COMPUTER ADAPTIVE TESTING WITH EMPIRICAL PRIOR INFORMATION: A GIBBS SAMPLER APPROACH FOR ABILITY ESTIMATION[1]

MARIAGIULIA MATTEUCCI
UNIVERSITY OF BOLOGNA
e-mail: m.matteucci@unibo.it

BERNARD P. VELDKAMP
UNIVERSITY OF TWENTE
e-mail: B.P.Veldkamp@gw.utwente.nl

[1] Corresponding author: Mariagiulia Matteucci, Statistics Department "P. Fortunati", University of Bologna, via Belle Arti 41, 40126 Bologna (Italy).

COMPUTER ADAPTIVE TESTING WITH EMPIRICAL PRIOR INFORMATION: A GIBBS SAMPLER APPROACH FOR ABILITY ESTIMATION

Abstract

In this paper, empirical prior information is introduced in computer adaptive testing. Despite its increasing use, the method suffers from a weak measurement precision, especially under particular conditions. Therefore, it is shown how the inclusion of background variables both in the initialization and the ability estimation is able to improve the accuracy of ability estimates. In particular, a Gibbs sampler scheme is proposed in the phases of interim and final ability estimation. By using simulated data, it is demonstrated that the method produces more accurate ability estimates, especially for short tests and when reproducing boundary abilities.

Keywords: adaptive testing, empirical prior information, Gibbs sampler, measurement precision.

1. Introduction

In recent years, we have assisted to a rapid development of computer-based testing in the field of educational assessment, especially in adaptive testing. Furthermore, the practice of conducting the test administration via adaptive testing is becoming more and more well-established. Since the early 1970s (Lord, 1970; 1971), studies have been conducted to develop the theoretical framework of computerized adaptive testing (CAT) (see e.g. van der Linden and Glas, 2000; Wainer et al., 2000). The basic idea of CAT is to simulate the behavior of a real oral examiner during a testing occasion. The most likely situation is that he/she would start with an initial item and, depending on the examinee's response, proceed with a more difficult or easier item, until the examinee's grade of proficiency becomes sufficiently precise. Analogously, in computer adaptive testing a first item is submitted to the test-taker: if the item is endorsed, a more difficult item is presented, otherwise an easier one is selected by the algorithm to be submitted. The procedure goes on until a pre-specified criterion is met and a measure of the examinee's proficiency is given. In this sense, the algorithm is adapted to the candidate because items are chosen exactly to met his/her specifications in terms of ability.

CAT relies strongly on item response theory (IRT), developed in order to estimate individual and item characteristics after a test administration (see e.g. Lord and Novick, 1968). In fact, the item pool is calibrated according to a particular IRT model, based on data nature and fit, and the response process is assumed to follow a chosen model.

Despite the wide use of computer adaptive testing, the method has a number of problems in use as item pool maintenance, test assembly and item exposure. Furthermore, technical issues as initialization, ability estimation, algorithm stopping rule should be improved, especially under particular conditions.

In this study, the focus is on the ability estimation for short CAT tests. Nowadays, in the area of psychological measurement but not only, examiners are more and more interested in obtaining as much information as possible about candidates by using restricted resources. This goal may be achieved by using short test versions, which allow a considerable saving of time and money. Moreover, an important consequence in adaptive testing is the decrease of item overexposure because, given a fixed number of items in the pool, items are selected less frequently. On the other hand, the main interest in CAT is in the precision of the ability estimates, which may be not sufficiently accurate adopting short tests. A possible solution to the estimation improvement is represented by the introduction of background information about individuals. In fact, besides the candidates' responses, more and more information about individuals is stored in databases (e.g. think about large scale educational assessments as PISA or TIMMS) and can be used in order to obtain more accurate estimates of candidates' degree of proficiency. Background variables may be included in CAT in two different stages. Firstly, the initialization of ability estimate can make use of prior information (see van der Linden, 1999). As a consequence, a better provisional ability estimate is provided and the first item is selected closer to the true ability of the person. Secondly, background variables may be included in the estimation process through an empirical prior distribution. A natural context this approach can be developed within is represented by Bayesian statistics, where likelihood and prior distributions are combined in order to obtain the posterior distribution of interest. Recently, Markov chain Monte Carlo (MCMC) methods, and particularly the Gibbs sampler (Geman and Geman, 1984), have been applied extensively in IRT estimation because they are able to provide flexible algorithms for a large variety of models. By introducing the empirical prior within MCMC, the posterior distribution is more informative about the candidate and better ability estimates can be obtained.

In the paper of van der Linden (1999) it is shown how prior information can be included in the ability initialization. On the other hand, the purpose of this paper is to show how collateral information can be used even more efficiently by introducing it both in initialization and ability estimation. Furthermore, the paper describes how the empirical prior can be integrated in the estimation process within the Gibbs sampler scheme.

The paper first gives an overview of how prior information can be included in CAT. Then, it is shown how the Gibbs sampler can be implemented in computer adaptive testing effectively in order to integrate information coming from both likelihood and prior distributions. The advantages of using background variables in CAT administration are discussed through a simulation study, which reports levels of ability precision when empirical prior is introduced instead of standard priors. A special attention is given to the case of short tests (e.g. test length less than 10 items) in order to show the potentialities of the algorithm.

