**The Bayesian Approach and Suggested Stopping Criterion in Computerized Adaptive Testing**

Loc Nguyen

University of Science, Ho Chi Minh city, Vietnam

Email: ng\_phloc@yahoo.com

**Abstract**

Computer-based tests have more advantages than the traditional paper-based tests when there is a boom of internet and computer. Computer-based testing allows examinees to perform tests at any time and any place and testing environment becomes more realistic. Moreover, it is very easy to assess examinees’ ability by using computerized adaptive testing (CAT). CAT is considered as a branch of computer-based testing but it improves the accuracy of test core when CAT systems try to choose items (tests, exams, questions, etc.) which are suitable to examinees’ abilities; such items are called adaptive items.

The important problem in CAT is how to estimate examinees’ abilities so as to select the best items for examinees. There are some methods to solve this problem such as maximization likelihood estimation but I apply the Bayesian method into computing ability estimates. In this paper, I suggest a stopping criterion for CAT algorithm: the process of testing ends only when examinee’s knowledge becomes saturated (she/he can’t do test better or worse) and such knowledge is her/his actual knowledge.

**Keywords**: Bayesian inference, computerized adaptive test.

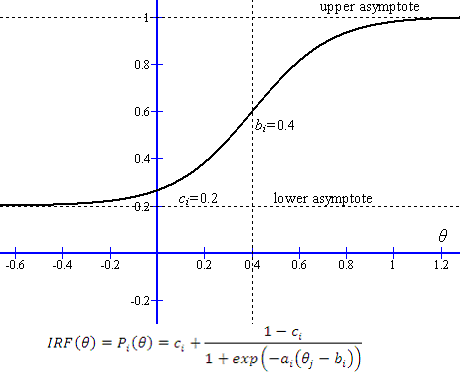
**1. Introduction**

**Item Response Theory** (IRT) is defined as a statistical model in which examinees can be described by a set of ability scores that are predictive. Based on mathematical models, IRT links together examinee’s performance on test items, item statistics, and examinee abilities (Rudner, 1998). Note that the term “*item*” indicates test, exam, question, etc. Users in IRT context are examinees. Examinee’s ability is often represented by variable *θ*. Given an examinee and item *i,* IRT is modeled as a function of a true ability *θ* of given examinee with three parameters of item *i* such as *ai, bi,* and *ci*. This function so-called Item Response Function (IRF) or Item Characteristic Curve (ICC) function computes the probability of a correct response of given examinee to item *i*. IRF is specified by equation 1 as follows:

|  |  |
| --- | --- |
|  | (1) |

Where *exp*(.) or *e*(.) denotes exponent function. Your attention please, IRF is function of examinee’s ability and it is essentially the probability of a correct response of a given examinee to an item. Suppose that *ai* is greater than 0.

IRF, a variant of logistic function, is plotted as the curve in following figure 1 with *ai=*6*, bi=*0.4*, ci=*0.2.



**Figure 1.** Item Response Function curve

The horizontal axis *θ* is the scale of examinee’s ability (Rudner, 1998). The vertical axis is the probability of correct response to the item specified by three parameters: *ai=*6*, bi=*0.4, *ci=*0.2. As seen in figure 1, the more the IRF shifts right, the more difficult item is. The lower asymptote at *ci=*0.2 indicates the probability of correct response for examinee with lowest ability and otherwise for the upper asymptote at 1.

IRF measures examinee’s proficiency based on her/his ability and some properties of item. Every item *i* has three parameters *ai*, *bi*, *ci* which are specified by experts or statistical data.

* The *ai* parameter called *discriminatory parameter* (Rudner, 1998) tells how well the item discriminates between examinees whose abilities are not different much. It defines the slope of the curve at the inflection point. The higher is the value of *ai*, the steeper is the curve. In case of steep curve, there is a large difference between the probabilities of a correct response for examinees whose ability is slightly below of the inflection point and examinees whose ability is slightly above the inflection point (Rudner, 1998).
* The *bi* parameter called *difficult parameter* (Rudner, 1998) indicates how difficult the item is. It specifies the location of inflection point of the curve along the *θ* axis (examinee’s ability). Higher value of *bi* shifts the curve to the right and implicates that the item is more difficult.
* The *ci* parameter called *guessing parameter* (Rudner, 1998) indicates that the probability of a correct response to item of low-ability examinees is very close to *ci*. It determines the lower asymptote of the curve. This parameter is called guessing parameter because it is the random probability that low-ability examinees guess a correct response to an item when they do not master the item. The upper asymptote always goes through value 1 because the probability that high-ability examinees give right response to an item is 1 (Rudner, 1998).

