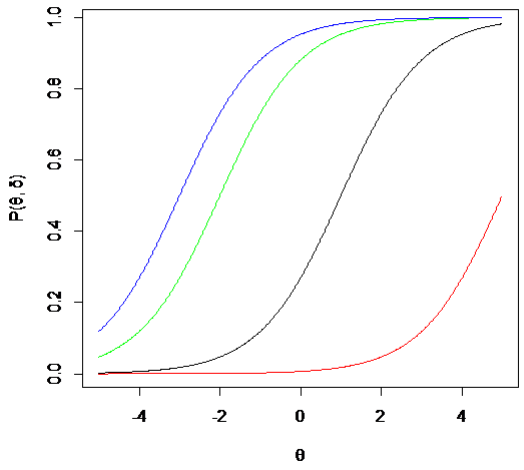


Figure 2: Item Characteristic Curves

In the item characteristic curves-plot, which is presented in Figure 2, the X-axis reflects a wide range of ability levels and the Y-axis reflects the probabilities of the correct response. Each item has a curve, and we can examine the curve of an item to find the likelihood that an individual with a particular trait level will answer the item correctly. Take a moment to study the curve for Item 1: what is the probability that an individual with an average level of mathematical ability will answer the item correctly? We find the point on the Item 1 (depicted in blue) curve that is directly above the 0 point on the X-axis (recall that the trait level is in z score units, so zero is the average trait level), and we see that this point lies between 0.80 and 0.90 on the Y-axis. Now have a look at the third item (depicted in black); an individual with an average level of mathematical ability has a probability of about 30% of answering it correctly. Thus, the item characteristic curves provide clues about the likelihoods with which individuals of any ability level would answer any of the five items correctly. Note that the order of

the curves, from left to right on the X-axis, reflects their difficulty levels. Item 1 (blue), with the left-most curve, is the easiest item, and Item 4(red), with the right-most curve, is the most difficult item.

3.3 The Gibbs sampler

In a Bayesian framework, the parameters of the model θ, δ, λ are viewed as random variables. Inferences about the parameters are made in terms of their posterior distribution. However, in many cases the simultaneous posterior distribution of all model parameters can be quite complicated. The major advantage of Bayesian estimation is that it allows us to study the properties of the posterior by drawing a sample from it. In fact, all the properties of a posterior distribution can be approximated to a degree of accuracy by drawing a sample that is sufficiently large. This approach to statistical estimation is called *sampling-based estimation*. Sampling-based estimation enables one to study distributions that are analytically intractable, given that one can sample from them. For the cases that the posterior distribution is intractable, there is extensive literature of remedies in order to render the intractable distributions tractable.

In the last decade, much attention has been paid to Markov chain Monte Carlo (MCMC) methods for generating a sample from a posterior (Gelman, Carlin, Stern, & Rubin, 2004). These methods involve (a) setting up a Markov chain which in the limit generates a dependent identically distributed (did) sample from the posterior and (b) the use of the Monte Carlo method for estimating properties of the did sample. Three important questions rise to the scientist wishing to use a MCMC-method for a particular problem:

- How to set up a Markov chain which converges to a did sample from the posterior
- How to assess whether the length of the Markov chain is sufficient for it to be adequately close to its stationary distribution
- How to assess whether the sample size (after convergence) is sufficient for the Monte Carlo estimates to be sufficiently precise

Within the scope of this thesis, we focus on the first question and more precisely in the estimation of the parameters of the Rasch model. Therefore for the cases that the posterior distribution is intractable, there is extensive literature of remedies in order to render the intractable distributions tractable. Some of the most popular approaches are: the acceptance-rejection sampling, the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, & Teller, 1953) and data augmentation schemes (Tanner & Wong, 1987).

Here, we focus on one MCMC method, Gibbs sampling, to generate a sample from the posterior distribution. The Gibbs sampler (introduced in the context of image processing by Geman & Geman, 1984 and also known as the *heat-bath* algorithm) is a Markov Chain Monte Carlo (MCMC) algorithm, special case of Metropolis-Hastings sampling wherein the random variable is always accepted. In each iteration, we generate a sample from each of the **full-conditional distributions** of the posterior; i.e., the distribution of a parameter (or set of parameters) conditional on the data and all the other parameters. Typically, the full conditionals are of moderate dimensionality. Hence the complete set of parameters is divided into a number of subsets in such a way that the distribution of every subset given all other parameters has a tractable form and can be easily simulated. Thus, one simulates n random variables sequentially from n univariate conditionals rather than generating a single n-dimensional vector in a single pass using the full joint distribution.[1]

As with other MCMC methods, draws produced at subsequent iterations are dependent and the method produces a Markov chain of samples. The Gibbs sequence converges to a stationary (equilibrium) distribution that is independent of the starting values and by construction this stationary distribution is the target distribution we are trying to simulate, under mild regularity conditions (see, e.g., Tierney, 1994).

In order to explain how the Gibbs sampler works, I provide an example in which we are interested in sampling from the posterior $p(\theta|y)$, where θ is a vector of three parameters, $\theta_1, \theta_2, \theta_3$. The steps to a Gibbs sampler are the following:

1. Pick a vector of starting values $\theta^{(0)}$. (Defining a starting distribution $\Pi^{(0)}$ and drawing $\theta^{(0)}$ from it.)