2. Adaptive Testing with Empirical Prior Information

In testing occasions, besides the candidates' responses on a target test, a set of individual covariates may be available. Background variables may include scores obtained by the examinees on other tests or testlets, socio-economic or demographical variables, and so on. Given the availability of such information, its inclusion in the investigation of candidates’ ability does make sense. Whether and how collateral information about examinees may be included in IRT ability estimation has been discussed by various authors (e.g. Zwinderman, 1991; 1997; van der Linden, 1999; Matteucci and Veldkamp, 2008). As reported in van der Linden and Pashley (2000), one reason for introducing collateral information about the candidates in adaptive testing is its weakness in the ability estimation when dealing with short tests, caused by a possible bad start in the ability initialization. Even if it is well known that the convergence of the algorithm is not affected by the choice of starting values, a rough initial inference about ability may cause a very slow convergence. In the following, the different steps of CAT with empirical prior are described. A particular section is dedicated to the ability estimation.

2.1 The Phases of CAT

Typically, in computerized adaptive testing, the item parameters are treated as known and the main purpose of test administration is the ability estimation of test takers. In the common practice of item pool calibration, the item parameters are estimated on the basis of a particular IRT model. The model should be able to reproduce the individuals' response process; therefore, it describes the mathematical function linking the response probability to a set of item parameters and ability. Once the item parameters have been estimated with sufficient precision, items with target features are included in the item pool to be administered. The choice of the model depends on different issues as item format, dimensionality specification, and fit. For the purpose of this study, the unidimensional two-parameter normal ogive (2PNO) model (Lord, 1952; Lord and Novick, 1968) is assumed to underlie the response process. The model has been designed for binary observed data, employing a cumulative standard normal distribution to express the probability of a correct response to an item *j*, with *j=1,...,k* items, as a function of ability and item parameters, as follows

$$P(Y_j = 1|\theta) = \Phi(\alpha_j\theta - \delta_j) = \int_{-\infty}^{\alpha_j\theta-\delta_j} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \qquad (1)$$

where Y_j is the random response variable for item *j*, taking the value *1* for a correct response and *0* otherwise, α_j and δ_j are the item discrimination and difficulty respectively, and θ is the unidimensional ability. Model (1) assumes unidimensionality, *i.e.* a single latent trait accounts for the individual responses. Depending on the data characteristics, other models are possible and have been employed in CAT.

Once the items have been calibrated according to an IRT model, computer adaptive testing works with the following steps:

1. Ability initialization.
2. Item selection.
3. Item administration.
4. Ability estimate update.

Steps 2-4 are repeated iteratively until a stopping rule is satisfied and a final estimate of the candidate's ability is obtained. Potentially, an empirical prior may be introduced both in the initialization of the algorithm (step 1) and in the interim-final ability estimation (step 4).

In the first step, an initial provisional ability value is required for the procedure to start. The ability initialization may be fixed, random or “adaptive”. When the initialization is fixed, each test-taker is assigned to the same ability starting value, typically equal to zero to reproduce the average value in the ability domain [-3; 3], while a random initialization provides a different but random allocation in the ability scale for each individual. Both solutions may come up with starting values which are very far from the true ability level. On the other hand, an adaptive initialization for the algorithm may be conducted providing ad hoc starting values based on background information about the candidate. In order to introduce empirical information, a relation between the ability θ and a set of P individual covariates $\{X_p\}$, with *p=1,…,P*, is assumed in the form of a linear regression, as follows

$$\theta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_P X_{iP} + \varepsilon_i, \tag{2}$$

where the error term are assumed to be independent and normally distributed as $\varepsilon_i \sim N(0, \sigma^2)$, with *i=1,…,n* individuals. The assumption of a linear regression model is translated into a normal conditional distribution of θ_i given the covariates, as

$$\theta_i | X_{i1}, \dots, X_{iP} \sim N(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_P X_{iP}; \sigma^2). \tag{3}$$

Equation (3) represents the informative prior distribution for ability. When regression (2) is estimated with satisfying precision and the quality of the background variables is good, *i.e.* they are high predictors, the estimated regression coefficients may be used in order to initialize the ability in CAT for a generic examinee *i* with realizations $(x_{i1}, \dots, x_{iP})$, as follows

$$\hat{\theta}_{i0} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_P x_{iP}. \tag{4}$$

The advantage of using *ad hoc* information to initialize the algorithm is mainly to shorten the procedure. Within this approach, initial values may be much more reliable and accurate initial inferences about ability are able to shorten time to convergence significantly.

Before proceeding with item selection (step 2), the following notation on CAT is introduced. Given *J* calibrated items in the pool, indexed by *j=1,…,J*, denote the rank of selected items as *k=1,…,K*. Hence, when choosing the *k*th item to be administered: j_k is the index of the chosen item, $S_{k-1}=\{j_1,j_2,\ldots,j_{k-1}\}$ is the set of selected items and $R_k=\{1,\ldots,J\}\backslash S_{k-1}$ is the set of remaining items in the pool. In the following, the index *i=1,…,n* of examinees is omitted and the test administration is referred to a generic candidate *i* implicitly.

In order to proceed with the item selection (step 2), various criteria have been proposed in the literature. A classical and straightforward method which is also applied in linear testing is the maximum-information criterion (Birnbaum, 1968). When selecting the *k*th item, the method works choosing the item which maximizes Fisher's expected information function at the current ability value $\theta = \hat{\theta}_{k-1}$, as follows

$$j_k \equiv \arg\max_j \{I_j(\hat{\theta}_{k-1}); j \in R_k\}. \tag{5}$$

The form of the information function depends on the particular chosen IRT model. According to model (1), the information function becomes

$$I_j(\hat{\theta}_{k-1}) = \alpha_j^2 \frac{\{(2\pi)^{-1/2}\exp(-\eta_j^2/2)\}^2}{\Phi(\eta_j)[1-\Phi(\eta_j)]}\ , \tag{6}$$

where $\eta_j = \alpha_j\hat{\theta}_{k-1} - \delta_j$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The method is widely used; nevertheless, the maximum-information criterion associated with a fixed ability initialization leads to the problem of item overexposure. In fact, when CAT is initialized at the average ability level $\hat{\theta}_0 = 0$, or at any fixed value, adopting Birnbaum's criterion means selecting the same first item for each examinee indifferently. On the other hand, the use of empirical information in the ability initialization, together with the maximum-information criterion for item selection, avoids this problem. Different criteria for adaptive item selection have been proposed, also adopting a Bayesian approach (see e.g. Owen's procedure; Owen, 1969; 1975). For the purpose of the current work, the classical maximum-information rule is adopted. In fact, prior information is not directly introduced at this phase and a classical criterion is considered sufficient. For a review of both classical and modern procedures for item selection, see van der Linden & Pashley, 2000.