In general, IRF is used by computerized adaptive testing for choosing the best item which is given to examinee and estimating examinee’ true ability *θ*. Computerized adaptive testing is described right after.

**Computerized Adaptive Testing** (CAT) (Rudner, 1998) is the iterative algorithm which begins providing examinee an (test) item so as to be best to her/his initial ability; after that the ability is estimated again and the process of item suggestion is continued until stopping criterion is met. This algorithm aims to make a series of items which are evaluated to become chosen items that suitable to examinee’s ability. The set of items from which system picks ones up is called as item *pool*. The items chosen and given to examinee compose the adaptive test. CAT includes following steps (Rudner, 1998) as shown in table 1:

|  |
| --- |
| 1. The initial ability of examinee must be defined and items in the pool that have not yet been chosen are evaluated. The best one among these items is the most suitable to examinee’s current ability estimate. Such best item will be given to examinee in the step 2. IRF is applied into evaluating items. 2. The best item is chosen and given to examinee and the examinee responds. Such item is moved from pool to adaptive test. 3. A new ability estimate of examinee is computed based on responses to all of the chosen items. IRF is applied into computing the ability estimate. It is explained that ability estimate is the estimated value of true ability *θ* of examinee at current time point. 4. Steps 1 through 3 are repeated until stopping criterion is met. |

**Table 1.** Computerized adaptive testing (CAT) algorithm

Note that the chosen item is also called the *administered item* and the process of choosing best item is also called the *administration process*. The ability estimate is the value of *θ* which is fit best to the model and reflects current proficiency of examinee in item but it is not imperative to define precisely the initial ability because the final ability estimate may not be closed to initial ability. The stopping criterion could be time, number of administered items, change in ability estimate, maximum-information of ability estimate, content coverage, a precision indicator (standard error), a combination of factors, etc. (Rudner, 1998).

In step 1, there is the question: “how to evaluate the items so as to choose the best one”. So each item *i* is qualified by the amount of information at given ability *θ*; such *information function* is denoted *Ii*(*θ*). The best next item is the one that gets most informative or provides highest value of *Ii*(*θ*). Equation 2 specifies information function for item *i* (Rudner, 1998).

|  |  |
| --- | --- |
|  | (2) |

Where *Pi*(*θ*) is the probability of a correct response to item *i* and so it is the IRF function specified by previous equation 1 and is the first-order derivative of *Pi*(*θ*). According to equation 1, we have:

The information function *Ii*(*θ*) reflects how much the item *i* matches examinee’s ability. The item should not be too easy or too difficult. In the step 1 of CAT algorithm, the best item is the one that maximizes the information function *Ii*(*θ*). It is easy to find out such best item by brute-force technique that browses all items.

In step 3 of CAT algorithm, it is required to compute the ability estimate. The next section discusses how to find out the ability estimate with maximization likelihood estimation (MLE) method (Baker, 2001, pp. 86-90) and Bayesian method (Linden & Pashley, 2002, pp. 3-7).

**2. Estimating examinee’s ability**

Let be the ability estimate of examinee, the goal of this section is to calculate Recall that the ability estimate is very important to step 3 of CAT algorithm; please see table 1 for more details about CAT algorithm.

Suppose there are *N* items given to an examinee; in other words, the size of item pool is *N*. Each item *i* has *qi* optional responses. For example, we have *qi*=4 when item is question with four possible answers such as *A*, *B*, *C*, and *D*. We have *qi*=10 when item is an exam whose resulted grade ranges from 1 to 10. Suppose the number of *correct responses* of given examinee to item *i* is *ri*.

For example, a given examinee do exam *i* whose grade ranges from 1 to 10 and she/he gains grade 9 then, we have *qi*=10 and *ri*=9. Let *Pi*(*θ*) be the cumulative probability of a correct response to given item. Exactly, *Pi*(*θ*) is the probability that examinee’s ability is less than or equal to *θ* with regard to item *i*. Note that the probability *Pi*(*θ*) is IRF function specified by equation 1. For convenience, let guessing parameter be zero (*ci*=0). It means that the probability that examinee guesses correct response equals 0. Equation 3 specifies *Pi*(*θ*) and its derivative *Pi’*(*θ*) with *ci*=0.

|  |  |
| --- | --- |
|  | (3) |

Where *a* and *b* are discriminatory parameter and difficult parameter, respectively. Without loss of generality, equation 3 implicates that guessing parameter is fixed.

According to Bernoulli trial (Montgomery & Runger, 2003, p. 72), the probability that examinee provides *ri* correct responses for given item *i* is:

Where probability *Pi*(*θ*) is specified by equation 3.