2. Start with any θ, the order does not influence the result but for the terms of the example I start with $\theta_1^{(1)}$. Draw a value from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, \mathbf{y})$.

3. Draw a value $\theta_2^{(1)}$ from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \mathbf{y})$. Note that we must use the updated value of $\theta_1^{(1)}$.

4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, \mathbf{y})$ using both updated values.

5. Draw $\theta^{(2)}$ using $\theta^{(1)}$ as in steps 2 to 4

6. Repeat until we get M draws, where each draw is a vector $\theta^{(t)}$.

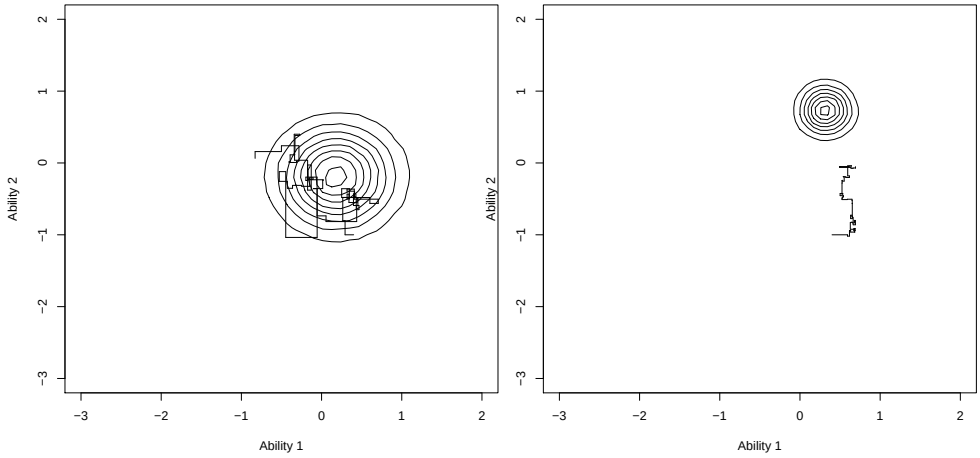
7. Optional burn-in and/or thinning.

[1] One iteration of all univariate distributions is often called a **scan** of the sampler.

Therefore, our result is a Markov chain with some draws of θ that are approximately from our posterior distribution.

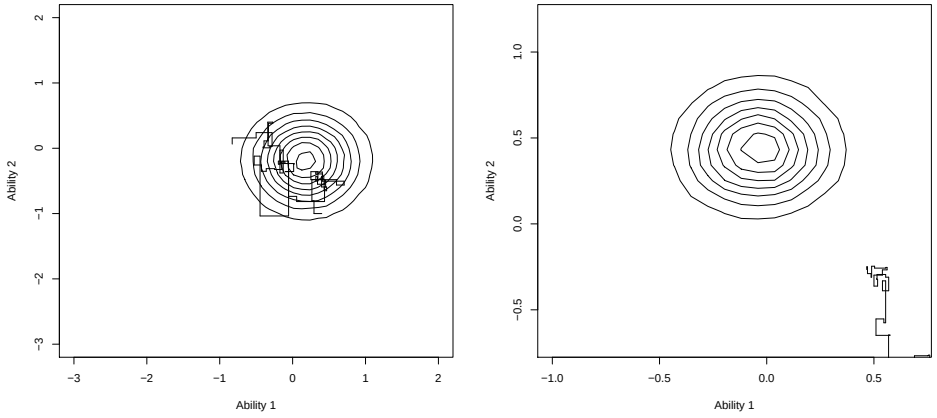
Gelfand & Smith, 1990 illustrated the power of the Gibbs sampler to address a wide variety of statistical issues while further details can be found in Casella & George (1992), and Tanner & Wong (1987). Finally, note that the Gibbs sampler can be thought of as a stochastic analogue to the EM (Expectation-Maximization) algorithm used to maximize likelihood functions when missing data are present. In the sampler, random sampling replaces the expectation and maximization steps.

As an illustration, Figure 1 shows contour plots of 50 draws of two abilities produced by a Gibbs sampler where successive draws are connected by a line segment. This example describes the situation where two students independently took a test with items designed to measure one specific ability and a Gibbs sampler to compute the joint posterior density of ability. The left panel of Figure 1 shows the result for a test with 20 items, and the right panel for a test that is five times longer as witnessed by the posterior density that is more concentrated around the true ability. Both plots depict simulations of the data augmentation-transformation Gibbs sampler of the Rasch model which is described in detail in section 3.1 of this thesis. In both plots, the sampler moves from an initial guess to the support of the posterior. A well-behaved Gibbs sampler is then expected to "walk around" producing a (dependent) sample from the posterior. The more item responses we observe the more we know about a person's ability and this is reflected in the tighter support of the posterior in the right panel. The interesting observation is that the sampler takes smaller steps in the right panel. This means 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 *relatively to the posterior variance*. That this might be the case here, is suggested by Figure 2. It would imply that sampling is less effective for the test of 100 items; even if the chain has converged and we start inside the posterior support. Furthermore, if it is true that the step-size diminishes with the number of observations at a faster rate than the posterior variance, the sampler will eventually grind to a halt.