Following CAT algorithm through step 3, the chosen item is administered to the test-taker and the answer is recorded. The response is subsequently used in step 4, when ability should be estimated. Steps 2-4 of the algorithm are repeated iteratively until a stopping rule is satisfied, as a fixed test length or a pre-specified level of precision for the ability estimate.

One crucial issue in CAT certainly is the measurement precision of ability estimates. Typically, standard errors of ability score estimates are not negligible and efforts in the direction of improving the accuracy of ability estimates should be done, especially under particular conditions. In fact, the task of obtaining an accurate ability estimate can be hard when poor information comes from the responses or when the examinee’s level of proficiency is extreme (very high or very low). As a consequence, in case the test is particularly short or when it is difficult to calibrate the items around the individual ability, the use of prior information is highly recommended.

In adaptive testing, a number of methods for the ability estimation are in use. These include maximum likelihood procedures or Bayesian methods (see van der Linden and Pashley, 2000). Due to its growing and relatively new use in IRT, a Gibbs sampler scheme is implemented for ability estimation in CAT. The algorithm, as shown in Matteucci and Veldkamp (2008), is able to integrate efficiently data coming from individual responses and empirical prior information. The method is illustrated in detail in the next section.

2.2 MCMC Ability Estimation

To perform a Bayesian ability estimation in CAT, the Gibbs sampler (Geman and Geman, 1984) is implemented. The algorithm belongs to the family of Markov chain Monte Carlo (MCMC) methods which introduce simulation for the purpose of reproducing a target distribution by using one or more sequences of correlated random variables. According to the Bayesian approach, both ability and item/regression parameters are regarded as random variables. Once all components of the joint posterior distribution of interest have been individuated, the single conditional distributions should be specified. The Gibbs sampler works creating suitable samples from each single conditional distribution iteratively until convergence. Among others, Albert (1999), Béguin and Glas (2001), Fox and Glas (2001), and Matteucci and Veldkamp

(2008) dealt with Gibbs sampler estimation within item response theory models. In the current work, the algorithm is modified in order to estimate ability in adaptive testing with the inclusion of an empirical prior.

Generally, the presence of the binary response variable Y_j can be modeled by introducing continuous underlying variables Z_j, which are independent and identically distributed as $Z_j \sim N(\alpha_j \theta - \delta_j; 1)$. The relation between the observed and the underlying variables is the following

$$Y_j = \begin{cases} 1 & if\ Z_j > 0, \\ 0 & if\ Z_j \leq 0. \end{cases} \tag{7}$$

According to Equation (7), the continuous variable Z is greater than zero if and only if the corresponding observed response Y is a success; the *underlying variable* approach (Bartholomew, 1987; Bartholomew and Knott, 1999) describes the partition of the continuous variable Z in order to represent the dichotomy of Y.

From a fully Bayesian perspective, the joint posterior distribution of interest is

$$P(\boldsymbol{Z}, \theta, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2 |\, \boldsymbol{Y}, \boldsymbol{X}) = P(\boldsymbol{Z} |\, \theta, \boldsymbol{\xi}, \boldsymbol{Y}) P(\theta | \boldsymbol{\beta}, \sigma^2, \boldsymbol{X}) P(\boldsymbol{\xi}) P(\boldsymbol{\beta}) P(\sigma^2), \tag{8}$$

where ξ is the vector including all item parameters. In linear testing, given the data on the responses and the observed covariates, the Gibbs sampler would have worked iteratively sampling from the following single conditional distributions:

1. $\mathbf{Z} \mid \theta, \xi$.
2. $\theta \mid \mathbf{Z}, \xi, \boldsymbol{\beta}, \sigma^2$.
3. $\xi \mid \theta, \mathbf{Z}$.
4. $\boldsymbol{\beta} \mid \theta, \sigma^2$.
5. $\sigma^2 \mid \theta, \boldsymbol{\beta}$.

On the other hand, in adaptive testing both item and regression parameters are treated as known; therefore, their conditional distributions are not needed in the scheme. In CAT, the Gibbs sampler works only with the conditional distribution of the underlying response variables Z_j (distribution in step 1) and the posterior distribution of the ability θ (distribution in step 2), in order to proceed with the ability estimation. The single conditional distributions, compared to the joint posterior, are treatable and easy to draw samples from.

With regard to the first conditional distribution, a classical result (see e.g. Johnson and Albert, 1999, chapter 3) is that the distribution of each Z_j given the ability and the item parameters is a truncated normal, as follows

$$Z_j|\theta,\xi\sim\begin{cases}N(\eta_j;1) & with\, Z_j>0\; if\; Y_j=1,\\ N(\eta_j;1) & with\, Z_j\leq 0\; if\; Y_j=0.\end{cases} \tag{9}$$

The conditional distribution of the underlying variables Z_j is normal, with expected value equal to $\eta_j=\alpha_j\theta-\delta_j$ and variance 1, truncated by 0 to the left if $Y_j=1$ (correct response to item *j*) and the right if $Y_j=0$ (incorrect response to item *j*).

