Evaluation of Relations between Scales in an IRT Framework

Khurrem Jehangir
Enschede, 2005

Supervisors
Prof Dr. C.A.W. Glas
Dr. H.J.Vos

Table of Contents

Introduction

In an ever changing world, psychological testing remains one of the flagships of applied psychology. Over the course of the past decade or two there have been currents of change in the domain of psychological testing. New testing techniques are emerging in response to contemporary needs in psychological testing. These techniques are also based on new principles underlying test development. One promising technique which has been gaining momentum since the last decade or two is called IRT or item response theory. It is the basis or mainstay for the creation of many a modern test. It is gradually phasing out the classical test theory (CTT) from the testing domain because of its more theoretically justifiable principles and greater potential to solve practical measurement problems.

IRT, also known as latent trait theory is model based measurement in which trait level estimates depend on both person responses and on the properties of items that were administered. The rules of measurement in IRT afford greater robustness, flexibility, efficiency and reliability in trait measurement than the classical test theory framework which was in use for most of the 20th century. The underlying principal used in IRT models for testing is that person and item parameters can be fully separated and this is brought to bear on measuring examinee traits and test characteristics with greater precision and flexibility.

IRT now underlies several major tests. Apart from educational testing to measure examinee ability, IRT has also been applied to personality trait measurements, as well as to attitude measurements and behavioral ratings. Computerized adaptive testing in particular relies on IRT. In computerized adaptive testing examinees receive items that are optimally selected to measure their potential. Different examinees may receive no common items. IRT principles are involved in both selecting the most appropriate items for an examinee and equating across different subsets of items.

Many diverse IRT models are now available for application to a wide range of psychological areas. Although early IRT models emphasized dichotomous item formats extensions to other formats has enabled applications in many areas; that is, IRT models have been developed for rating scales, partial credit scoring and multiple category scoring.

This report studies the application of a new procedure for measuring across-scales relationship in a multi-dimensional IRT test. A multi-dimensional test is one in which a test is divided into sub-sets and latent variables are measured separately for each scale. The latent variables are assumed to correlate and the new procedure called 'limited information maximum likelihood estimation' proposed by Rubin and Thomas (2001) is used to estimate this correlation. The authors did not indicate that they had empirically tested this new procedure and in this report empirical tests are carried out to observe the effects of different models and values of item and person parameters.

Chapter 1 of this report makes a comparison between the new IRT framework and the old Classical Test Theory (CTT) framework. Apart from highlighting the conceptual differences between IRT and CTT it also lists the array of advantages that are afforded by IRT to examiners. After stressing the importance of IRT as a testing framework, an explanation of the different IRT models for both dichotomous and polytomous scoring is given.

Chapter 2 of this report is about the methodology that is tested in this report. It begins with an explanation of important concepts that are used for estimation of IRT model parameters like maximum likelihood scoring for single variable models. It also presents other estimation methods that are used in this report like marginal maximum likelihood (MML) and the EM algorithm. Then the new methodology used for estimating across scales relationship is described.

Chapter 3 is the concluding chapter; the results of the simulations carried out are presented followed by a discussion of the results and the feasibility of the methodology used for different IRT model configurations.

CHAPTER 1

Introduction to IRT

Introduction to IRT

In this chapter I give an introduction to the IRT framework. I begin by making a comparison between the old testing framework called the Classical Test Theory (CTT) and IRT. I discuss the advantages of IRT over CTT and how it is better equipped to meet the requirements of modern testing. After that I briefly trace the history of IRT evolution and present the basic IRT model in use today for dichotomous scoring called the Rasch model and its extensions (i.e. 2PL). After that I present three IRT models for polytomous scoring and describe how they can be evolved from the basic Rasch model.

IRT and the Classical Test Theory

The two testing theories that are in use today are the Classical Test Theory (CTT) and the IRT. CTT predates IRT and has been the mainstay of psychological testing for most part of the 20th century. However since the last decade or two IRT has become the dominant currency of testing, while CTT is become less popular. IRT is based on fundamentally different principles than CTT. IRT is a model based measurement that controls various confounding factors in score comparisons by a more complete parameterization of the measurement situation.

IRT differs substantially from CTT as a model based system of measurement. Unlike IRT in CTT item properties and person properties are confounded in the basic model which has many implications. Firstly, practical testing problems such as equating different test forms have been solved by using IRT. Furthermore, thanks to justifiable measurement scale properties of IRT inferential statistics about group differences as well as test score comparisons within or between persons have become possible.

One powerful feature of IRT is that IRT trait levels have meaning for any set of calibrated items (where the same model holds) because IRT models include invariant item properties. By contrast, in CTT the true score has meaning only for a fixed population and item set; by not including item properties in the model, a true score can apply only to a particular set of items or their equivalent. Thus the properties of items are not explicitly linked to behavior in CTT but in IRT they are. The outcome of this is that for example the relative impact of difficult items on trait level estimates and item responses can be known from an IRT model.

Another weak link in CTT is that a persons true and error score may not be decomposed from a single test administration. The standard error of measurement in CTT applies to all scores in a particular population i.e. it is the same for all examinees in the population whereas in IRT based testing the standard error differs across scores or response patterns and gives a more accurate estimate of the conditional standard error for any particular examinee.

Another outcome of the IRT based testing is that tests do not have to be as lengthy as they were in CTT. A short test that is based on IRT can be more accurate than a longer test based on CTT. This is made possible by using adaptive testing in which the examinees are administered items which match their estimated abilities using the IRT framework. Rather than testing examinees with items that are spread out over the entire ability spectrum, the test items can be focused around the estimated ability estimates of the examinees thus resulting in a more accurate estimate coupled with a shorter test.

IRT is thus gradually phasing out CTT because of the main reasons which have been explained above. However IRT based testing is truly effective when it is done in the form of adaptive testing using computers. Because in IRT examinees are not supposed to receive the same questions but rather those questions that give maximum information about their estimated trait levels at that point in time. Modern computers possess the computing power to perform the necessary calculations to estimate trait levels continuously and select the next appropriate item to administer from the item pool. Computer adaptive testing is not yet pervasive but it is gradually phasing out paper and pencil based tests which are based on the CTT. IRT in its capacity as the framework on which computer adaptive testing relies will be the currency of future testing.