(a) n=20 Items (b) n=100 Items

Figure 3: First 50 samples produced by the Gibbs sampler connected by a line segment. Contour plots of the posterior densities of two abilities, where $\theta_1 = 0.25$ and $\theta_2 = 0.5$.



(a) n=20 Items (b) n=100 Items

Figure 4: First 50 samples produced by the Gibbs sampler connected by a line segment. Contour plots of the posterior densities of two abilities, where $\theta_1 = 0.25$ and $\theta_2 = 0.5$. Note that the scales in the two panels are different; the right panel is plotted in a smaller scale. We can easily notice that the steps get smaller faster than the posterior tightens.

3.4 The Data Augmentation Algorithm

Unfortunately, straightforward application of the Gibbs sampler to IRT models does not lead to tractable full conditional distributions. Albert (1992) noted that the Gibbs sampler becomes feasible if we include the latent responses. Though the joint distribution of (Z, θ, δ) has an intractable form, the full conditional distributions of each of the three parameters are easy to sample from. The introduction of latent data is often useful in Gibbs sampling and is called *Data Augmentation (DA)*. The DA Gibbs sampler alternates samples from the conditional distributions of the latent data denoted Z, and parameters θ and δ, given the remaining variables as seen in Algorithm 1.

Algorithm 1 A Gibbs sampler algorithm for a marginal IRT model

for $iter = 1$ to $nIter$ **do**
  {Sample latent data.}
  for $i = 1$ to M **do**
    Sample $z_{pi}^{(iter)}$ from $f(z_{pi}|x_{pi}, \underline{\theta}^{(iter-1)}, \underline{\delta}^{(iter-1)})$.
  end for
  {Sample person parameters.}
  for $p = 1$ to N **do**
    Sample $\theta_p^{(iter)}$ from $f(\theta_p|x_{pi}, \underline{z}^{(iter)}, \underline{\delta}^{(iter-1)})$.
  end for
  {Sample item parameters.}
  for $i = 1$ to M **do**
    Sample $\delta_i^{(iter)}$ from $f(\delta_i|x_{pi}, \underline{z}^{(iter)}, \underline{\theta}^{(iter)})$.
  end for
end for

From the algorithm we can see that the step of sampling first the latent data can be seen as the imputation step and the steps of sampling from the ability parameter distribution and the difficulty parameter distribution as the posterior steps.

In the context of this report, we are only focused on sampling from the full conditional of the ability parameter θ assuming that δ is known. Therefore, the iterative scheme alternates the steps of first sampling from the full conditional of the latent data and then from the full conditional of θ. We should also note that we assume that the distribution of ability is known.

4 Transformation and Asymptotic Behavior

4.1 The Data Augmentation-Transformation Algorithm for the Rasch model

As we described in the previous section, the sampling scheme is iterative and alternates between sampling from the full conditional of the latent responses and the full conditional of the ability parameter. At this point, note that when conjucacy between the distribution of the latent data and the prior of ability does not hold or when the product of distributions is not tractable, sampling from the full conditionals is not possible. Whereas Albert (1992) found that the full conditionals for the DA-Gibbs sampler for the Normal Ogive model were all simple, this is not the case for the Rasch model. For the Normal Ogive model, the distribution of the latent data and the prior of the parameters combine in an nice conjugate manner, for each parameter, whereas this is not the case for the Rasch model. To tackle this problem Maris and Maris (2002) proposed the Data augmentation-transformation (DA-T) algorithm, which, as its name implies, uses a transformation of the latent data in order to render the intractable full conditionals tractable. The latent data are transformed as follows: $Y_{pi} = Z_{pi} - (\theta_p - \delta_i)$. With all item-parameters equal to zero ($\delta_i = 0$), the full conditional of the latent data is:

$$\begin{aligned} f(y_{pi}|x_{pi}, \theta_p) &\propto f(y_{pi})(y_{pi} \leq \theta_p)^{x_{pi}}(y_{pi} > \theta_p)^{1-x_{pi}} \\ &= \frac{\exp(y_{pi})}{(1+\exp(y_{pi}))^2}(y_{pi} \leq \theta_p)^{x_{pi}}(y_{pi} > \theta_p)^{1-x_{pi}}, \end{aligned} \tag{2}$$

which is a truncated logistic distribution: truncated on the left at θ_p if $x_{pi} = 1$ and on the right at θ_p otherwise. The full conditional distribution of ability if we assume that θ has a prior logistic distribution is:

$$\begin{aligned} f(\theta_p|x_{pi}, y_{pi}) &\propto \left[\prod_i (\theta_p \geq y_{pi})^{x_{pi}}(\theta_p < y_{pi})^{1-x_{pi}}\right] f(\theta_p) \\ &\propto \left(\max_{i:x_{pi}=1}(y_{pi}) \leq \theta_p < \min_{i:x_{pi}=0}(y_{pi})\right) \frac{\exp(\theta_p)}{(1+\exp(\theta_p))^2}. \end{aligned} \tag{3}$$

This is a doubly truncated logistic distribution. The distribution is truncated on the left at $\max(y_{pi})$ for all i for which $x_{pi} = 1$ and on the right at $\min(y_{pi})$ for all i for which $x_{pi} = 0$. From equations (2) and (3) we can observe that the full conditionals of the transformed latent data are truncated distiributions of the transformed latent data and the full conditionals for the ability parameter θ is a truncated prior.