The second conditional distribution is obtained combining the likelihood and the informative prior distribution, according to Bayesian conjugate families of distributions. Starting from the normal regression model $Z_j=\alpha_j\theta-\delta_j+\upsilon_j,$ for $j=1,\dots,J$, we obtain

$$Z_j+\delta_j=\alpha_j\theta+\upsilon_j, \tag{10}$$

where υ_j are *i.i.d.~N(0;1)*. Equation (10) is simply the regression of the terms on the left side $Z_j+\delta_j$ on the independent variable α_j, where θ is the regression coefficient. Hence, the likelihood function of the ability θ follows a normal distribution, as

$$\theta\sim N(\hat{\theta};\, v), \tag{11}$$

where $\hat{\theta}=(\alpha_j'\alpha_j)^{-1}\alpha_j'(Z_j+\delta_j)$ is the least square estimate of θ and $v=(\alpha_j'\alpha_j)^{-1}$ is the variance. Practically, the variance can be calculated as $v=1/\sum_{j=1}^{J}\alpha_j^2$ and the expected value as $\hat{\theta}=\sum_{j=1}^{J}\alpha_j(Z_j+\delta_j)/\sum_{j=1}^{J}\alpha_j^2.$ The prior distribution for the ability is the empirical normal prior (3) and the combination of likelihood and prior leads to a normal posterior distribution, as follows

$$\theta|\boldsymbol{Z},\boldsymbol{\xi},\boldsymbol{\beta},\sigma^2\sim N\left(\frac{\hat{\theta}/v+\boldsymbol{X\beta}/\sigma^2}{1/v+1/\sigma^2};\frac{1}{1/v+1/\sigma^2}\right). \tag{12}$$

After the *k*th item has been administered, the Gibbs sampler is able to simulate ability as follows:

1. Start with known item parameters ξ and a provisional estimate of $\theta_k^{(0)}$, $\theta_k^{(0)} \equiv \theta_{k-1}$, and sample $\boldsymbol{Z}^{(0)}$ from distribution (9), with $j \in S_k$.
2. Use $\boldsymbol{Z}^{(0)}$ and known $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, σ^2 to sample $\theta_k^{(1)}$ from distribution (12).
3. Repeat steps 1-2 with the updated values, iteratively.

The steps describe the estimation of the interim ability. Simply, after the last item has been administered, the same steps may be applied with the updated likelihood in order to obtain the final ability estimate. The Gibbs sampler has been implemented in the software MATLAB 7.1 (The MatWorks Inc., 2005) .

3. Simulation Studies

The design of an adaptive test may be very complicated and several decisions should be taken into account in order to proceed with the algorithm. A popular choice in standard CAT, is to initialize the ability at a fixed value (e.g. zero), to adopt the maximum-information item selection and to estimate the ability on the sole basis of the individual responses. The evident consequences of using this approach are the overexposure of the first item, which is repeatedly administered, and the absence of a useful source of information as it is prior knowledge in the ability estimation.

In order to compare the accuracy of ability estimates in adaptive testing by using different criteria for the initialization and the ability estimation, simulation studies are conducted under different conditions. The first simulation is designed to compare the performances of the algorithm with and without empirical prior for different test lengths. In the second study, different settings are evaluated for a very short test of length equal to 5. In particular, the estimation results are compared for the MCMC CAT proposed by the authors, CAT without empirical prior, and CAT with only empirical ability initialization. Finally, the issue of the algorithm convergence is taken into account.

3.1 Prior in Use: a Comparison with Different Test Length

The purpose of the first simulation study is to show the potentiality of the empirical prior use in the parameter recovery within the Gibbs sampler scheme. To this aim, two different CAT designs are compared: the first one follows the common practice of initializing the ability at zero and assuming a standard normal as a prior for the ability distribution, while the second one adopts an empirical prior both in the initialization and in the ability estimation, as shown in the previous section. For simplicity of description, the former approach is denominated *standard* while the latter is called *empirical*. In both cases, item selection is conducted by using the maximum-information criterion. This is equivalent to choose the item which difficulty is the closest to the provisional ability estimate of the simulee.

In the study, an item bank of 200 items is employed, with item parameters sampled as $\alpha_j \sim \mathcal{U}(0.7;2)$ and $\delta_j \sim \mathcal{U}(-3;3)$, for *j=1,...,*$k$. When the empirical approach is adopted, the linear relation $\theta = 0.2 + 0.7X + \varepsilon$ with $\varepsilon \sim N(0{,}0.3)$ is assumed between the ability θ and a single covariate *X*. Since θ is known in the simulation, the distribution of *X* is normal with parameters depending on the linear combination of the normal distribution of ε. Response generation is simulated for different levels of ability from -3 to 3 according to model (1). In order to get results for tests consisting of a different number of items, the CAT stopping rule is defined as fixed test length of 5, 10, 15 and 20 items. The Gibbs sampler with a chain length of 5000 iterations and burn-in of 500 is employed for the ability estimation. The output consists in the mean and standard deviations sampled from the posterior distribution of ability. The choice of the chain length and the number to discard iterations are motivated by the convergence study described in Section 3.3. A number of 100 replications have been conducted in the simulation. Besides the expected a posterior estimate and the standard deviation, also the average bias and the root mean square errors (RMSE) have been calculated. Both indicators compare the distance between the true and the simulated values in each replication. In particular, given Q replications, bias is defined as

$$Bias(\hat{\theta}) = 1/Q \sum_{q=1}^{Q} (\hat{\theta}_q - \hat{\theta}_{true}), \tag{13}$$

Where $\hat{\theta}_q$ is the estimated ability for replication *q*, with *q=1,…,*Q, and $\hat{\theta}_{true}$ is the simulee’s true ability. RMSE is calculated as follows

$$RMSE(\hat{\theta}) = \left[1/Q \sum_{q=1}^{Q} \left(\hat{\theta}_q - \hat{\theta}_{true}\right)^2\right]^{1/2}. \tag{14}$$

Table 1 provides the results of the simulation study in case of a very short test consisting of 5 items.