IRT Models

Item response theory (IRT) contains a large family of models. IRT development may be traced back to Lawley(1943) and Lord(1952). An important theoretical breakthrough was made by Georg Rasch(1960), a Danish mathematician who developed a family of IRT models that were applied to develop measures of reading. Rasch was particularly interested in the scientific properties of measurement models and he separated person and item parameters fully in his models. His student Anderson consequently elaborated estimation methods for the person and item parameters in Rasch's model. Rasch developed many IRT models but his most famous model which is known by his name is the building block for many more complex or advanced models used in IRT today. Though there are many IRT models in use today I will only discuss the five IRT models that I use in this report. Two of these models are models for dichotomously scored items (in which there are only two possible outcomes) and three models are for modelling polytomously scored items (in which there are more than two possible outcomes).

Dichotomous IRT Models

The Rasch Model or the 1-PL Model.

The simplest IRT model which belongs to the exponential family that is in use today is the Rasch model which is also known as the one parameter logistic model (1PL). For the simple Rasch model, the dependent variable is the response to a dichotomously scored item. The independent variables are the person's trait score and the item difficulty level. Linking the independent variable to the dependent variable requires a non linear function which is the logistic function. The prediction provided by this logistic function is as follows:

$$P(X_{si}=1) = \frac{\exp(\theta_s - \beta_i)}{1+\exp(\theta_s - \beta_i)}$$

where X_{si} is the response of person s on item i, θ_s the trait level of person s and β_i the item difficulty of item i . P (X_{si}=1) is the probability of a correct response on item i for person s.

This is known as the Rasch model or the 1 parameter logistic (1PL) measurement model, due to the inclusion of only one item parameter (i.e. difficulty) to represent item differences.

The 2-PL Model.

The 2-PL model is a more 'complete' or complex IRT model than the Rasch model. This model includes two parameters to represent item properties. One is the item difficulty which was also there in the Rasch model and the other new parameter is called the discrimination parameter denoted by the symbol alpha. The discrimination parameter is also known as the slope parameter. The Rasch model can be extended to the two-parameter logistic model (2-PL) by allowing different discrimination parameters for each dichotomous item. Both item difficulty and item discrimination are included in the exponential form of the logistic model as follows:

$$P(X_{si}=1) = \frac{\exp\alpha_i(\theta_s - \beta_i)}{1+\exp\alpha_i(\theta_s - \beta_i)}$$

where X_{si} is the response of person s on item i, θ_s the trait level of person s and β_i the item difficulty of item i . α_i is the discrimination parameter. P (X_{si}=1) is the probability of a correct response on item i for person s.

The discrimination or slope parameters indicate the steepness of the response function .An outcome of this property is that it is possible to observe how good the response function discriminates θ-values in the neighborhood of β_{si} .When the discrimination parameters are high, they give more information about the trait level of a person if he/she is administered a question whose difficulty level is equal to or close to the trait level of an examinee. Tests can be shorter if the discrimination parameter is also included in the model because its inclusion means that questions if carefully selected can give more information about the trait level of an examinee which can thus be more quickly identified.

Polytomous IRT Models

In this section I will discuss three models for polytomous item scoring. The models that were discussed earlier were for dichotomous scoring, i.e. there were only two possible response categories, either a correct score or an incorrect score. In polytomous scoring there are more than two possible response categories, i.e. there are more possibilities than either scoring correct or wrong. In other words a person may score an item partially correct. To model response behaviour on a test where there are more than 2 possible response categories models known as polytomous models have to be used. These models can be applied to any situation in which performances on an item or an assessment criterion are recorded in two or more ordered categories (e.g., rating scales) and there is an intention to combine results across items/criteria to obtain measures on some underlying variable. There are many models for polytomous scoring that have been proposed. In this section I discuss three of the more commonly used models which I will later use in the simulation studies.

The Generalized Partial Credit Model (GPCM)

The PCM or the partial credit model is a uni-dimensional model for the analysis of responses recorded in two or more ordered categories (Masters, 1982). The Partial Credit Model is an application of Rasch's model for dichotomous scoring to polytomously scored items (Masters, 1982). When an item provides only two scores 0 and 1 (i.e., wrong and correct), the probability of scoring 1 rather than 0 is expected to increase with the ability being measured. In Rasch's model for dichotomous scoring, this expectation is modelled as:

$$\frac{P_{ij1}}{P_{ij0} + P_{ij1}} = \frac{\exp(\theta_j - \delta_i)}{1 + \exp(\theta_j - \delta_i)},$$

where $P_{\mathrm{ij}1}$ is the probability of person j scoring 1 on item $i.P_{ij0}$ is the probability of person j scoring 0 on item i, and θ_{j} is the ability of person j. Furthermore, δ_{i} denotes the difficulty of item i defined as the location on the ability scale at which a score of 1 on item i is as likely as a score of 0 (i.e., $P_{ij0} = P_{ij1} = 0.5$). The larger δ_{i}, the smaller the probability of scoring 1 rather than 0 on item i.

The model is written here as a conditional probability to emphasize that it is a model for the probability of person j scoring 1 *rather than* 0. The above formula for the probability of scoring 1 rather than 0 also expresses the probability of person j scoring 1 on item i (i.e., the item response function P_{ij1}), since P_{ij0} and P_{ij1} must obviously sum up to 1.