Note that a product of indicator functions corresponds to an intersection of intervals. In general,

$$\prod_i (l_i \leq x < u_i) = (\max_i(l_i) \leq x < \min_i(u_i)),$$

where either $l_i = \infty$ or $u_i = \infty$. Hence each term $(l_i \leq x < u_i)$ restricts the range of x_i to a half open interval extending to either plus or minus infinity. As illustrated, their product is the intersection of these intervals, ranging from $\max_i(l_i)$ to $\min_i(u_i)$. Thus, $\max_i(l_i)$ and $\min_i(u_i)$ are the truncation constants for the full conditional.

It can be proven that this intersection is never empty and $\max_i(y_{pi})$ and $\min_i(y_{pi})$ are valid truncation constants for the full conditional, see Maris and Bechger (2005).

4.2 The Data-Augmented Gibbs sampler for the Normal Ogive model

For the DA algorithm by Albert, the latent data z_{pi} follow a Normal distribution. It easily follows that the data augmented model is:

$$\begin{aligned} f(x_{pi}, z_{pi}|\theta_p) &\propto \prod_p \prod_i P(x_{pi}|z_{pi}) f(z_{pi}|\theta_p) \\ &= \prod_p \prod_i (z_{pi} \geq 0)^{x_{pi}} (z_{pi} < 0)^{1-x_{pi}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_{pi} - \theta_p)^2}{2} \right) \end{aligned}$$

where $(z_{pi} \geq 0)$ is an indicator function for the condition within parenthesis. Note that the distribution of the observed responses given the latent data is simple:

$$X_{pi} = \begin{cases} 1 & \text{if } z_{pi} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

For the Normal Ogive model, the distribution of the latent data and the prior of the ability parameter θ_p combine in a nice conjugate manner, hence we can easily derive the full conditional of ability parameter θ_p as follows:

$$\begin{aligned} f(\theta_p|x_{pi}, z_{pi}) &\propto \prod_i f(z_{pi}|\theta_p) f(\theta_p) \propto \exp\left(-\frac{(z_{pi} - \theta_p)^2}{2} \right) \exp\left(-\frac{\theta_p^2}{2} \right) \\ &= \exp\left(-\frac{\sum_i (z_{pi} - \theta_p)^2 + \theta_p^2}{2} \right) \\ &\Downarrow \text{ by completing the squares} \\ &= \frac{1}{\sqrt{\pi \frac{2}{n+1}}} \exp\left(-\frac{\left(\theta_p - \frac{\sum_i z_{pi}}{n+1}\right)^2}{\frac{2}{n+1}} \right) \end{aligned} \tag{4}$$

We can easily notice that the full conditional of ability is a normal distribution from which is easy to obtain a sample from.

Finally, the full conditional distribution of the latent data is:

$$f(z_{pi}|x, \theta) \propto (z_{pi} \geq 0)^{x_{pi}} (z_{pi} < 0)^{1-x_{pi}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_{pi} - \theta_p)^2}{2} \right) \tag{5}$$

which is a truncated normal distribution. Depending on the observed response we sample from a normal distribution truncated at zero from the left or from the right. Specifically, if $x_{pi} = 0$, the distribution of Z_{pi} is left truncated at zero, and if $x_{pi} = 0$ it is right truncated at zero, respectively. Note that DA sampler does not show the behaviour of the DA-T Gibbs sampler. Its autocorrelation is constant with respect to the number of observations. The example that is described in this thesis is a case where the asymptotic behavior depends on the parameterization of the Gibbs sampler.

4.3 Data Augmentation-Transformation Algorithm for the Normal Ogive model

For the DA sampler, the simplicity of the full conditional distribution of the ability parameter θ_p depends on the conjugacy of the distribution of the latent responses and the prior of ability. This approach, as has been previously described, breaks down for the Rasch model. As a remedy, Maris and Maris (2002) proposed the following transformation of the latent data: $Y_{pi} = Z_{pi} - (\theta_p - \delta_i)$. Hence, this remedy is called DA-T where T stands for *Transformation*. With all items parameters equal to zero ($\delta_i = 0$) the joint posterior density is the following:

$$f(x_{pi}, y_{pi}, \theta_p) = \prod_p \left(\prod_i (y_{pi} \leq \theta_p)^{x_{pi}} (y_{pi} > \theta_p)^{1-x_{pi}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_{pi}^2}{2} \right) \right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta_p^2}{2} \right) \quad (6)$$

Hence the full conditional of the latent data is:

$$\begin{aligned} f(y_{pi}|x_{pi}, \theta_p) &\propto f(y_{pi})(y_{pi} \leq \theta_p)^{x_{pi}} (y_{pi} > \theta_p)^{1-x_{pi}} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_{pi}^2}{2} \right) (y_{pi} \leq \theta_p)^{x_{pi}} (y_{pi} > \theta_p)^{1-x_{pi}}, \end{aligned} \quad (7)$$

which is a truncated normal distribution: truncated on the left at θ_p if $x_{pi} = 1$ and on the right at θ_p otherwise. The full conditional distribution of ability is:

$$\begin{aligned} f(\theta_p|x_{pi}, y_{pi}) &\propto \left[\prod_i (\theta_p \geq y_{pi})^{x_{pi}} (\theta_p < y_{pi})^{1-x_{pi}} \right] f(\theta_p) \\ &= \left(\max_{i:x_{pi}=1} (y_{pi}) \leq \theta_p < \min_{i:x_{pi}=0} (y_{pi}) \right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta_p^2}{2} \right) \end{aligned} \quad (8)$$

This is a doubly truncated normal distribution. The distribution is truncated on the left at $\max(y_{pi})$ for all i for which x_{pi}=1 and on the right at $\min(y_{pi})$ for all i for which $x_{pi} = 0$.

