Genetics and Molecular Biology, 27, 4, 679-685 (2004)
Copyright by the Brazilian Society of Genetics. Printed in Brazil
www.sbg.org.br

Research Article

Detection of determinant genes and diagnostic via Item Response Theory

Héliton Ribeiro Tavares[1], Dalton Francisco de Andrade[2] and Carlos Alberto de Bragança Pereira[3]

[1]*Universidade Federal do Pará, Departamento de Estatística, Belém, PA, Brazil.*
[2]*Universidade Federal de Santa Catarina, Departamento de Informática e Estatística, Florianópolis, SC, Brazil.*
[3]*Universidade de São Paulo, Departamento de Estatística, São Paulo, SP, Brazil.*

Abstract

This work presents a method to analyze characteristics of a set of genes that can have an influence in a certain anomaly, such as a particular type of cancer. A measure is proposed with the objective of diagnosing individuals regarding the anomaly under study and some characteristics of the genes are analyzed. Maximum likelihood equations for general and particular cases are presented.

Key words: Item Response Theory (IRT), predisposition, propensity, logistic model.

Received: September 22, 2003; Accepted: May 12, 2004.

Introduction

In many practical situations, decisions have to be taken based upon individual quantities that cannot be observed directly. These quantities are referred to by latent variables that are given different names according to the areas in which they are applied: *ability* or *proficiency* in educational and psychological areas; *purchasing power* in marketing; *life quality* or *predisposition* to a certain disease in the biological and medical areas (see Andrade *et al.*, (2000), Paas (1998), for example). These types of analysis are, in general, based upon the responses of a set of variables often referred to as *items* that comprise the measuring tool. In educational evaluations, for example, items are represented by questions in a test that might have their answers categorized as right/wrong, A/B/C/D/E with only one correct alternative or in a way where A is the least correct, and E is the most correct alternative. Other extensions are available, such as for each item a weight like 1 (right) or 0 (wrong) is attached. These types of study were, for sometime, based upon scores for each individual, that is, upon the number of items with weight one. However, this type of approach has many drawbacks mainly because it does not make a difference among the items which lead to the development of a theory based upon the items themselves and not upon the overall results, named Item Response Theory. In such a theory each item has a set of well defined characteristics that are estimated. The estimation procedure of the latent variable of an individual takes into account each one of the items of the test and reveals, for example, the level of knowledge of that individual in a certain area or his purchasing power as related to a certain product.

Some times there is more than one population being studied. For instance, in the educational area the interest can be the estimation of the average proficiencies regarding sex or geographical location.

In a similar situation, a set of genes is studied in order to appraise the predisposition of an individual related to a certain illness. A set of items (genes) are taken into account and their answers can be activated or deactivated or in the categorized form as A/B/C/D/E representing different levels of activity of the genes. Genes have peculiar characteristics that need to be incorporated into a model so that they can be evaluated. Suggestions have been advanced on the way to pinpoint genetic influences (Vanyukov and Tarder, 2000), but with some shortcomings. For example, the conclusions reached depend upon the sample chosen.

Models for Response Functions

The Item Response Theory is based upon models that represent the probability of response to an item as function of the parameters of the item and of the individual predisposition. These functions are treated as *Item Response Functions (IRF) or Item Charasteristic Curve (ICC)*. The different models proposed in the literature depend basically upon the type of item.

For explanatory reasons we will consider that there are *K* populations in study and each of them has the same *n*

Send correspondence to Heliton R. Tavares. Universidade Federal do Pará, Departamento de Estatística, 66075-900 Belém, Pará, Brazil. E-mail: heliton@ufpa.br.

genes being analyzed. The sample related to the population k is composed by N_k individuals, $k = 1,..., K$. Following, the model used in this paper is the unidimensional logistic model of 4 parameters for each item of two categories (of the type activated/deactivated). Its expression is given by

$$P(U_{ijk} = 1|\theta_{jk}, \varsigma_i) = c_i + (\gamma_i - c_i)\frac{1}{1 + e^{-Da_i(\theta_{jk} - b_i)}} \quad (1)$$

with $\zeta_i = (a_i, b_i, c_i, \gamma_i)'$, $i = 1,..., n$, $j = 1,..., N_k$ and $k = 1, 2,..., K$, where

U_{ijk} is a dichotomous variable that takes on the values 1,when the individual j of the population k has gene i activated, or 0 when the gene is deactivated.