When an item provides more than two responses categories (e.g., three ordinal categories scores 0, 1 and 2), a score of 1 is not expected to be increasingly likely with increasing ability. This is because beyond some points on the ability scale, a score 1 should become less likely as a score 2 becomes a more likely result. Nevertheless, it follows from the intended order 0 < 1 < 2 …. < m_i of a set of response categories that the *conditional* probability of scoring x rather than x-1 on an item i (i.e., given only two possible scores) should increase monotonically throughout the ability range. In the PCM, this expectation is modelled using Rasch's model for dichotomous scoring:

$$\frac{P_{ijx}}{P_{ijx-1} + P_{ijx}} = \frac{\exp(\theta_j - \delta_{ix})}{1 + \exp(\theta_j - \delta_i x)}, \qquad x = 1, 2, \dots, m_i$$

where P_{ijx} is the probability of person j scoring x on item i, P_{ijx-1} is the probability of person j scoring x-1 on item i. θ_j is the ability of person j, and δ_{ix} is an item parameter (denoted also as difficulty parameter or sometimes as category or threshold parameter) governing the probability of scoring x rather than x-1 on item i. However, in the polytomous case, unlike in the dichotomous case, it does not hold any longer that P_{ijx-1} and P_{ijx} sum up to 1 since there are now more than two probabilities involved.

Since it must hold that P_{ij0} + P_{ij1} + … + P_{ijx} + … + P_{ijm_i} = 1, as shown by Masters (1982), it can readily be derived that the item response functions (mostly denoted as category response functions) for P_{ijx} (x = 1, 2, …, m_i) and P_{ij0} in the PCM can be formulated as follows:

$$P_{ijx} = \frac{\exp(x\theta_j - \sum_{k=1}^{x} \delta_{ik})}{1 + \sum_{h=1}^{m_i} \exp(h\theta_j - \sum_{k=1}^{h} \delta_{ik})},$$

$$P_{ij0} = \frac{1}{1 + \sum_{h=1}^{m_i} \exp(h\theta_j - \sum_{k=1}^{h} \delta_{ik})}.$$

Defining for notational convenience $\sum_{k=0}^{0} (\theta_j - \delta_{ik}) \equiv 0$, the above two expressions above for the response category functions can be formulated more compact as follows:

$$P_{ijx} = \frac{\exp \sum_{k=0}^{x} (\theta_j - \delta_{ik})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h} (\theta_j - \delta_{ik})} \qquad x = 0, 1, \ldots, m_i.$$

The PCM simplifies to the Rasch model if m_i = 1, that is, for an item with only two response categories. In other words, Rasch's model for dichotomous scoring can be considered as a special case of the PCM or in other words the PCM belongs to the Rasch family of models.

It can be verified from the above formulas that δ_{ik} indicates the location on the θ-scale where the probabilities of responding to categories k-1 and k (k = 1,…,m_i) on item i are equal. Furthermore, it holds that the larger δ_{ik}, the smaller the probability of scoring k rather than k-1 on item i. It is also evident that the category response function for the lowest ordered category (i.e., category 0) is decreasing in θ, whereas the one for the highest ordered category (i.e., category m_i) is increasing in θ. Since the sum of the category response functions must sum up to 1 for each value of θ, it follows that all other response category functions (i.e., categories 1, 2,…, $m_{i\text{-}1}$) must necessarily first be increasing in θ and next be decreasing in θ.

In order to allow for different slopes for the category response functions the PCM can be modified to the *Generalized Partial Credit Model* (GPCM) by incorporating the discrimination parameter in the model for the PCM. The GPCM which was originally developed by Muraki in 1992 and can be formulated in compact form as follows:

$$P_{ijx} = \frac{\exp \sum_{k=0}^{x} a_i(\theta_j - \delta_{ik})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^{h} a_i(\theta_j - \delta_{ik})} \qquad x = 0, 1, \ldots, m_i,$$

where a_i denotes the discrimination parameter for item i. The GPCM will boil down to the PCM when all discrimination parameters are equal. The GPCM is a more flexible model than the PCM implying that a better fit to the data can be achieved, but at the expense of statistical elegance (i.e. no sufficient statistics).

The Sequential Model

The Sequential model is an alternative to the GPCM. Verhelst, Glas and de Vries(1997) develop the model by assuming that a polytomous item consists of a sequence of item steps. It views polytomous data as a special case of data resulting from a multistage testing design with dichotomous items where every test consists of one dichotomous item only. The choice of a follow up test is a function of the responses on previous items. Every step corresponds with a conceptually dichotomous Rasch item. The student is only administered the next conceptual Rasch item if a correct response was given to the previous one. It is assumed that if a conceptual item is administered, the Rasch model holds, so the probability of taking a step is given by:

$$p(Y_{ikm} = 1 \mid d_{ikm}) = 1, \theta_i, b_{km}) = \frac{\exp(\theta_i - b_{km})}{1 + \exp(\theta_i - b_{km})}$$

Where d_{ikm} is a design variable for dichotomous items given by:

d_{ikm} = {1, if a response of person i to item k is available}

d_{ikm} = { 0 , if a response of person i to item k is not available}

b_{km} is the difficulty parameter of step m within item k. If we denote the number of steps within item k by r_{ik} then:

$$\mathrm{r}_\mathrm{k} = \sum_{m=1}^{M_k} d_{km} y_{km}$$

Using these formulas in the table given below are all possible responses for an item with M_k=3 and the associated probabilities $\mathrm{P}(y_k \mid \theta, b_k)$

Table 1.0

y_k	r_k	$\mathbf{P}(y_k \mid \theta, b_k)$
0,c,c	**0**	$\frac{1}{1 + \exp(\theta_i - b_{k1})}$
1,0,c	**1**	$\frac{\exp(\theta_i - b_{k2})}{[1 + \exp(\theta_i - b_{k1})][1 + \exp(\theta_i - b_{k2})}$

1,1,0	**2**	$\frac{\exp(\theta_i - b_{k1})\exp(\theta_i - b_{k2})}{[1+\exp(\theta_i - b_{k1})][1+\exp(\theta_i - b_{k2})][1+\exp(\theta_i - b_{k3})]}$
1,1,1	3	$\frac{\exp(\theta_i - b_{k1})\exp(\theta_i - b_{k2})\exp(\theta_i - b_{k2})}{[1+\exp(\theta_i - b_{k1})][1+\exp(\theta_i - b_{k2})][1+\exp(\theta_i - b_{k3})}$

From inspection of the above table it can be verified that in general:

$$P(y_k \mid \theta, b_k) = \frac{\exp\left[r_k\theta - \sum_{m=1}^{M_k} b_{km} \right]}{\prod_{h=1}^{\min(M_k, r_k+1)} [1+\exp(\theta - b_{km})}$$

Where min(M_k, r_k +1)stands for the minimum of M_k and r_k +1.