5 Asymptotic Behavior

The introduction of latent continuous variables renders the Gibbs sampler for IRT models tractable but it comes with a cost: the amount of latent data increases when the number of persons and/or items increases, raising the autocorrelation between successive draws.

In this Section, we are concerned to show how the step size of the sampler diminishes when the number of items increases. The interesting observation is that the step sizes correspond to the rate that the truncation limits of the full conditional of the ability parameter shrink. Our goal is to derive the distribution of the truncated area, which is the difference between the upper and the lower bound. Note that the rationale used to derive the asymptotic distribution of the DA-T samplers for both the Normal Ogive model and the Rasch model is the same, yet we choose to present them in two separate subsections for the sake of completeness.

5.1 Asymptotic behavior of the DA-T for the Normal Ogive model

For the Normal Ogive model, we look carefully at the at the full-conditional distribution of ability for the DA-T sampler (8). This is a doubly truncated normal distribution where the truncation limits correspond to the step sizes of the algorithm.

We wish to know how the step size diminishes when the number of items increases. To this aim, we consider a second transformation of variables: $U_{pi} = \Phi(Y_{pi})$ and $\Pi_p = \Phi(\Theta_p)$, where Φ denotes the cumulative normal distribution function, which gives full conditional distributions:

$$\begin{aligned} f(\pi_p|x,u,\pi^{(p)}) &\propto \left(\max_{i:x_{pi}=1}(\Phi^{-1}(u_{pi})) \leq \Phi^{-1}(\pi_p) < \min_{i:x_{pi}=0}(\Phi^{-1}(u_{pi})) \right) \\ &= \left(\max_{i:x_{pi}=1}(u_{pi}) \leq \pi_p < \min_{i:x_{pi}=0}(u_{pi}) \right) \end{aligned} \tag{9}$$

and $f(u_{pi}|x,u^{(p)},\pi) \propto (0 \leq u_{pi} < \pi_p)^{x_{pi}}(\pi_p < u_{pi} \leq 1)^{1-x_{pi}}$. It follows that:

1. For $i : x_{pi} = 0$
$$\frac{U_{pi} - \pi_p}{1 - \pi_p} \sim \mathrm{U}(0,1)$$
2. For $i : x_{pi} = 1$
$$\frac{U_{pi}}{\pi_p} \sim \mathrm{U}(0,1)$$

After we have obtained these full conditionals and by using the results explained in Appendix (see section "The Beta Distribution") and (15) and (16) it can be seen that:

$$\frac{\max_{i:x_{pi}=1}(U_{pi})}{\pi_p} \sim \text{Beta}(x_{p+}, 1)$$

$$\frac{\min_{i:x_{pi}=0}(U_{pi})}{1-\pi_p} \sim \text{Beta}(1, n - x_{p+})$$

It follows that:

$$\mathcal{E}[\min_{i:x_{pi}=0}(U_{pi}) - \max_{i:x_{pi}=1}(U_{pi})] = \pi_p + (1-\pi_p)\frac{1}{n - x_{p+} + 1} - \pi_p \frac{x_{p+}}{x_{p+} + 1} = o(1/n) \tag{10}$$

which means that for large n the sequence $\pi_p + (1-\pi_p)\frac{1}{n-x_{p+}+1} - \pi_p \frac{x_{p+}}{x_{p+}+1}$ is an order of magnitude smaller than $1/n$. Simply stated, this sequence tends to 0 faster than $1/n$. Thus we find that the maximum (expected) step size in an iteration of the Gibbs sampler is of order $1/n$, whereas the posterior standard deviation decreases with rate $1/\sqrt{n}$ (Stout, 2002), which implies that the Markov chain gets stickier as the sample size increases until it ultimately grinds to a halt.

5.2 Asymptotic behavior of DA-T for the Rasch model

In order to study the asymptotic behavior of the DA-T sampler for the Rasch model we need to make use of theorems. To this aim we make use of the following theorem:

Theorem 1. *Suppose that $X_1, \ldots, X_n$ are independent and identically distributed positive random variables from a distribution with probability density f which is continuous and positive at $x = 0$. Then $n \cdot \min_i(X_i)$ is asymptotically exponentially distributed with rate $\lambda = f(0)$.*

Note that the assumption about a continuous density implies that the distribution function is differentiable and in particular, its derivative from the right at 0, $F'(0)$, equals the density function at $(0, f(0))$. To this aim we use the survival function of n times the minimum.

Proof Sketch.