θ_{jk} represents a predisposition of the jth individual of the population k.

b_i is the inactivity (or position) parameter of the gene i, measured at the same scale of the predisposition

a_i is the discrimination (or of inclination) parameter of the gene i

c_i is the probability of gene i being active for individuals with low predisposition,

γ_i is the probability of gene i being deactivated for individuals with high predisposition,

D is a scale factor, constant and equal to 1. The 1.7 value is used when it is desired that the logistic function yield results similar to that of the normal function.

N is the number of individuals involved in the study.

Defining the Parameters of the Genes

In a general way, the proposed model is based upon the fact that predisposed individuals are more likely to have the gene i activated, and that this relation is not linear. As a matter of fact, it can be perceived from Figure 1 that the IRF has the form of "*S*" with inclination and displacement defined by the gene parameters. However, only a subset of genes has to satisfy this situation that occurs only when $a_i > 0$. Chances are that some genes are deactivated in high propensity individuals, and therefore the IRF curve should have an inverted form, expressing that individuals with high propensity are less likely to get the gene activated, and this is expressed by $a_i < 0$. When $a_i = 0$, we have that $P(U_{ijk} = 1|\theta, \zeta_i) = (c_i + \gamma_i)/2$, constant for all θ, indicating that the gene i does not interfere in the occurrence of the anomaly.

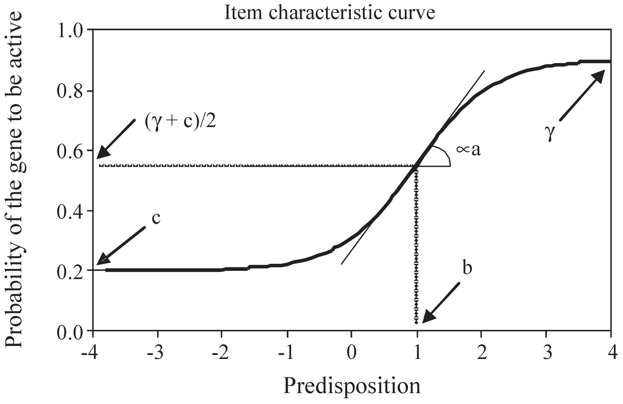

Figure 1 - Example of a graphic of an IRF

Parameter b_i is, perhaps, the most important of the four. The greater this parameter is, less likely it is that a given individual has the gene i activated. This is a valid conclusion only for $a_i > 0$, and the opposite is true for $a_i < 0$.

It is safe to say that individuals with low predisposition are prone to have the gene i active, and this information is conveyed by the parameter c_i. On the other hand, high predisposition individuals can also have the gene i inactive, and this information is conveyed by 1 - γ_i. These conclusions are valid only for $a_i > 0$, and the opposite is valid for $a_i < 0$.

Scale of Measurement/Indetermination

Predisposition can theoretically take any real value between -∞ e +∞. Thus, it is necessary to establish an origin and a unit of measurement for defining the scale. When only one population is under study the scale of measurement can be defined in such a way as to represent the mean value and the standard deviation of the individual predispositions of the population under study. For the graphs shown earlier the scale used had a mean of 0 and a standard deviation of 1, that will be referred from now on as scale (0,1). In practice, it does not make any difference to set these or any other values. What is paramount are the existent order relations between scale points. For example, in the scale used above an individual with a predisposition of 1.2 in fact is 1.2 standard deviations above the predisposition mean. This same individual would have a predisposition of 92, and therefore would also be 1,2 standard deviations above the predisposition mean, if the scale used for this population would have been the scale (80,10).

When various populations are present, one of them can be adopted as a *Reference Population*, and only the scale for this population will have to be refereed. The obtained predisposition values for other populations will have to be directly compared with those of the Reference Population. One such example consists of taking healthy individuals in the Reference Population and the population with a certain anomaly as the other. Other populations can be taken into account.