The Graded Response Model

In 1969, the general graded response model was proposed by Samejima. It is different from adjacent category models or Continuous-ratio models .In Adjacent category models the definition of the probability that the score say R_k is equal to m conditional on the event that it is either *m* or *m-1* and is given by:

$$P(R_k = m \mid R_k = m, or, R_k = m-1) = \psi(a_k(\theta - b_{km}))$$

Where ψ is a logistic function. Above it was shown that this assumption leads to the GPCM.

Continuous ratio models on the other hand are based on the definition of probability of scoring equal to or higher than m given that the score is at least m-1:

$P(R_k \geq m | R_k \geq m\text{-}1) = \psi(a_k(\theta - b_{km}))$

In the Graded response model however the probability is defined by:

$P(R_k \geq m) = \psi(ak(\theta - b_{km}))$

It follows that the probability of scoring in a response category m is given by

$P(R_k = m) = P(Y_{ikm} = 1 | \theta, b_k) = \psi(a_k(\theta - b_{km})) - \psi(a_k(\theta - b_{k(m+1)}))$

for m=1,……M_{k-1}.Since the probability of obtaining a score M_{k+1} is zero and since everyone can at least obtain a score 0, $P(R_k \geq M_k + 1) = 0$ and $P(R_k = 0) = 1$.Thus follows that

$P(R_k = 0) = P(Y_{ik0} = 1 | \theta, b_k) = 1 - \psi(a_k(\theta - b_{k1}))$

and

$P(R_k = M_k) = P(Y_{ikMk} = 1 | \theta, b_k) = \psi(a_k(\theta - b_{kMk}))$

For this model to work it must hold that $\psi(a_k(\theta - b_{km})) > \psi(a_k(\theta - b_{k(m+1)}))$, which implies that $b1 < b2 <$,…….,$< b_{Mk}$.. Furthermore the discrimination parameter a_k must be the same for all steps.

CHAPTER 2

Evaluation of relations between scales

Evaluation of relations between scales in an IRT framework

In this chapter I discuss the methodology that is tested in this report. It begins with an explanation of important concepts that are used for estimation of IRT model parameters like maximum likelihood scoring for single variable models. It also presents the marginal maximum likelihood estimation method (MML) that is used when there is more than one variable in the model. Then the EM or the estimation-maximization algorithm is presented as it is used in the methodology studied in this report. The new methodology used for estimating across scale relation is described next. It is called 'limited information maximum likelihood' estimation.

Maximum Likelihood estimation of trait levels in IRT

The relationship between item responses and trait level is fundamentally different in IRT and CTT. Under CTT, trait levels are scored by combining responses across items. Typically, responses are summed into a total score and then converted into a standard score. However in IRT, determining the person's trait level is not a question of how to add up the item responses.

In a sense the IRT process of estimating trait levels is analogous to the clinical inferences process. In models of the clinical inference process, a potential diagnosis or inference is evaluated for plausibility. That is, given the observed behaviors how plausible is a certain diagnosis. Thus, the behaviors (and test responses) are symptoms of a latent variable whose value must be inferred. Given the limited context in which the persons behavior can be observed by the clinical and knowledge of how behaviors are influenced by a latent syndrome, what diagnosis is most likely to explain the presenting behaviors? The IRT process is akin to clinical inference; given the properties of the items and knowledge of how item properties influence behavior (i.e. and IRT model), what trait level is most likely to explain the persons responses.

Supposing a person received very hard items on a test and answered nearly all of them correctly; this response pattern is not very likely if a person has a low trait level. The likelihood that a person with moderate trait level could have answered those questions correctly is somewhat higher, but the likelihood of the response pattern is even higher for a person who has a high trait level.

Finding the IRT trait level for a response pattern requires a search process rather than a scoring procedure. That is, the trait level that yields the highest likelihood for the responses is sought. In some cases, collateral information may be incorporated into the estimation procedure (e.g. trait distribution in the relevant procedure) so that more information than the response pattern is available. For instance when estimating the trait level distribution in a large class of examinees it is often assumed that the trait levels are normally distributed (i.e. pattern of distribution is like a normal distribution which has a bell shape). In IRT, trait levels are estimated in a model for a person's responses, controlling for the characteristics of the items. Typically, trait levels are estimated by the maximum likelihood method; specifically the estimated trait level for a person maximizes the likelihood of his or her response pattern given the item properties. Thus to find the appropriate trait level ,one must(a) represent the likelihoods of a response pattern under various trait levels and (b) conduct a search process that yields the trait level that gives the highest likelihood.

Maximum Likelihood estimation of trait level from response patterns

To find the most likely trait score, first the likelihood of the person's response pattern must be expressed in the model that contains the properties of the items that were administered. Once so expressed, the likelihood of the person's response pattern may be computed for any hypothetical trait level. Then the likelihoods may be plotted by the trait level so that the trait level with the highest likelihood can be observed.

Instead of plotting a graph of likelihoods (probabilities) of a response pattern against different values of the ability parameter to see which value of the ability parameter yields the highest likelihood (probability), one can simply find the maximum of function of the response pattern with respect to the trait level parameter in the function. The first step in calculating maximum likelihood of a response pattern is to translate the item response pattern into the probability of that item response pattern occurring under the given model. All item parameter values are considered known in the model with the exception of the unknown parameter 'trait level' which is yet to be estimated. The next step is to multiply all these probabilities for every item in the test. The resultant probability function is the probability of getting that response pattern over the whole test. This probability function contains one unknown parameter i.e. the trait level parameter. The next step is to maximize this probability function for the ability parameter so as to find out the value of the ability parameter for which the function gives a maximum value. (It can be shown that the probability function is single peaked so it returns only a single maximum value.).However finding the maximum of the likelihood function in this form is tedious from a mathematical point of view. So in practice the log likelihood of this function is calculated and maximized; because the log of the likelihood function also maximizes at the same trait level as the likelihood function in the original product form and is easier to calculate. The only exception to calculating maximum likelihood estimates are response patterns that are all correct or all wrong. These patterns will yield the value of 'infinite' ability for an all correct response pattern and a value of 'negative infinity' for an all wrong response pattern.