$S_n(a) = \Pr(n \cdot \min X_i \geq a) = \Pr(X_i \geq \frac{a}{n} \text{ for all i}) = (1 - F(\frac{a}{n}))^n$ □

The rationale behind this idea is that for $n \to \infty$, $F(\frac{a}{n})$ is approximately equal to $\frac{a}{n} f(0)$ and that $(1 - f(0)\frac{a}{n})^n \to \exp(-af(0))$ as $n \to \infty$. We also use the log survival function. This is $n \cdot \log(1 - F(\frac{a}{n}))$ and we need to show that it converges to $-af(0)$ when $n \to \infty$. There are numerous ways to prove

this but in this thesis we make use of the first order Taylor expansion.[2] Now, $n\log(1 - F(\frac{a}{n})) = n\log(1 - f(0))\frac{a}{n} + o(\frac{1}{n}) = n[-f(0)\frac{a}{n} + o(\frac{1}{n})] = -af(0)$ as $n \to \infty$.

Therefore, as trying to derive the distribution of the truncated area by using the minimum and the maximum, we realize that the two distributions are both truncations (to the right and to the left, respectively) of the same distribution; and the two sample sizes are asymptotically proportional to the probabilities to the right and to the left in the same distribution. Thus, the normalized densities(i.e., after the truncations) stand for the inverse proportions to the sample sample sizes. The two exponentials have asymptotically the same rate when we use the same normalization on both sides; the sum converges in distribution to a Gamma (shape=2)

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

where $\beta = 1$ and $\alpha = 2$.

The distribution of the scaled truncation constants is a Gamma$(2, 1)$ distribution. Walker (1969) and Chang and Stout (1993) show that the posterior distribution converges to a normal distribution with mean equal to the maximum likelihood (ML) estimator of θ, i.e., $\hat{\theta}_n$ and variance equal to $\sigma_n^2 = -L''(\hat{\theta_n})^{-1}$, i.e., the second derivative of the log-likelihood function evaluated at the ML estimator. In our case this boils down to $\sigma_n^2 = (npq)^{-1}$, where p is the probability of a correct response and q the probability of an incorrect response. It follows that the posterior standard deviation $\sigma_n \approx (npq)^{-\frac{1}{2}}$, which is of order $n^{\frac{1}{2}}$. Properties of the Gamma(α, β) show that on average:

$$\frac{\alpha}{\beta} = \frac{2}{npq},$$

where n is the number of items of the test, and p and q are the probabilities of a correct and incorrect response for a person with ability equal to θ, respectively and converges to 0 as $n \to \infty$. If we compare the rate of convergence for the average difference with that of the posterior standard

[2] A Taylor series is a series expansion of a function about a point. A one-dimensional Taylor series is an expansion of a real function $f(x)$ about a point $x = a$ is given by:

$$f(x) = f(\alpha) + f'(\alpha)(x - \alpha) + \frac{f''(\alpha)}{2!}(x - \alpha)^2 + \frac{f^{(3)}(\alpha)}{3!}(x - \alpha)^3 + ...$$

If $\alpha = 0$, the expansion is known as a Maclaurin series.

deviation, we are:

$$\begin{aligned}\lim_{n\to\infty} \frac{\frac{2}{npq}}{\frac{1}{\sqrt{n}}} &= \lim_{n\to\infty} \frac{2\sqrt{n}}{npq} \\ &= \lim_{n\to\infty} \frac{2}{\sqrt{n}pq} \\ &= 0.\end{aligned}$$

Thus, the posterior standard deviation shrinks slower than the truncated area under the Gamma approximation. As a result, autocorrelation in the Markov chain increases as the number of items in the test grows.

6 Autocorrelation

Looking at the step sizes is an intuitive but indirect way to look at the autocorrelation of the Markov chain. By definition the autocorrelation is:

$$\rho(\theta^{(t)}, \theta^{(t+1)}) = \frac{\mathrm{Cov}(\theta^{(t)}, \theta^{(t+1)})}{\sigma_{\theta^{(t)}} \sigma_{\theta^{(t+1)}}} \tag{11}$$

where $\theta^{(t)}$ and $\theta^{(t+1)}$ are the values of the ability parameter for time t and $t+1$, respectively. If the chain has converged it holds that

$$\mathrm{Cov}(E[\theta^{(k)}|Z, X], E[\theta^{(k+1)}|Z, X]) = \mathrm{Var}(E[\theta|Z, X]),$$

where X denotes the observed data, Z the latent data and θ an ability. Since, $\mathrm{Cov}(\theta^{(k)}, \theta^{(k+1)}|Z, X) = 0$, it follows from the *covariance decomposition formula* that the autocorrelation can be computed as:

$$\rho(\theta^{(t)}, \theta^{(t+1)}|X) = \frac{\mathrm{Var}(E[\theta|Z, X])}{\mathrm{Var}(\theta|X)} = 1 - \frac{E(\mathrm{Var}[\theta|Z, X])}{\mathrm{Var}(\theta|X)} \tag{12}$$

This expression known as the (Bayesian) fraction of missing information (FMI), was derived by Liu and shows that the autocorrelation depends on the ratio of the *augmented* posterior variance and the posterior variance.

As an intuition of this is that the higher the fraction of missing information, the more "sticky" the sample outputs from the data augmentation and vice versa. The fraction of missing information is also important for one to decide how many imputations should be provided. Also, in terms of the augmented-data Fisher information, the less we augment the faster the algorithm will be as measured by its theoretical rate of convergence. On the other hand, the less we augment the more difficult the implementation is expected to be.