[INSERT TABLE 1 ABOUT HERE]

As can be easily noticed, compared with the standard version of CAT, the parameter recovery of empirical CAT is more accurate in terms of posterior mean and bias, especially when deviating from $\theta = 0$. Particularly, for boundary abilities as -3, -2.5, 2.5 and 3, standard CAT produces seriously biased estimates while the introduction of ad hoc prior information leads to smaller bias. Clearly, few items bring very little information about the individual trait and the prior distribution really dominates the estimation. Across the different true ability values, standard deviation and RMSE are always smaller in the empirical solution than in the standard one.

Analogous conclusions can be drawn from Table 2, where the simulated values are reported for a test consisting of 10 items.

[INSERT TABLE 2 ABOUT HERE]

When increasing the number of items, an improvement in the precision of estimates is denoted for standard CAT in terms of bias, even if standard deviations and RMSE are always larger than the empirical solution. Table 3 and 4 show the results of the simulations conducted for adaptive tests of 15 and 20 items, respectively.

[INSERT TABLE 3 ABOUT HERE]

[INSERT TABLE 4 ABOUT HERE]

Due to the increasing number of items, standard CAT becomes more precise, and slightly outperforms the empirical approach in terms of posterior mean and bias. However, the empirical CAT shows a better performance in terms of standard deviations and RMSE. The comparison of true and simulated values for central abilities suggests that there are not considerable differences in reproducing the ability values between the two approaches.

The whole simulation study suggests that the introduction of an informative prior leads to an improvement of the measurement precision in the individual ability assessment. This improvement becomes very evident for short tests and when shifting to boundary ability values. This evidence cannot be generalized to the case of longer test (e.g. more than 20 items): it is well known that, when the test length increases, the prior distribution lacks in strength and the two solutions become more and more similar.

3.2 Introduction of Prior Information at Different Levels

According to the findings of the previous study, the use of prior information in CAT shows its maximum effectiveness in case of very short tests. In the current simulation study, the focus is on the comparison of different levels of prior information for a target test consisting of 5 items. Results of Table 1 regarding empirical and standard CAT are compared to an intermediate solution, named *semi-empirical*, where empirical information is used only in the initialization of the ability estimate. In fact, when problems of fairness arise, it can be very difficult to justify the introduction of background variables directly related to the evaluation of the performances in the estimation. Table 5 illustrates the results of the simulation.

[INSERT TABLE 5 ABOUT HERE]

The empirical initialization CAT shows an intermediate behavior respect to the other two approaches. In fact, from the comparison between the estimated and the true ability values, it can be seen that the results are more precise than standard CAT but less accurate than empirical CAT, when deviating from 0. Also, standard deviations are always intermediate between the correspondent values of the standard and empirical approaches. Figure 1 shows the trend of the RMSE across the ability true values, for the three approaches.

[INSERT FIGURE 1 ABOUT HERE]

Figure 1 provides a clear visualization of the potentiality of using empirical information both in the initialization and in the ability estimation when dealing with short tests. For the empirical solution, the RMSE curve is always below than the curves associated with the standard and the semi-empirical approaches. The difference in precision is particularly significant for ability levels that satisfy $|\theta| \geq 2$.

3.3 A Note on the Algorithm Convergence

One of the most critical issues in MCMC estimation, is assessing the convergence of the algorithm. A large number of researchers have approached the problem turning out with different solutions, sometimes conflicting (for a review, see Cowles and Carlin, 1996). When simulating a MCMC chain, the first thing is to check the trace plot of the simulated random draws. Even if convergence cannot be ensured by simply looking at the iteration history, a clearly critical situation of not-convergence can be detected immediately. After computing the posterior mean and the standard deviation, a measure of the standard error of estimate should be computed. As suggested in Gelman, Carlin, Stern and Rubin (2004, chap. 10), an approximate measure of the accuracy of the sample mean estimate is the standard deviation divided by the square root of the number of simulations, which is nothing but the posterior deviance. Moreover, an estimate of the Monte Carlo standard error should be computed. As a rule of thumb, the estimated Monte Carlo error should be less than 5% of the standard deviation.

In order to decide the necessary number of iterations for obtaining an acceptable accuracy, a study has been conducted by simulating single chains. In particular, the simulation design of Section 3.1 is drawn on in the case of ability $\theta=0$ and test length $T=5$. The purpose of the study is to evaluate the accuracy of the posterior mean in simulations conducted by using different number of iterations (1000, 2000, 5000 and 10000). Table 6 shows the results both for the empirical and the standard approaches.

[INSERT TABLE 6 ABOUT HERE]

The number of iterations is specified in the first column, while the number to discard iterations (burn-in phase) is contained in column 2. Besides the posterior mean and the standard deviation, an estimate of the Monte Carlo error (MC error) is reported. The estimate is calculated as the square root of the spectral density variance estimate divided by the number of actual iterations (time-series estimate), and it was proposed by Geweke (1992) as an estimate of the asymptotic standard error. The MC error has been calculated by using the R package BOA.