The likelihood function (Czepiel, 2002, pp. 4-5) of examinee’s ability when she/he responses *N* items in the pool is specified by equation 4 as follows:

|  |  |
| --- | --- |
|  | (4) |

Note, *θ* becomes variable of the likelihood function *L*(*θ*). The notation denotes combination taken *ri* of *qi* elements and so we have .

It is required to estimate the ability of examinee *θ* so that the likelihood function takes maximum value. Let be the ability estimate of *θ*; of course, is the maximum value of likelihood function *L*(*θ*). Thus, this method is called maximum likelihood estimation (MLE) and the goal of MLE is to find out the ability estimate .

Because it is too difficult to work with the likelihood function in the form of product of condition probabilities, it is necessary to take logarithm of *L*(*θ*) so as to transform the likelihood function from repeated multiplication into repeated addition. The natural logarithm of *L*(*θ*) called log-likelihood function is denoted *LnL*(*θ*), according to equation 5 as follows:

|  |  |
| --- | --- |
|  | (5) |

Where *ln*(.) denotes natural logarithm function, *θ*0 is examinee’s initial ability and *ri* is examinee’s response. The notation denotes combination taken *ri* of *qi* elements and so we have .

Maximizing the likelihood function is equivalent to maximizing *LnL*(*θ*).

The maximization can be done by setting first-order partial derivatives of *LnL*(*θ*) with respect to *θ* to 0 and solving this equation to find out the ability estimates . The first-order derivative of *LnL*(*θ*) with respect to *θ* is:

Due to:

We have:

Setting this first-order derivative to 0, we have equation 6 for solving the estimate .

|  |  |
| --- | --- |
|  | (6) |

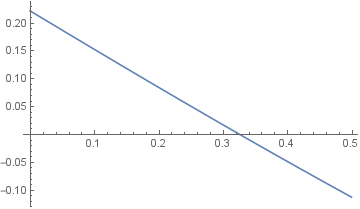
The Newton-Raphson method (Burden & Faires, 2011, pp. 67-69) is used to find solution of equation 6along with the tangent of *LnL’*(*θ*). It starts with an arbitrary value of *θ*0 as a solution candidate. Suppose the current value is *θk*, the next value *θk*+1 is calculated based on equation 7 (Baker, 2001, p. 87):

|  |  |
| --- | --- |
|  | (7) |

Where *LnL’’*(*θ*) is the second-order derivative of *LnL*(*θ*) with respect to *θ* as follows:

The value *θk* is solution of equation 6 if *LnL’*(*θk*) = 0 which means that *θk*+1=*θk*. In practice, *θk* is an acceptable solution if the absolute bias |*θk* – *θk*–1| is significantly small. For example, given three items (*a*1=1.0, *b*1=–1), (*a*2=1.2, *b*2=0), and (*a*3=0.8, *b*3=1), an examinee gives three respective responses *r*1=1, *r*2=0, and *r*3=1, respectively with suppose that all items are binary such that *q*1 = *q*2 = *q*3 = 1. This example is extracted from the book “The Basic of Item Response Theory” by author Frank B. Baker (Baker, 2001, pp. 88-90). Within this example, we have:

Figure 1 shows the curve *y* = *LnL’*(*θ*). The best ability estimate which is solution of equation 6 is intersection point of the curve *y* = *LnL’*(*θ*) and horizontal axis *y* = 0.



**Figure 1.** Curve equation *LnL’*(*θ*)=0

By applying Newton-Raphson method according to equation 7 with initial ability *θ*0=1, we get 4 estimates after 4 times to run such as , , , and . Because of , the best ability estimate is . The standard error of is 1.2296. The concept of standard error will be discussed in next section. Here we know that the smaller the standard error is, the more accurate the estimate is.

Ability *θ* has no prior distribution in MLE method. Thus, the initial ability *θ*0 for Newton-Raphson algorithm is set as arbitrary value, which causes that convergence of Newton-Raphson algorithm may be slowly. If *θ* has prior distribution π(*θ*), the initial ability *θ*0 will be set as a value which conforms the prior distribution π(*θ*), which can improve speed of convergence. Moreover, by taking advantages of such prior distribution we can produce more accurate estimate. Given *N* responses *r*1, *r*2,…, *rN*, the probability of such responses given ability *θ* according to Bernoulli trial (Montgomery & Runger, 2003, p. 72) is:

According Bayes’ rule (Wikipedia, 2017), the posterior distribution of *θ* with prior distribution π(*θ*) is:

The maximum a posteriori probability (MAP) method aims to determine an estimate that maximizes the posterior density function *f*(*r*1, *r*2,…, *r*N | *θ*). Please refer to Wikipedia website (Wikipedia, 2017) to know MAP method. In fact, MAP method is similar to MLE method except that MAP method follows Bayesian approach (Linden & Pashley, 2002, p. 6).