Local Independence

An often used hypothesis in IRT is the local independence (or conditional independence). It states that the probability that a certain gene is active depends only on its predisposition; that is, it offers all the necessary information to determine an activation/deactivation of the gene. In this fashion it does not mean that the quantities U_{kji} e U_{kjl}, $i \neq l$, are independent, but given the individual predisposition θ_{jk} they will be considered conditionally independent. However, there are models for the case when conditional independence is not met, but we have to model this possible dependence.

Parameter Estimation of the Genes and Predispositions

One of the most important stages of the IRT is the parameter estimation of the genes and/or of the individual predispositions. In some cases we can consider that the parameters of the genes are already known and what is wanted is to estimate the predispositions; in other, less common, predispositions of the individuals are known and what is wanted is the estimation of the parameter of the genes. However, *the most common cases are those in which not only the parameters of the genes are to be estimated but also the individual predisposition simultaneously*. In all these cases, the proposed model is assumed as true, and from the set of responses obtained for a certain number of individuals from one or more populations, parameters and/or predispositions are estimated using either likelihood or Bayesian methods. Both methods require iterative procedures involving very complex calculations and, therefore, specific computer codes. It is important to point out that, in any of these cases, the predisposition values and those of the gene parameters will all be in the same scale of measurement and therefore they can be compared.

Before outlining some points about the estimation process, some arrangements are in order. The set of genes involved in the analysis will be ordered in a fashion such that they will be represented by $\zeta = (\zeta_1,..., \zeta_n)$. Let $U_{kj.} = (U_{kj1}, U_{kj2},..., U_{kjn})$ be a random vector of answers from individual j from group k; $U_{k..} = (U_{k1.}, U_{k2.},..., U_{kn.})$ the random vector of answers from group k and $U_{...} = (U_{1..}, U_{2..},..., U_{k..})$ the whole vector of answers. In a similar fashion, observed answers will be represented by u_{kji}, $u_{kj.}$, $u_{k..}$ and $u_{...}$. This notation and local independence allow us to write the probability associated with the vector of answers U_{kj} as

$$P(u_{kj.}|\theta_{kj},\zeta) = \prod_{i \in I_k} P(u_{kji}|\theta_{kj},\zeta_\iota) \qquad (2)$$

Generally, it is considered that the predispositions of the individuals of population k, θ_{jk}, $j = 1,..., N_k$, are accomplishments of a random variable θ_k, with continuous distribution and probability density function $g(\theta|\eta_k)$, twice differentiable, with the components of η_k finite. In the case where θ_k has a Normal distribution, we have $\eta_k = (\mu_k, \sigma_k^2)$, where μ_k is the mean and σ_k^2 the variance of the predispositions of the individuals of the population k, $k = 1,..., K$. This hypothesis carries a great advantage: only the parameters of the genes have to be estimated, as the likelihood will not depend on the individuals' predispositions. Therefore, the estimation is a two-stage process, where in the first only the parameters of the genes are estimated, after which these parameters are considered as known for the estimation of the predispositions.

Estimation of Gene Parameters

With the above defined notations we have determined that the marginal probability of U_{kj} is given by

$$P(u_{kj.}|\varsigma,\eta_k) = \int_{\Re} P(u_{kj.}|\theta,\varsigma,\eta_k) g(\theta|\eta_k) d\theta =$$

$$\int_{\Re} P(u_{kj.}|\theta,\zeta) g(\theta|\eta_k) d\theta$$

where in the last inequality we use that the distribution of $U_{kj.}$ is not a function of parameters η_k.

Utilizing the independence between answers of different individuals, we can see that the associated probabilities to the vector of answers $U_{...}$ as

$$P(u_{...}|\zeta,\eta_k) = \prod_{k=1}^{K} \prod_{j=1}^{N_k} P(u_{kj.}|\zeta,\eta_k)$$

Even though the likelihood can be written as (2), the approach has often been used of *Response Patterns*. As we have n genes, with two possible answers for each item (0 or 1), there are $S = 2^n$ possible response vectors (response patterns). Let r_{kj} be the number of distinct occurrences of the answer pattern j in group k, and yet $S_k \leq \min(N_k, S)$ the number of response patterns with $r_{kj} > 0$. It follows that

$$\sum_{j=1}^{S_k} r_{kj} = N_k$$

By the independence between the answers of different individuals, we have that the data follows a *Product-Multinomial* distribution, that is,

$$L(\zeta,\eta) = \prod_{k=1}^{K} \left\{ \frac{N_k!}{\prod_{j=1}^{S_k} r_{jk}!} \prod_{j=1}^{S_k} \left[P\left(u_{kj.}|\zeta,\eta_k\right) \right]^{r_{jk}} \right\}$$