Marginal Maximum Likelihood estimation

The maximum likelihood estimation procedure described above is computed for a likelihood function of a single variable. However sometimes there is more than one variable in the likelihood function, as will be the case in this report. In such a case, in order to maximize the likelihood function of two or more variables, marginal likelihoods with respect to a subset of the variables can be defined. Let *a* denote the subset of variables marginalized (i.e., integrated). Let *b* denote the other variables. Let *x* denote observed data. Given the likelihood function *p*(*x*|*a*, *b*), the marginal likelihood of *b* is

$$\mathbf{P(x \mid b)} = \int p(x \mid a,b)\, p(a \mid b)\, da \,\mathbf{,}$$

where *p*(*a*| *b*) is the distribution of *a* conditional on *b*. In practice, *b* can be the item parameters or *b* can be the covariance matrix between ability dimensions (Bock & Aitkin, 1981)

The EM algorithm

The EM or the expectation-maximization algorithm is an algorithm for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables or missing data. EM alternates between performing an expectation (E) step, which computes the expected value of the latent variables, or missing data, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters given the data and setting the latent variables or missing data to their expectation(in an exponential family model) or to their distribution (in other models).

The EM algorithm can be viewed as an iterative method for finding the mode of the marginal posterior density p($\theta \mid y$) and is useful for many common models for which it is hard to maximize p($\theta \mid y$) directly but easy to work with p($\gamma \mid \theta, y$) and p($\theta \mid y, \gamma$) .If one thinks of θ as the parameters in the problem and γ as the missing data the EM algorithm formalizes the following idea, handling missing data starting with a guess of the parameters. (1) replace missing values by their expectations (or distributions) given the guessed parameters, (2) estimate parameters assuming the missing data are given by their estimated values (or distributions), (3) re-estimate the missing values assuming the new parameter estimates are correct, (4) re-estimate parameters and so forth iterating until convergence.

Measurement of between scales covariance

In the previous section I described how the IRT theory can be used to determine the ability level of examinees on a test. Usually, in IRT models it is assumed that there is one (dominant) latent variable θ that explains test performance. However, it may be a priori clear that multiple latent variables are involved. A test with items related to more than one latent variable is often labeled a within-item-multidimensional test (Adams, et al., 1997) and a multi-dimensional model is required for such tests. Adams et al. (1997) also write about another class of multidimensional IRT models, between-item-multidimensional models, in which one test can be divided into subtests or scales where the responses to the items of each scale can be described by a uni-dimensional IRT model. The latent variables, measured separately for each scale are assumed to correlate. It can be of interest to know the relationship between the different cognitive sub scale abilities. For instance it may be interesting to know the relationship between a persons I.Q. and math ability or I.Q. and language ability or math and language ability. The aim of this report is to test a new methodology for estimating the between scale relationships for a multi-dimensional test using a number of one-dimensional models. A simulation is carried out in which a group of students is tested on a test with three dimensions or areas of ability, like math I.Q. and language. A matrix of 3 dimensional abilities is generated using a model in which we incorporate a known covariance matrix for configuring the between scales relationship. We then test the new method to find the between scales covariance from the available data using the new method and compare it with the known covariance used to model the data set.

One method for calculating the covariance between scales is to use the so called full information maximum likelihood estimation in which the item parameters and the covariance matrix are concurrently estimated using marginal maximum likelihood (MML) and the EM-algorithm. However this is a tedious process and is infeasible for high-dimensional theta spaces because the computation of the integrals is cumbersome. It requires the calculation of a multiple integral (over the number of scales) in the probability model for every element of the covariance matrix. In this study I use limited information maximum-likelihood estimation in which the model is estimated in two steps: first the item and person parameters are estimated per dimension and secondly the covariance matrix is estimated given the estimates obtained in the first step. If the limited information maximum likelihood estimation works well it could then be considered a viable alternative for cases when the full information maximum likelihood estimates become too complex to compute.

The limited information maximum likelihood for calculating between scales covariance

The limited information maximum likelihood method used here is based upon an application of the EM algorithm presented by Rubin and Thomas (2001). Rubin and Thomas (2001) discuss a two-stage procedure where the first stage consists of calibrating the uni-dimensional subscales using a uni-dimensional IRT model such as the GPCM and the second stage consists of estimating the covariance-matrix between the latent variables using a combination of parameter expansion and the EM-algorithm

The method calculates the covariance between scales by using observed data. These variance estimates include variance due to measurement error. The relationship between the missing data and the estimated abilities is given by the equation 2.0.

Equation 2.0

$$y_i = \boldsymbol{K}\theta_i + \varepsilon_i$$

In this equation the parameter y_i represents the estimates abilities of the students. The parameter θ_i represents the actual abilities of the students and the parameter K is a matrix of the regression coefficients. The term ε_{i} represents the measurement error.

Equation 2.0 can be used in the EM algorithm as described in the following equations 2.1. and 2.2.

Equation 2.1

$$\theta^{t} = K^{-1} \; y^{t}$$

Equation 2.2

$$K_j = \boldsymbol{inv}\,(\theta\, W_j\, \theta^t)\, *\, (\theta^t\, W_j\, Y_j)$$

where $W_j = \mathrm{diag}(\tau_{1j}, \ldots\ldots, \tau_{Nj})$

Note that K_j is both the maximum likelihood estimate and the least squares estimate of K

In equation 2.1 the parameter y represents the abilities of students estimated earlier in step 1 whereas the parameter θ represents the expectation of θ given the model in equation 2.0 and are treated as the missing data.