Because the marginal probability is positive and independent from *θ*, it is possible to remove such marginal probability from the expression of maximization as follows (Wikipedia, 2017):

Equation 8 expresses the MAP problem:

|  |  |
| --- | --- |
| Where, | (8) |

Thus, *g*(*θ*) is function of *θ*. The natural logarithm function of *g*(*θ*) is:

As a convention, *lg*(*θ*) is also called log-likelihood function for MAP method. Maximizing *g*(*θ*) is equivalent to maximizing *lg*(*θ*).

The maximization can be done by setting first-order partial derivatives of *lg*(*θ*) with respect to *θ* to 0 and solving this equation to find out the ability estimates . The first-order derivative of *lg*(*θ*) with respect to *θ* is:

As a convention, *ln*’(π(*θ*)) is the first-order derivative of *ln*(π(*θ*)).

Setting *lg’*(*θ*) to 0, we have equation 9 for solving the estimate .

|  |  |
| --- | --- |
|  | (9) |

The Newton-Raphson method (Burden & Faires, 2011, pp. 67-69) is used to find solution of equation 9. It starts with initial ability *θ*0 following the distribution π(*θ*). Suppose the current value is *θk*, the next value *θk*+1 is calculated based on equation 10:

|  |  |
| --- | --- |
|  | (10) |

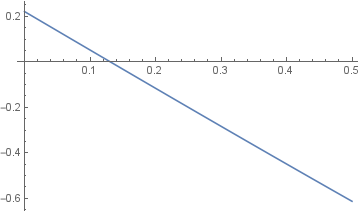
Where *lg’’*(*θ*) is the second-order derivative of *lg*(*θ*) with respect to *θ* as follows:

As a convention, *ln’’*(π(*θ*)) is the second-order derivative of *ln*(π(*θ*)).

The value *θk* is solution of equation 10 if *lg’*(*θk*) = 0 which means that *θk*+1=*θk*. In practice, *θk* is an acceptable solution if the absolute bias |*θk* – *θk*–1| is significantly small.

Going back the aforementioned example, given three items (*a*1=1.0, *b*1=–1), (*a*2=1.2, *b*2=0), and (*a*3=0.8, *b*3=1), an examinee gives three respective responses *r*1=1, *r*2=0, and *r*3=1, respectively with suppose that all items are binary such that *q*1 = *q*2 = *q*3 = 1. Suppose ability of examinee conforms standard normal distribution with mean *μ*=0 and variance *σ*2=1. This example is extracted from the book “The Basic of Item Response Theory” by author Frank B. Baker (Baker, 2001, pp. 88-90). Within this example, we have:

Figure 2 shows the curve equation *y* = *lg’*(*θ*). The best ability estimate which is solution of equation 9 is intersection point of the curve *y* = *lg’*(*θ*) and horizontal axis *y* = 0.



**Figure 2.** Curve equation *lg’*(*θ*)=0

By applying Newton-Raphson method according to equation 10 with initial ability *θ*0= *μ* = 0, we get 3 estimates after 3 times to run such as , , and . Because of , the best ability estimate is . The standard error of is 0.7705. The concept of standard error will be discussed in next section. Here we know that the smaller the standard error is, the more accurate the estimate is. By MAP method, the speed of convergence is faster with 3 run times and standard error is smaller because MAP takes advantages of prior distribution of *θ*.

**3. Suggested stopping criterion**

In normal the stopping criterion in step 4 of CAT algorithm is often the number of (test) items, for example, if the test has 10 items then the examinee’s final estimate is specified at 10th item and the test ends. This form is appropriate to examination in certain place and certain time and user is the examinee who passes or fails such examination.

Suppose in situation that user is the learner who wants to gains knowledge about some domain as much as possible and she/he does not care about passing or failing the examination. In other words, there is no test or examination and the learners prefer to study themselves by doing exercise. There is an exercise and items are questions that belong to this exercise. It is possible to use another stopping criterion in which the exercise ends only when the learner cannot do it better or worse. At that time her/his knowledge becomes saturated and such knowledge is her/his actual knowledge. The *ability error* is used to assess the saturation of learner’s knowledge. The ability error is difference between current ability estimate and previous examinee’s ability *θ*. Given threshold *ξ*, if the ability error is less than *ξ* then the CAT algorithm terminates; hence this is the new stopping criterion for CAT algorithm. Equation 11 specifies ability error denoted *Err*.

|  |  |
| --- | --- |
|  | (11) |

However, the ability error can be defined as standard error of ability estimate . Because *Pi*(*θ*) specified by equation 3 with *ci*=0 is a cumulative probability function, its derivative *Pi’*(*θ*) is a density probability function.