And, therefore, the log-likelihood is

$$\log L(\zeta,\eta) = \sum_{k=1}^{K} \log \left\{ \frac{N_k!}{\prod_{j=1}^{S_k} r_{jk}!} \right\} + \sum_{k=1}^{K} \sum_{j=1}^{S_k} r_{jk} \log P\left(u_{kj.}|\zeta,\eta_k\right) \qquad (3)$$

The estimation equations for the item parameters are given by

$$\frac{\partial \log L(\zeta,\eta)}{\partial \zeta_i} = 0, \quad i = 1,...,n,$$

with

$$\frac{\partial \log L(\zeta,\eta)}{\partial \zeta_i} = \frac{\partial}{\partial \zeta_i}\left\{\sum_{k=1}^{K}\sum_{j=1}^{S_k} r_{jk}\ \log P\left(u_{kj.}\middle|\zeta,\eta_k\right)\right\} =$$

$$\sum_{k=1}^{K}\sum_{j=1}^{S_k} r_{jk}\ \frac{1}{P\left(u_{kj.}\middle|\zeta,\eta_k\right)}\ \frac{\partial P\left(u_{kj.}\middle|\zeta,\eta_k\right)}{\partial \zeta_i} =$$

$$\sum_{k=1}^{K}\sum_{j=1}^{S_k} r_{jk}\int_{\Re}\left[\frac{(u_{kji}-P_i)}{P_iQ_i}\left(\frac{\partial P}{\partial \zeta_i}\right)\right] g^*_{kj}(\theta)d\theta,$$

where

$$g^*_{kj}(\theta) \equiv g(\theta|u_{kj.},\zeta,\eta_k) = \frac{P(u_{kj.}|\theta,\zeta)g(\theta|\eta_k)}{P(u_{kj.}|\zeta,\eta_k)}$$

and P_i represents the IRF adopted. The specific equations for each parameter of the vector $\zeta_i = (a_i, b_i, c_i, \gamma_i)$' can thus be obtained from above.

Application to the 4-parameter Logistic Model

For convenience, let

$$W_i = \frac{P^*_i Q^*_i}{P_i Q_i} \tag{4}$$

where $P^*_i = \left\{1 + e^{-Da_i(\theta - b_i)}\right\}^{-1}$ and $Q_i^* = 1 - P_i^*$.

In sum, the estimation equations for a_i, b_i, c_i and γ_i are, respectively,

$$D(1-c_i)\sum_{k=1}^{K}\sum_{j=1}^{S_k} r_{kj}\int_{\Re}(u_{kji}-P_i)(\theta-b_i)W_i g^*_{kj}(\theta)d\theta = 0,$$

$$-Da_i(1-c_i)\sum_{k=1}^{K}\sum_{j=1}^{S_k} r_{kj}\int_{\Re}(u_{kji}-P_i)W_i g^*_{kj}(\theta)d\theta = 0,$$

$$\sum_{k=1}^{K}\sum_{j=1}^{S_k} r_{kj}\int_{\Re}(u_{kji}-P_i)\frac{W_i}{P^*_i} g^*_{kj}(\theta)d\theta = 0,$$

$$\sum_{k=1}^{K}\sum_{j=1}^{S_k} r_{kj}\int_{\Re}(u_{kji}-P_i)\frac{W_i}{Q^*_i} g^*_{kj}(\theta)d\theta = 0,$$

which do not have explicit solutions. Therefore, such estimations are arrived at by iterative processes, such as Newton-Raphson, BFGS, Fisher's Scoring or EM algorithm.