One single replication, depending on the number of iterations in the chain, took only few seconds to complete (from 2 to 6 seconds) on a 2.27 GHz Intel Core2 laptop. The simulations conducted by using 1000 iterations do not satisfy the accuracy condition of MC error less than 5% of the standard deviation, while the solution with 2000 iterations slightly satisfy it. On the other hand, running 5000 or 10000 iterations turns out with MC errors significantly lower than the 5% of standard deviation and are considered a good standard of accuracy.

As a consequence of these results, the adopted number of iterations was settled to 5000. The chosen chain length represents a good compromise between the estimate accuracy and the time needed to complete the algorithm. When running 100 replications, each one with 5000 iterations, only 6 minutes are needed to complete. Therefore, this solution allows a fast implementation in the real practice of testing.

Figure 2 shows the trace plot of the simulation with 5000 iterations, when prior information is included.

[INSERT FIGURE 2 ABOUT HERE]

Clearly, the plot shows a random fluctuation of the sample values around the mean. The absence of autocorrelation (at least at a lag higher than 5) is confirmed by the autocorrelation plot reported in Figure 3.

[INSERT FIGURE 3 ABOUT HERE]

Usually, one of the main drawbacks of MCMC is the time consuming and slow convergence of the algorithm; however, adopting the above mentioned features for the chain, the simulation represents a good compromise between speed and accuracy. Of course, we should also mention that the model implemented is rather simple, because it is a unidimensional model for binary indicators. Probably, the extension of the algorithm to more complicated model, as multidimensional models, would come out with a slower convergence.

4. Discussion

The study introduced the problem of ability estimation in computer adaptive testing under particular situations of uncertainty about the candidate’s level of proficiency. Examples are CAT consisting of a small number of items or candidates with latent ability far from average. A solution to these cases is represented by the introduction of prior information in the algorithm in order to obtain more accurate ability estimates. This approach is developed within the MCMC methods, particularly adopting the Gibbs sampler to integrate likelihood with empirical prior information about the candidate.

The findings of simulation studies suggest that the introduction of informative priors is effective in improving the accuracy of ability estimates, especially when dealing with short tests and when the ability is far from zero. In particular, the measurement precision is improved when empirical priors are introduced both to initialize and to estimate ability. The use of empirical information is highly recommended with short tests (e.g. length equal to 5), when the standard approaches based on a standard normal prior fail to reproduce the true ability values. Despite the great availability of background variables concerning the individuals, the quality of information remains a fundamental issue. The usefulness of the described approach depends highly on the predictive capability of the collateral variables.

When problems of fairness arise and empirical information cannot be used in the ability estimation, an initial inference which is as close as possible to the true ability value is recommended, i.e. an empirical CAT initialization is desiderable.

References

Bartholomew, D.J. (1987). *Latent variable models and factor analysis*. New York: Oxford University Press.

Bartholomew, D.J., & Knott, M. (1999). *Latent variable models and factor analysis.* London: Arnold Publishers.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Cowles, M.K., & Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91, 883-904.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis, 2nd edition*. Boca Raton, Florida: Chapman and 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.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J.M. Bernardo, J. Berger, A.P. Dawid & A.F.M. Smith (Eds.), *Bayesian statistics 4* (pp. 169-193). Oxford,U.K.: Oxford University Press.

Johnson, V.E., & Albert, J.H. (1999). *Ordinal data modeling*. New York: Springer-Verlag.

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

Lord, F.M. (1970). Some test theory for tailored testing. In W.H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139-183). New York: Harper and Row.

Lord, F.M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement*, *8*, 147-151.

Matteucci, M., & Veldkamp, B.P. (2008). Including empirical prior information in test administration. In B. Fishet, D. Piccolo & R. Verde (Eds.), *Book of short papers, First Joint Meeting SFC-CLADAG of the Italian Statistical Society* (pp. 97-100). Napoli: Edizioni Scientifiche Italiane.

Owen, R.J. (1969). *A Bayesian approach to tailored testing* . Research Report 69-92. Princeton, NJ: Educational Testing Service.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

van der Linden, W.J. (1999). Empirical initialization of the trait estimation in adaptive testing. *Applied Psychological Measurement*, *23*, 21-29.

van der Linden, W.J., & Glas, C.A.W. (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer Academic Publishers.

van der Linden, W.J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25) . Boston, MA: Kluwer Academic Publishers.

Wainer, H., Dorans, N.J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R.J., & Steinberg, L. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Zwinderman, A.H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589-600.

Zwinderman, A.H. (1997). Response models with manifest predictors. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 245-256) . New York: Springer-Verlag.

TABLE 1

Ability parameter recovery for empirical and standard solutions (T=5).

True θ	Empirical				Standard			
	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE
-3	-3.09	0.57	-0.09	0.42	-2.33	0.88	0.67	0.69
-2.5	-2.61	0.53	-0.11	0.37	-2.21	0.82	0.29	0.39
-2	-2.06	0.46	-0.05	0.38	-1.88	0.70	0.12	0.41
-1	-1.02	0.41	-0.02	0.28	-0.91	0.51	0.09	0.30
0	-0.08	0.40	-0.08	0.31	0.04	0.46	0.04	0.33
1	1.04	0.40	0.04	0.27	0.96	0.50	-0.04	0.35
2	2.07	0.47	0.07	0.34	1.94	0.77	-0.06	0.43
2.5	2.58	0.53	0.08	0.34	2.19	0.87	-0.31	0.36
3	3.13	0.56	0.13	0.44	2.23	0.89	-0.77	0.78

TABLE 2

Ability parameter recovery for empirical and standard solutions (T=10).