For the EM algorithm to begin an initial value of the matrix K is guessed. In our case it is the identity matrix. The E step then estimates a value for the missing data θ given the guessed parameter K in equation 2.1. This is followed by the M-step (given in equation 2.2.) in which the parameter K is estimated assuming the missing data θ are given by their estimated values in the E-step. This ‘updated’ estimate of the parameter K which is the matrix of regression coefficients is then reinserted into the E-step to get new values for the missing data θ in the E-step which is then reinserted into the M-step for calculating an improved estimate of K and so forth till convergence is achieved. It can be shown mathematically that with every iteration the values of the estimates of estimates of K and $\sum$ are nearer to maximum likelihood estimates of these parameters.

Equation 2.3

$$\sum = 1/\mathrm{N} \sum_{i=1}^{N} \theta_i\, \theta_i^{t}$$

Once convergence is achieved, the final value of the covariance matrix $\sum$ is calculated using equation 2.3. On right hand side of the equation every column of the transpose of the estimated matrix of missing data is multiplied by its transpose. This is done over all persons and the average of the result will yield the final covariance matrix.

Simulation setup

The objective of this simulations carried out in this report is to study the viability of a limited information EM algorithm technique proposed by Rubin and Thomas(2001) for finding the covariance between scales. The scales in the empirical example used in this report correspond to the multidimensional ability sub sets. The test is split in a number of sub sets and every subset relates to a specific ability parameterθ_t. The relationship or covariance between the ability estimates for the different sub-sets is represented in the form of a covariance matrix.

The simulation study begins with randomly drawing values of theta for different scales from a multivariate normal distribution with a known covariance matrix. Thus we know from the start the covariance used to model the multidimensional ability matrix. The value of this known covariance matrix is used later for comparison with the covariance matrices resulting from the estimation procedure of Rubin & Thomas (2001).

The values of 'abilities' obtained so far represent the real abilities of the students. The next step is to convert these real ability values into estimates of these real abilities along with estimates of the measurement error. This is done in two steps. First, by generating a response pattern for all examinees over all items in each sub scale of the test using the actual sub scale abilities under the respective response model. Secondly, by estimating the sub-scale abilities from the response patterns using 'maximum likelihood' which will then yield the estimates of the sub scale abilities and the associated measurement errors. The estimates of the abilities obtained using maximum likelihood will serve as the input matrix Y (in equation 2.1) of the EM algorithm and likewise the estimates of the measurement error will serve as the diagonal elements of the matrix W (in equation 2.2).

We then calculate the covariance matrix using the EM algorithm and compare the values obtained with the original known covariance matrix for various input parameter combinations of the five different response models. We can thus observe the performance of the limited information EM algorithm in calculating the covariance between scales for the different situations.

.

CHAPTER 3

Results & Conclusions

CHAPTER 3

In this concluding chapter the results of the simulations carried out are presented. The Chapter is divided into two sections, one for dichotomous testing and one for polytomous models. Results are tabulated for both sections followed by a discussion of those results. The chapter ends by concluding about the feasibility of the 'limited information maximum likelihood estimation' for various dichotomous and polytomous models.

Results of the Simulation Study

The simulation setup described in the previous section was implemented for the five models described earlier. In the section below we can observe the results of the simulation exercises for the five different models. The tables in this section shows the between the simulated covariance matrix and the original covariance matrix for various input configurations. This is followed by a discussion of these results.

The test parameters that are altered for the five different models (as shown in the tables) include the sample size denoted by M , test length denoted by K and the value of the non-diagonal elements of the original covariance matrix denoted by $\sum$. The results of the simulation are represented in the form of the average of differences between the diagonal and off-diagonal elements of the original covariance matrix and the simulated covariance matrices. The simulated values of the covariance matrices represent an average of those values over 100 replications.

The simulations study is divided into two main sections, one for dichotomous scoring models and the other for polytomous scoring models. There are two sorts of effects that are studied, the main effects, like those resulting from alteration of test parameters like sample size, test length and item parameters like item difficulty. Secondly, the across model effects, i.e. to see how the simulation results are affected by the choice of model under which the simulations are done.

The Dichotomous Models

For the dichotomous models, four simulations setups were created. The first three simulation setups were for the Rasch model or the 1PL dichotomous model in which the values of the item difficulty parameter were varied. The last simulation setup was for the 2PL model. The aim was to observe any main effects of varying the item parameters and other test characteristics like sample size & test length. For every model in this section a table of results is presented followed by a discussion of those results. At the end of the section there is a general discussion of the differences in the results of the different models .

Simulation 1.1: The 1-PL model for β=0

The first simulation was carried out for the 1PL model with the fixed values of the item difficulty parameter β = 0.The results are tabulated in the Table 1.1 below.

Table 1.1: *Mean absolute error of the estimates of the covariance matrix (diagonal $\sum$ =1.0) for dichotomous items generated using the 1PL with β =0.*

M	ρ **(off-diagonal $\sum$)**	*K*	*Diagonal*	*Off-Diagonal*
1000	.8	21	0.556	0.136
	.8	63	0.119	0.042
	.4	21	0.558	0.072
	.4	63	0.124	0.027
100	.8	21	0.563	0.144
	.8	63	0.196	0.078
	.4	21	0.566	0.075
	.4	63	0.194	0.061

By examining the above table for values of β fixed at zero a main effect of test length can be observed. The estimates for both the diagonal and off diagonal elements are more precise when the test is longer, i.e. 63 items. There is a secondary effect of the sample size; the estimates are more precise when the sample size is larger. There seems to be no effect of the covariance parameter on the diagonal elements but it seems to effect the off-diagonal values; the off-diagonal differences are almost twice as less for the cases when the covariance parameter is smaller, i.e. 4.

Simulation 1.2: The 1-PL model for β=1

The second simulation was carried out for the 1PL model with the fixed values of the item difficulty parameter β = 1.The results are tabulated in Table 1.2 below.