Estimation of the Population Parameters

Considering the log-likelihood obtained in (3), the estimation equations for the mean predispositions and population variances are obtained by

$$\frac{\partial \log L(\zeta,\eta)}{\partial \mu_k} = 0$$

and

$$\frac{\partial \log L(\zeta,\eta)}{\partial \sigma^2_k} = 0$$

$k = 1,..., K$. However,

$$\frac{\partial \log L(\zeta,\eta)}{\partial \eta_k} = \sum_{j=1}^{S_k} r_{kj}\int_{\Re}\left(\frac{\partial \log g(\theta|\eta_k)}{\partial \eta_k}\right) g^*_{kj}(\theta)d\theta.$$

If we use the distribution $N(\mu_k, \sigma_k^2)$ for θ_k, we have

$$\frac{\partial \log g(\theta|\eta_k)}{\partial \mu_k} = \frac{\theta-\mu_k}{\sigma^2_k}$$

and

$$\frac{\partial \log g(\theta|\eta_k)}{\partial \sigma^2_k} = \frac{\sigma^2_k-(\theta-\mu_k)^2}{2\sigma^4_k}$$

Thus, the final forms of the estimation equations for μ_k and σ_k^2 are, respectively,

$$\left(\sigma^2_k\right)^{-1}\sum_{j=1}^{S_k} r_{kj}\int_{\Re}(\theta-\mu_k)g^*_{kj}(\theta)d\theta = 0,$$

$$-\left(2\sigma^4_k\right)^{-1}\sum_{j=1}^{S_k} r_{kj}\int_{\Re}\left[\sigma^2_k-\left(\theta-\mu_k\right)^2\right]g^*_{kj}(\theta)d\theta = 0.$$

Estimation of the Predispositions

Once the parameters of the genes are set, individual predispositions can be estimated. In addition, such predispositions can also be estimated for individuals whose data were not considered in the item parameters estimation. The usual methods for estimating the predispositions are the maximum likelihood (ML) as well as Bayesian methods such as *maximum a posteriori (MAP)* and the *expected a posteriori (EAP)*.

Estimation by ML

In this case, the estimation of the predispositions is done iteratively by the Newton-Raphson algorithm maximizing the likelihood in (2), or of the equivalent form, the function

$$\log L(\theta) = \sum_{k=1}^{K}\sum_{j=1}^{S_k}\sum_{i=1}^{n}\left\{u_{kji}\log P_{kji} + \left(1-u_{kji}\right)\log Q_{kji}\right\}$$

The Maximum Likelihood Estimator (MLE) of θ_{kj} is that which maximizes the likelihood, or equivalently, is the solution of the equation

$$\frac{\partial \log L(\theta)}{\partial \theta_{kj}} = 0,\ j = 1,..., N_k, k = 1,..., K. \tag{5}$$

Note, from (5), that

$$\frac{\partial \log L(\theta)}{\partial \theta_{kj}} = \sum_{i=1}^{n} \left\{ u_{kji} \frac{\partial (\log P_{kji})}{\partial \theta_{kj}} + (1 - u_{kji}) \frac{\partial (\log Q_{kji})}{\partial \theta_{kj}} \right\} =$$

$$\sum_{i=1}^{n} \left\{ u_{kji} \frac{1}{P_{kji}} - (1 - u_{kji}) \frac{1}{Q_{kji}} \right\} \left(\frac{\partial P_{kji}}{\partial \theta_{kj}} \right) = \qquad (6)$$

$$\sum_{i=1}^{n} \left\{ \frac{u_{kji} - P_{kji}}{P_{kji} Q_{kji}} \right\} \left(\frac{\partial P_{kji}}{\partial \theta_{kj}} \right) = \sum_{i=1}^{n} \left\{ (u_{kji} - P_{kji}) \frac{W_{kji}}{P^*_{kji} Q^*_{kji}} \right\} \left(\frac{\partial P_{kji}}{\partial \theta_{kj}} \right)$$

where the last equality follows by (4), and when plugged in the respective quantities. As

$$\frac{\partial P_{kji}}{\partial \theta_{kj}} = D a_i (1 - c_i) P^*_{kji} Q^*_{kji} ,$$

it is obtained

$$h(\theta_{kj}) \equiv \frac{\partial \log L(\theta)}{\partial \theta_{kj}} = D \sum_{i=1}^{n} a_i (1 - c_i)(u_{kji} - P_{kji}) W_{kji}$$

It follows then that the estimation Eq. (5) for θ_{kj}, $j = 1,..., N_k$, is

$$D \sum_{i=1}^{n} a_i (1 - c_i)(u_{kji} - P_{kji}) W_{kji} = 0.$$

Again, this equation does not have an explicit solution for θ_{kj} and, for this reason it is necessary for some iterative method in order to obtain the desired estimation. Following, the necessary expressions are obtained for applications of the Newton-Raphson iterative processes.