7 Discussion

For our sampler, we have shown that the variance of the augmented posterior may decrease at a much faster rate than the posterior variance thus raising the autocorrelation of the chain. This result can be interpreted as: with increasing amount of data, the complete data information increases faster than the the observed data information. Note that in order to dermine whether a Gibbs sampler may suffer "asymptotic impotence" is the asymptotics of the augmented posterior. More precisely, whether or not the augmented posterior follows the usual central limit theorem asymptotics or something else: e.g., extreme value.

Hence, we can conclude that although data augmentation is a powerful tool, it is not guaranteed that it is efficient especially for large datasets. The fact that the constructed Markov chain may grind to a halt as the sample size increases provides a cautionary argument in the implementation of the algorithm. This thesis focused only in one-dimensional IRT models therefore provides evidence of caution when using data augmented schemes for models with one parameter. Maris and Bechger (2005) provided also an expository account of the DA-T Gibbs sampler for the two-parameter (2PL) model.

Finding an efficient augmentation scheme is however difficult since it needs to be worked out on a case-by-case basis. For example, while the "Slice" sampling[3] is a general strategy, it can be challenging to implement when $p(x|z)$ is not easy to sample from and can result in extremely slow algorithms when certain asymmetries arise in the target density. The DA-T sampler is also a slice sampler. Much recent work has been devoted to the development of general strategies for construction MCMC algorithms that are both fast and simple; for example, the work of Damien, Wakefield and Walker (1999).

[3]Slice sampling is a type of Markov chain Monte Carlo algorithm for pseudo-random number sampling, i.e., for drawing random samples from a distribution. The method is based on the observation that to sample a random variable one can sample uniformally from a region under the graph of its density function.

References

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

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

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.

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.

Gelfand, A. E., & Smith, A. F. M. (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.

Liu, J. (n.d.). *Fraction of missing information and convergence rate of data augmentation.* (Department of Statistics, Harvard University, Cambridge, MA 02138)

Maris, G., & Bechger, T. (2005). An introduction to the DA-T gibbs sampler for the two-parameter logistic (2pl) model and beyond. *Psicológica*, *26*, 327-352.

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

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.

R Core Team. (2013). 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.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, *67*, 485-518.

Swendsen R., W. J. (1987). Nonuniversal critical dynamics in monte carlo simulations. *Physics Review Letters*, *58*, 86-88.

Tanner, M. A., & Wong, W. H. (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.
Walker, A. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Society. Series B*, *31*, 80-88.

Appendices

A Background information on Beta Distribution and Order of Uniform Variables

A.1 The Beta Distribution

The beta distribution is a continuous distribution for random variables defined on the interval [0, 1] parametrized by two positive shape parameters, denoted by α and β. In Bayesian statistics it is the conjugate prior distribution for the binomial and geometric distribution. The probability density function is:

$$f(x) \propto x^{\alpha-1}(1-x)^{\beta-1} \tag{13}$$

and the expected value of a Beta distribution random variable is $E[x] = \frac{\alpha}{\alpha+\beta}$.

A.2 Order Statistics of Uniform variables

Assume $U_1, U_2, ..., U_n$ i.i.d uniformally distributed on (0,1). Hence, $F_U(x) = x$ and $f_U(x) = 1$. Let $U_{i:n}$ denote the i-th order statistic. In particular, $U_{1:n} = \min\{U_i\}$ and $U_{n:n} = \max\{U_i\}$. The distribution of the maximum is easily derived to be:

$$P(U_{n:n} \leq x) = \prod_i P(U_i \leq x) = [F_U(x)]^n \tag{14}$$

Similarly we see that $P(U_{1:n} \leq x) = 1 - [1 - F_U(x)]^n$. Differentiating we find that:

$$f_{U_{n:n}}(x) = n[F_U(x)]^{n-1} f_U(x) = nx^{n-1}$$

and

$$f_{U_{1:n}}(x) = n[1 - F_U(x)]^{n-1} f_U(x) = n(1-x)^{n-1}$$

Comparison to the density of the beta distribution shows that both the minimum and maximum of i.i.d uniform variables are distributed according to a beta distribution: That is

$$U_{1:n} \sim \text{Beta}(1, n) \tag{15}$$

$$U_{n:n} \sim \text{Beta}(n, 1) \tag{16}$$

B The Probability Transform

Suppose that a random variable X has a continuous distribution for which the cumulative distribution function is F. Then the random variable $Y = F(x)$ has a uniform distribution. This follows from the fact that:

$$F(X) \leq F(x) \Leftrightarrow (X \leq x) \cup (X > x, F(X) = f(x))$$

Since $P\Big(X > x, F(X) = f(x)\Big) = 0$ it follows that

$$P\Big(F(X) \leq F(x)\Big) = F(x) \tag{17}$$

Thus if $F(x) = p$, where p lies between 0 and 1, we have $\mathrm{P}(F(X) \leq p) = p$. This fact is often used in derivations and also implies a way to simulate data:

$$\mathrm{P}(F^{-1}(U) \leq x) = \mathrm{P}(U \leq F(x)) = F(x) \tag{18}$$

This method is called inverse probability sampling.