Table 1.2: *Mean absolute error of the estimates of the covariance matrix Σ (diagonal Σ =1.0) for dichotomous items generated using the 1PL with β =1.*

M	ρ **(off-diagonal Σ)**	*K*	*Diagonal*	*Off-Diagonal*
1000	.8	21	0.803	0.182
	.8	63	0.129	0.031
	.4	21	0.810	0.079
	.4	63	0.121	0.020
100	.8	21	0.805	0.189
	.8	63	0.164	0.059
	.4	21	0.807	0.082
	.4	63	0.162	0.052

From the above table it can be seen that there is a main effect of varying the test length. When the test is longer the estimates are more accurate. There is a main effect of the covariance parameter for the non diagonal values which becomes greater when the test length is 21; the estimates are more precise when the covariance is lower. There is also a secondary effect of the sample size when the test is longer i.e. consists of 63 items; the estimates are slightly more accurate in all corresponding diagonal and off-diagonal cases.

Simulation 1.3: The 1-PL model for varying β

The third simulation was carried out for the 1PL model with varying values of the item difficulty parameter β .The item difficulties were uniformly varied between the values -

1.5 and + 1.5 such that the average value was zero. The results are tabulated in Table 1.3 below.

Table 1.3: *Mean absolute error of the estimates of the covariance matrix* $\sum$ *(diagonal* $\sum$ *=1.0) for dichotomous items generated using the 1PL for varying values of* β.

M	ρ **(off-diagonal** $\sum$ **)**	***K***	***Diagonal***	***Off-Diagonal***
1000	. 8	21	0.394	0.090
	. 8	63	0.081	0.031
	. 4	21	0.394	0.047
	. 4	63	0.076	0.021
100	. 8	21	0.405	0.105
	. 8	63	0.157	0.064
	. 4	21	0.373	0.066
	. 4	63	0.162	0.052

The results in the table above show that there is a main effect of test length. When the number of items is greater the estimates are significantly more accurate for both diagonal and off-diagonal elements of the covariance matrix. There is a secondary effect of the sample size when the test is longer; the estimates are more precise when the sample size is bigger .The covariance parameter does seem to have a main effect for the off-diagonal cases; when the covariance parameter is low the accuracy of the estimates is higher in each case, more so when the test is shorter.

Simulation 1.4: The 2-PL model

The simulation for the 2PL model was carried out while varying both the values of the item difficulty parameter and the item discrimination parameter. The item difficulty parameters were varied between -1.5 and + 1.5 with the average being zero. The item discrimination parameters were drawn uniformly between the values 0.75 and 2.25.The results of the outcomes are tabulates below in Table 2.

Table 1.4 *: Mean absolute error of the estimates of the covariance matrix $\sum$ (diagonal $\sum$ =1.0) for dichotomous items generated using the 2PL model.*

M	ρ **(off-diagonal** $\sum$ **)**	*K*	*Diagonal*	*Off-Diagonal*
1000	.8	21	0.790	0.205
	.8	63	0.190	0.066
	.4	21	0.792	0.110
	.4	63	0.182	0.033
100	.8	21	0.791	0.217
	.8	63	0.247	0.089
	.4	21	0.819	0.153
	.4	63	0.242	0.067

A main effect of test length is visible; when the test is longer the estimates are significantly more precise (for both diagonal and off-diagonal elements). There is a secondary effect of sample size when the test is longer, the estimates are more accurate for a bigger sample size. The covariance parameter has a main effect for the off-diagonal elements; when the covariance is low the accuracy of the estimates is better.

Discussion of results for dichotomous models

For all the dichotomous scoring cases it is evident that there is a main effect of the test length. When the test is longer the estimates are significantly more accurate. There is also a secondary effect in all the models: when the test is longer, the sample size affects the estimates. The estimates are more precise when the sample size is greater. The covariance parameter also seems to have a main effect on the off-diagonal cases especially more so when the number of items is lesser; when the covariance is lower the accuracy of the estimates is better.

Besides the main effects of input parameter variation there are also across-model effects. i.e. the model used seems to affect the quality of the results produced. The best estimates were made for the 1PL model when the item difficulty parameter was varied uniformly between -1.5 and 1.5 for the test while its average value was kept at zero. Results were also relatively good for the 1PL model with item difficulties for all items assigned a zero. The results were poorer for the 1PL model with all item difficulties being 1.0. The reason for this is that the abilities of the examinees were normally distributed with the average being zero. Thus a set of items with difficulty levels matching the average ability, i.e. zero, ought to yield less error in ability estimation as was the case in our simulations. This would in turn result in a more accurate estimation of the covariance between sub-scale abilities. Thus the 1PL with difficulty parameters fixed at 1.0 yielded poor results as questions were not concentrated in the vicinity where actual abilities were concentrated.

Results for varying item difficulties with average zero were better than for the simulation in which every item difficulty was fixed at zero. The plausible explanation is that though abilities were also averaged at zero, there was variation in the ability distribution and the variation in difficulties of the items administered in the tests was concurrent with the variation in the tested abilities. This provided more accuracy in ability estimation because more information can be gained from the response to an item whose difficulty is closer to the real ability of an examinee that is being tested.

Results were also relatively poor for the 2PL model with varying item difficulties and varying discrimination parameters. The distribution of the item discrimination parameters caused this effect. A plausible explanation is that the item discrimination parameters were assigned such that they were generally low for the cases when the item difficulties matched the actual abilities and generally higher in the case when difficulties of items administered were further from the actual ability level of the examinees. This would results in lesser information being gathered about the actual abilities of the examinees and thus a larger error term. This could be rectified to a certain degree by a more appropriate choice of item discrimination parameters.

.

The Polytomous Models

In the second part of the simulation study three models for polytomous item scoring were considered and simulations were carried out to observe any effects of model selection. The results of the simulation study were tabulated for the three models followed by a discussion of the differences in the results.

The set up of the simulation was as follows. In order to be able to compare the results for the three models the selection of the item and ability parameters had to be such that they are themselves not an underlying and unwarranted cause for variation in the simulations across the three models. To achieve this, the following steps were taken. For all three models the ability parameters were drawn from a standard normal distribution with mean zero as described earlier. The discrimination parameters were fixed at one. For the GPCM drawing the item parameters from a distribution was not considered, because the dependence between these parameters may result in very unfavorable values with the consequence of item categories without responses (Wilson & Masters, 1993). Therefore the β parameter values were fixed. The values of the first five items are given in the table below.