True θ	Empirical				Standard			
	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE
-3	-3.08	0.64	-0.08	0.31	-2.88	0.98	0.12	0.34
-2.5	-2.66	0.59	-0.16	0.35	-2.50	0.80	-0.00	0.40
-2	-2.09	0.48	-0.09	0.29	-1.98	0.62	0.02	0.41
-1	-1.04	0.41	-0.04	0.22	-0.99	0.47	0.01	0.25
0	0.03	0.39	0.03	0.17	0.03	0.42	0.03	0.22
1	1.00	0.39	0.00	0.22	0.99	0.45	-0.01	0.26
2	2.12	0.48	0.12	0.28	1.97	0.62	-0.04	0.33
2.5	2.66	0.59	0.16	0.33	2.44	0.81	-0.06	0.35
3	3.08	0.64	0.08	0.28	2.83	0.97	-0.17	0.34

TABLE 3

Ability parameter recovery for empirical and standard solutions (T=15).

True θ	Empirical				Standard			
	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE
-3	-3.11	0.71	-0.11	0.31	-2.96	0.94	0.04	0.33
-2.5	-2.61	0.61	-0.11	0.30	-2.43	0.71	0.07	0.27
-2	-2.04	0.49	-0.04	0.23	-1.96	0.58	0.04	0.27
-1	-1.01	0.42	-0.01	0.18	-1.04	0.46	-0.04	0.21
0	0.02	0.39	0.02	0.16	0.02	0.42	0.02	0.19
1	1.02	0.40	0.02	0.17	0.99	0.44	-0.01	0.18
2	2.09	0.50	0.10	0.27	2.02	0.58	0.02	0.24
2.5	2.64	0.62	0.14	0.28	2.52	0.77	0.02	0.30
3	3.11	0.71	0.11	0.34	2.95	0.94	-0.05	0.34

TABLE 4

Ability parameter recovery for empirical and standard solutions (T=20).

True θ	Empirical				Standard			
	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE
-3	-3.16	0.77	-0.16	0.38	-2.93	0.90	0.06	0.27
-2.5	-2.62	0.64	-0.12	0.28	-2.53	0.75	-0.03	0.30
-2	-2.07	0.52	-0.07	0.21	-2.04	0.60	-0.04	0.25
-1	-1.04	0.44	-0.04	0.18	-1.01	0.46	-0.01	0.19
0	0.03	0.39	0.03	0.15	0.02	0.41	0.02	0.16
1	1.01	0.41	0.01	0.18	0.97	0.44	-0.03	0.18
2	2.09	0.51	0.10	0.21	2.07	0.60	0.07	0.24
2.5	2.68	0.66	0.18	0.30	2.54	0.76	0.04	0.26
3	3.18	0.78	0.18	0.33	2.98	0.93	-0.02	0.28

TABLE 5

Ability parameter recovery for empirical, semi-empirical and standard solutions (T=5).

True θ	Empirical				Semi-empirical				Standard			
	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE	$\hat{\theta}$	s.d.	Bias	RMSE
-3	-3.09	0.57	-0.09	0.42	-2.40	0.77	0.60	0.71	-2.33	0.88	0.67	0.69
-2.5	-2.61	0.53	-0.11	0.37	-2.27	0.70	0.23	0.47	-2.21	0.82	0.29	0.39
-2	-2.06	0.46	-0.05	0.38	-1.77	0.58	0.23	0.48	-1.88	0.70	0.12	0.41
-1	-1.02	0.41	-0.02	0.28	-0.96	0.50	0.04	0.32	-0.91	0.51	0.09	0.30
0	-0.08	0.40	-0.08	0.31	0.00	0.47	0.01	0.29	0.04	0.46	0.04	0.33
1	1.04	0.40	0.04	0.27	0.93	0.50	-0.07	0.34	0.96	0.50	-0.04	0.35
2	2.07	0.47	0.07	0.34	1.93	0.62	-0.07	0.38	1.94	0.77	-0.06	0.43
2.5	2.58	0.53	0.08	0.34	2.16	0.71	-0.34	0.53	2.19	0.87	-0.31	0.36
3	3.13	0.56	0.13	0.44	2.40	0.77	-0.60	0.72	2.23	0.89	-0.77	0.78

TABLE 6

Estimated accuracy of simulation across different number of iterations.

N. iter	Burn-in	Empirical				Standard			
		$\hat{\theta}$	s.d.	5% s.d.	MC error	$\hat{\theta}$	s.d.	5% s.d.	MC error
1000	100	-0.389	0.386	0.019	0.022	0.113	0.510	0.025	0.030
2000	200	-0.001	0.379	0.019	0.013	0.254	0.482	0.024	0.023
5000	500	0.118	0.389	0.019	0.009	-0.111	0.445	0.022	0.010
10000	1000	0.078	0.390	0.019	0.006	-0.305	0.451	0.023	0.007

FIGURE 1

Root mean square error (RMSE) for the three different approaches (empirical, semi-empirical and standard) when the test consists of 5 items.