C R Code

```
k <- 20 # number of items
theta <- 0.25 # ability parameter 1
x <- 1*(rlogis(k)<=theta)

r.trunc <- function(l, h, n)
{ p <- runif(n)*(1/(1+exp(-h)) - 1/(1+exp(-l))) + 1/(1+exp(-l))
  return(log(p)-log(1-p)) }

niter<-1000000
z <- vector(length=k)
pv <- rep(-3, niter)
for(iter in 2:niter)
 {
 z <- x*r.trunc(-Inf, pv[iter-1], k) + (1-x)*r.trunc(pv[iter-1], Inf, k)
 pv[iter] <- r.trunc(max(z[x==1]), min(z[x==0]), 1)
 }
theta1 <- pv

k <- 20
theta <- 0.5 # ability parameter 2
x <- 1*(rlogis(k)<=theta)

r.trunc <- function(l, h, n)
{ p <- runif(n)*(1/(1+exp(-h)) - 1/(1+exp(-l))) + 1/(1+exp(-l))
  return(log(p)-log(1-p)) }

niter <- 1000000
z <- vector(length=k)
pv <- rep(-3, niter)
```

```
for(iter in 2:niter)
 {
  z <- x*r.trunc(-Inf, pv[iter-1], k) + (1-x)*r.trunc(pv[iter-1], Inf, k)
  pv[iter] <- r.trunc(max(z[x==1]), min(z[x==0]), 1)
 }
theta2 <- pv

library(MASS)
kde.dat=kde2d(theta1, theta2, n=50) # joint density

z_1 <- vector(length=k)
z_2 <- vector(length=k)
pv_1 <- rep(0.4, niter)
pv_2 <- rep(-1, niter)
contour(kde.dat,x\lim = c(-3, 2), y\lim = c(-3, 2), xlab="Ability 1",
ylab="Ability 2", drawlabels=FALSE)
for(iter in 2:50)
 {
  z_1 <- x*r.trunc(-Inf, pv_1[iter-1], k) + (1-x)*r.trunc(pv_1[iter-1], Inf, k)
  z_2 <- x*r.trunc(-Inf, pv_2[iter-1],k) + (1-x)*r.trunc(pv_2[iter-1], Inf, k)
  pv_1[iter] <- r.trunc(max(z_1[x==1]), min(z_1[x==0]), 1)
  pv_2[iter] <- r.trunc(max(z_2[x==1]), min(z_2[x==0]), 1)
  segments(pv_1[iter-1], pv_2[iter-1], pv_1[iter], pv_2[iter-1])
  segments(pv_1[iter], pv_2[iter-1], pv_1[iter], pv_2[iter])
 }
##############################################################################
k <- 100 # number of items
theta <- 0.25
x <- 1*(rlogis(k)<=theta)

r.trunc <- function(l, h, n)
 { p <- runif(n)*(1/(1+exp(-h)) - 1/(1+exp(-l))) + 1/(1+exp(-l))
   return(log(p)-log(1-p)) }

niter <- 1000000
z <- vector(length=k)
pv <- rep(-3, niter)
for(iter in 2:niter)
 {
 z <- x*r.trunc(-Inf, pv[iter-1], k) + (1-x)*r.trunc(pv[iter-1], Inf, k)
 pv[iter] <- r.trunc(max(z[x==1]) , min(z[x==0]),1)
 }
theta1 <- pv
```

```
k <- 100
theta <- 0.5
x <- 1*(rlogis(k)<=theta)

r.trunc <- function(l, h, n)
 { p <- runif(n)*(1/(1+exp(-h)) - 1/(1+exp(-l))) + 1/(1+exp(-l))
   return(log(p)-log(1-p)) }

niter <- 1000000
z <- vector(length=k)
pv <- rep(-3, niter)
for(iter in 2:niter)
 {
  z <- x*r.trunc(-Inf, pv[iter-1], k) + (1-x)*r.trunc(pv[iter-1], Inf, k)
  pv[iter] <- r.trunc(max(z[x==1]), min(z[x==0]), 1)
 }
theta2 <- pv


library(MASS)
kde.dat <- kde2d(theta1, theta2, n=50)

z_1 <- vector(length=k)
z_2 <- vector(length=k)
pv_1 <- rep(0.4, niter)
pv_2 <- rep(-1, niter)
contour(kde.dat, x\lim=c(-3, 3), y\lim=c(-3, 3), xlab="Ability 1",
ylab="Ability 2", drawlabels=FALSE)
for(iter in 2:50)
 {
  z_1 <- x*r.trunc(-Inf, pv_1[iter-1],k) + (1-x)*r.trunc(pv_1[iter-1], Inf, k)
  z_2 <- x*r.trunc(-Inf, pv_2[iter-1],k) + (1-x)*r.trunc(pv_2[iter-1], Inf, k)
  pv_1[iter] <- r.trunc(max(z_1[x==1]), min(z_1[x==0]), 1)
  pv_2[iter] <- r.trunc(max(z_2[x==1]), min(z_2[x==0]), 1)
  segments(pv_1[iter-1], pv_2[iter-1], pv_1[iter], pv_2[iter-1])
  segments(pv_1[iter], pv_2[iter-1], pv_1[iter], pv_2[iter])
  }
```