Table 2.0: Item parameter values used for simulating data using the GPCM

Item	Category			
	1	2	3	4
1	-2.0	-1.5	-0.5	0.0
2	-1.5	-1.0	0.0	0.5
3	-1.0	-0.5	0.5	1.0
4	-0.5	0.0	1.0	2.5
5	0.0	0.5	1.5	2.0

Note: The difficulty levels of item 3 are located in such a way that the category –bounds are located symmetric with respect to the standard normal ability distribution. The first *two items are shifted to the left on the latent scale, the last two items are shifted to the right.*

Data were then generated under the GPCM and using this data the item parameters of the SM and the GRM were estimated using maximum marginal likelihood (Bock & Atkin, 1981).The resulting item parameters of the SM and GRM were such that item category response curves for all three models were close. These estimated values were then used for generating data using the SM and the GRM which were then compared for all three models.

Simulation 2.1: The Generalized Partial Credit Model (GPCM)

Table 2.1: *Mean absolute error of the estimates of the covariance matrix $\sum$ (diagonal $\sum$ =1.0) for dichotomous items generated using the GPCM.*

M	ρ **(off-diagonal $\sum$)**	***K***	***Diagonal***	***Off-Diagonal***
1000	.8	24	0.078	0.032
	.8	48	0.066	0.027
	.4	24	0.066	0.017
	.4	48	0.059	0.016
100	.8	24	0.140	0.059
	.8	48	0.138	0.058
	.4	24	0.150	0.056
	.4	48	0.144	0.045

Simulation 2.2: The Sequential Model(SM)

Table 2.2*: Mean absolute error of the estimates of the covariance matrix $\sum$ (diagonal $\sum$ =1.0) for dichotomous items generated using the SM.*

M	ρ **(off-diagonal $\sum$)**	***K***	***Diagonal***	***Off-Diagonal***
1000	.8	24	0.094	0.033
	.8	48	0.055	0.022
	.4	24	0.093	0.024
	.4	48	0.052	0.016

100	.8	24	0.183	0.070
	.8	48	0.131	0.058
	.4	24	0.174	0.061
	.4	48	0.138	0.047

Simulation 2.3: The Graded Response Model (GRM)

Table 2.3: *Mean absolute error of the estimates of the covariance matrix $\sum$ (diagonal $\sum$ =1.0) for dichotomous items generated using the GRM.*

M	ρ **(off-diagonal** $\sum$ **)**	***K***	***Diagonal***	***Off-Diagonal***
1000	.8	24	0.085	0.027
	.8	48	0.047	0.019
	.4	24	0.080	0.021
	.4	48	0.051	0.017
100	.8	24	0.178	0.071
	.8	48	0.148	0.062
	.4	24	0.204	0.067
	.4	48	0.148	0.049

Discussion of results for Polytomous models

By observing the above tables it can be seen that there is no significant difference between the results obtained for the three models; in other words they work equally well with the limited information EM algorithm. However there are similar main effects of input parameter variation within the three models. The first main effect that is quite significant is that of the ‘number of persons’ parameter. When the number of persons is greater the estimates are significantly more accurate. There is also another main effect which is less pronounced. It is caused by the number of items. When the number of items is greater the estimates are more precise. The covariance parameter has a relatively small ‘main’ effect on off-diagonal values; the estimates are slightly more accurate for lower covariance. These results are consistent with the results obtained using dichotomously scored items.

Conclusions

The aim of this study was to see how well the limited-information maximum-likelihood estimation technique suggested by Rubin and Thomas (2001) works as a function of test length, sample size, parameter choice for dichotomous items and model choice for polytomous items. Rubin and Thomas (2001) only give an empirical example in their paper which does not answer these questions. This study employs their technique in a simulation study for calculating between scales co-variance in order to know its feasibility via-a-vis parameter variation and model selection.

The setup of the study was divided into two parts, for dichotomous models and polytomous models. In the first instance simulations were carried out on three cases of the 1PL model in which the difficulty parameters were varied. A fourth simulation of the dichotomous case was of the 2PL model. For the polytomous cases model comparison was done between the GPCM, GRM and the SM.

Some results were as expected. There was a main effect of test length. The effect was stronger in the dichotomous models than in the polytomous models. In the polytomous models the strongest main effect was caused by the sample size; when the number of persons was greater there was a large improvement. In the dichotomous models there wasn’t any main effect caused by sample size. However there was a secondary effect of the sample size when the test length was longer. The covariance parameter also has a main effect in both polytomous and dichotomous model; the differences being more pronounced for the dichotomous models. The effect was that when the covariance is lesser the estimates are more precise. For the dichotomous models the improvement was even more when the number of items was lesser.

However in the dichotomous models case there was a marked difference produced by parameter selection. The best results were for the 1PL cases where the average values of the difficulty parameter matched the average value of the abilities which were drawn from a normal distribution. In the 1PL case where the average value of the difficulty parameter was further away from the average of the drawn ability values the estimates were less accurate because of the larger measurement error in estimating the abilities. The results for the 2PL model were also poor which were caused due to selection of inappropriate discrimination parameters in relation to the abilities of the students and items administered to them.

In the polytomous models the estimates were similar for all three models and no significant change could be detected.

Thus based on the results obtained it would be fair to conclude that the limited information marginal maximum likelihood technique employed in the study works equally good for the three polytomous models with sample size having a significant main effect on results within each model. For the dichotomous cases the method works better when item parameters are appropriately selected. Furthermore the test length has a significant main effect within the dichotomous models; the long tests producing significantly more accurate estimates than short tests.

Bibliography

Dempster, A., Laird, N., and Rubin, D.(1977). Maximum Likelihood estimates from incomplete data via the EM algorithm(with discussion),*Journal of the Royal Statistical Societ,Series* B,39,1-38

Masters, G.N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*, No. 2, pp. 149-174.

Masters, G.N. & Wright, B.D. (1997). *Handbook of Modern Item Response Theory*. van der Linden, W.J & Hambleton, R.K. (eds.). New York : Springer.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*, No. 2, pp. 159-176.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph 7.*

Lord, F.M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18,* 57-75.

Lord, F.M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-548.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.