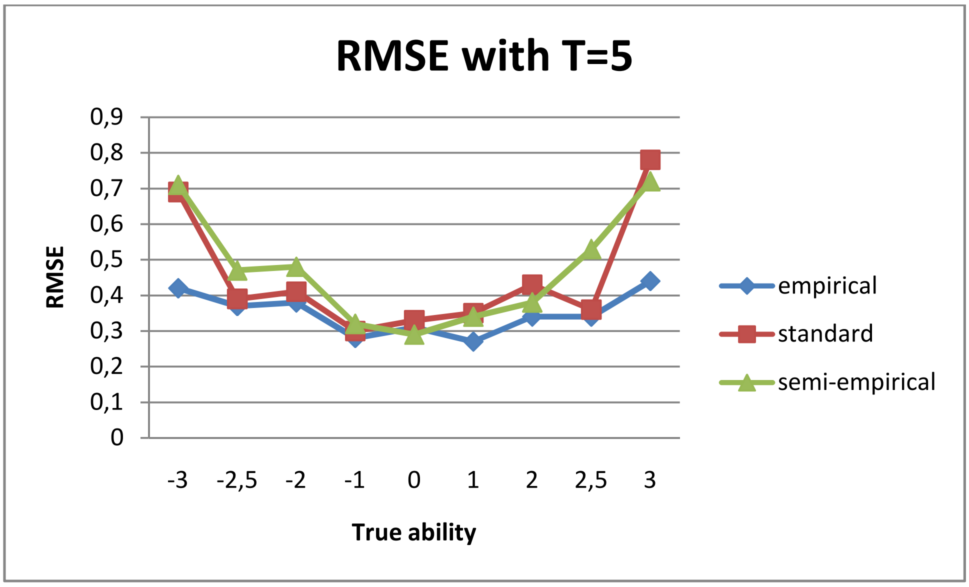


FIGURE 2

Trace plot of a single chain, in the case of T=5 and empirical information introduced.

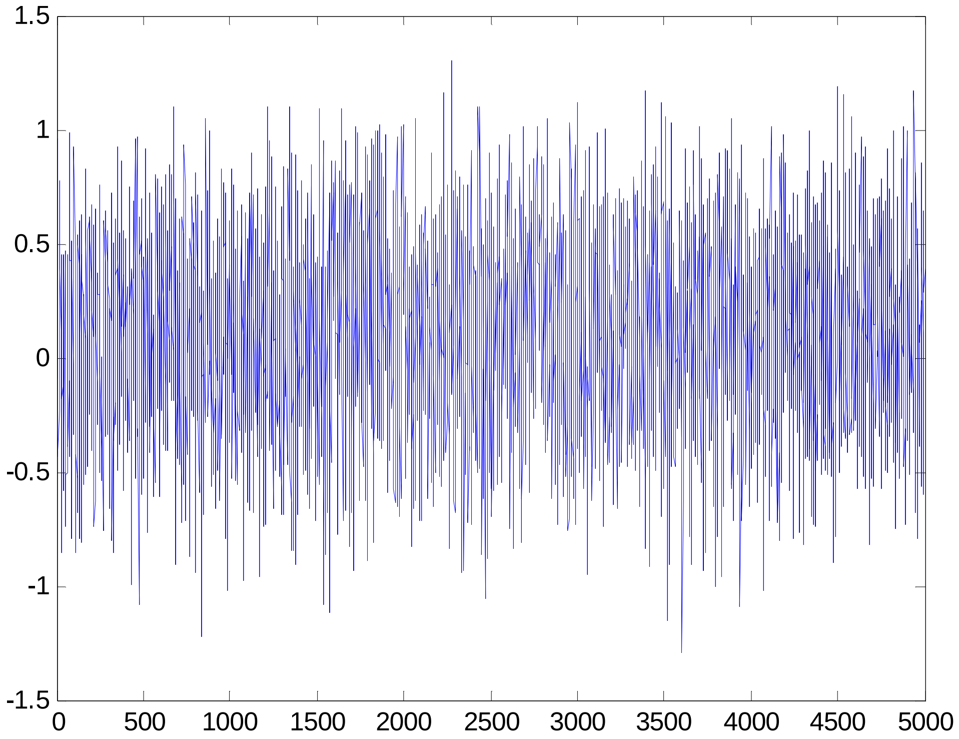
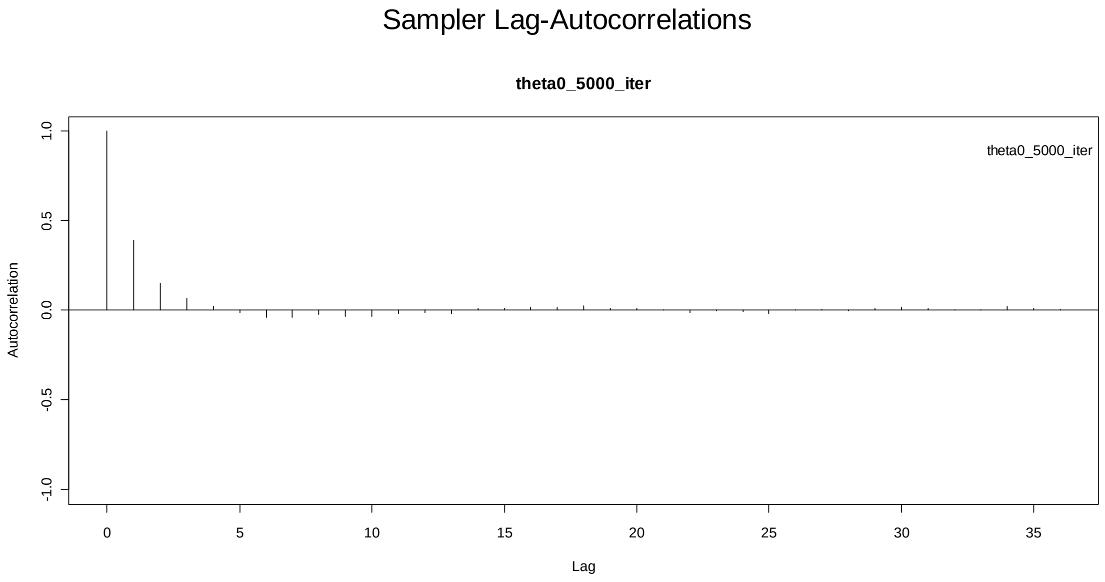


FIGURE 3

Autocorrelation plot.


Sampler Lag-Autocorrelations
theta0_5000_iter
theta0_5000_iter
Autocorrelation
Lag