For ability *θ*, negative expectation of the second-order derivative of log-likelihood function is called information value of *θ* (Lynch, 2007, p. 40).

Because *LnL’*’(*θ*) does not depends on *ri*, we have:

The information value *I*(*θ*) conveys the amount of information at ability *θ* over all items. It is sum of *N* terms and each term is:

With binary item (*qi*=1), each term is actually the information function *Ii*(*θ*) of item *i* at given ability *θ* according to equation 3:

It implies

Given estimate which is resulted from MLE or MAP, the lower bound of variance of is inverse of information value *I*(*θ*) according to theorem of Cramer-Rao inequality (Zivot, 2009, p. 11):

If estimate is unbiased, variance of is equal to Cramer-Rao lower bound as follows (Zivot, 2009, p. 12):

Standard deviation of which is squared root of is called standard error of , which is denoted . Suppose is already determined, is calculated as follows:

The smaller the standard error is, the more accurate the estimate is. Therefore, is an important metric to evaluate accuracy of . Equation 12 specifies the standard error with regard to MLE and MAP.

|  |  |
| --- | --- |
|  | (12) |

Now the ability error can be defined as standard error .

In fact, if is small enough, the current ability of examinee represented by the estimate indicates her/his actual knowledge.

**4. Conclusion**

I recognized that CAT gives us the excellent tools for assessing examinee’s ability. The CAT algorithm includes four steps in which step 3 is most important when examinee’s ability estimate is determined. There is an advantage of Bayesian method (MAP method) when prior probability of examinee’s ability is used to enhance how to estimate examinee’s ability with high accuracy and high speed of convergence. However, quality of the posterior probability depends on the prior probability which may be pre-defined by experts. In the future trend, I intend to find out the technique for learning training data so as to specify precisely prior probability. Note that I only apply MLE method and MAP method into estimating examinee’s ability. Both methods are traditional and popular in statistical literature and so I does not invent or improve them in this research.

In general, I only propose the stopping criterion for CAT algorithm in which given threshold ξ, if ability error of examinee is less than ξ then the CAT algorithm stops. The goal of this technique is that the exercise ends only when the examinee can’t do it better or worse. It means that her/his knowledge becomes saturated and such knowledge is her/his actual knowledge. This method is only suitable to training exercises because there is no restriction for the number of (question) items in exercises. Conversely, in the formal test, the examinee must finish such test right before the decline time and the number of items in formal test is fixed. The idea of ability error is not actually new but I hope that it may be useful for researchers.

**References**

Baker, F. B. (2001). *The Basic of Item Response Theory* (2nd ed.). (C. Boston, & L. Rudner, Eds.) USA: ERIC Clearinghouse on Assessment and Evaluation. Retrieved from http://files.eric.ed.gov/fulltext/ED458219.pdf

Burden, R. L., & Faires, D. J. (2011). *Numerical Analysis* (9th Edition ed.). (M. Julet, Ed.) Brooks/Cole Cengage Learning.

Czepiel, S. A. (2002). *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation.* Czepiel's website http://czep.net.

Linden, W. J., & Pashley, P. J. (2002). Item Selection and Ability Estimation in Adaptive Testing. In W. J. Linden, G. A. Glas, W. J. Linden, & G. A. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (p. 323). Kluwer Academic Publishers.

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists.* Springer Berlin Heidelberg NewYork.

Montgomery, D. C., & Runger, G. C. (2003). *Applied Statistics and Probability for Engineers* (3rd Edition ed.). New York, NY, USA: John Wiley & Sons, Inc.

Rudner, L. M. (1998, November). *An On-line, Interactive, Computer Adaptive Testing Tutorial.* Retrieved 2009, from Lawrence M. Rudner's website: http://edres.org/scripts/cat

Wikipedia. (2017, March 2). *Maximum a posteriori estimation*. (Wikimedia Foundation) Retrieved April 15, 2017, from Wikipedia website: https://en.wikipedia.org/wiki/Maximum\_a\_posteriori\_estimation

Zivot, E. (2009). *Maximum Likelihood Estimation.* Lecture Notes on course "Econometric Theory I: Estimation and Inference (first quarter, second year PhD)", University of Washington, Seattle, Washington, USA.