Considering $\hat{\theta}^{(t)}_{kj}$ an estimation of θ_{kj} in iteration t, then, in iteration $t + 1$ we have

$$\hat{\theta}^{(t+1)}_{kj} = \hat{\theta}^{(t)}_{kj} - \left[H\left(\hat{\theta}^{(t)}_{kj} \right) \right]^{-1} h\left(\hat{\theta}^{(t)}_{kj} \right),$$

where

$$H\left(\theta_{kj}\right) = \sum_{i=1}^{n} (u_{kji} - P_{kji}) W_{kji} \times \left[H_{kji} - (u_{kji} - P_{kji}) W_{kji} h^2_{kji} \right]$$

with

$$h_{kji} = \left(P^*_{kji} Q^*_{kji} \right)^{-1} \left(\frac{\partial P_{kji}}{\partial \theta_{kj}} \right) = D a_i (1 - c_i) \qquad (7)$$

and

$$H_{kji} = \left(P^*_{kji} Q^*_{kji} \right)^{-1} \left(\frac{\partial^2 P_{kji}}{\partial \theta^2_{kj}} \right) = D^2 a^2_i (1 - c_i)(1 - P^*_{kji}) \qquad (8)$$

Estimation by MAP

Such as in the marginal likelihood estimation, the Bayesian estimation of the predispositions is done on the second stage, considering the fixed parameters of the genes. Through the hypothesis of independence between the predispositions of different individuals, estimations can be done separately for each individual.

Let us assume that the prior distribution for θ_{kj}, $j = 1,..., N_k$, is Normal with known vector $\eta_k = (\mu_k, \sigma^2_k)$ of parameters. The posterior distribution for the ability of the individual j of the population k can be written as

$$g^*_{kj}(\theta_{kj}) \equiv g(\theta_{kj} | u_{kj.}, \varsigma, \eta) \propto P(u_{kj.} | \theta_{kj}, \varsigma) g(\theta_{kj} | \eta_k) \qquad (9)$$

Some characteristic of $g^*_{kj}(\theta_{kj})$ can be adopted as estimator of θ_{kj}, where the most frequently adopted are the mean and the mode. Following, we deal with how to obtain each of these characteristics.

Estimation of the mode of the posterior distribution - MAP

The Bayesian modal estimation consists in obtaining the maximum of (9). For easiness, we work with the logarithm of the posteriori

$$\log g^*_{kj}(\theta_{kj}) = C + \log P(u_{kj.} | \theta_{kj}, \varsigma) + \log g(\theta_{kj} | \eta_k),$$

where C is a constant. It follows that the estimation equation for θ_{kj} is

$$\frac{\log g^*_{kj}(\theta_{kj})}{\partial \theta_{kj}} = \frac{\partial \log P(u_{kj.} | \theta_{kj}, \varsigma)}{\partial \theta_{kj}} + \qquad (10)$$

$$\frac{\partial \log g(\theta_{kj} | \eta_k)}{\partial \theta_{kj}} = 0$$

By local independence, we have that

$$\log P(u_{kj.} | \theta_{kj}, \varsigma) = \log \left[\prod_{i=1}^{n} P(u_{kji} | \varsigma_i, \theta_{kj}) \right] =$$

$$\sum_{i=1}^{n} \log P(u_{kji} | \varsigma_i, \theta_{kj}).$$

Therefore,

$$\frac{\partial \log P(u_{kj.} | \theta_{kj}, \varsigma)}{\partial \theta_{kj}} = \sum_{i=1}^{n} \frac{\partial \log P(u_{kji} | \varsigma_i, \theta_{kj})}{\partial \theta_{kj}}$$

$$= \sum_{i=1}^{n} \frac{\partial P(u_{kji} | \varsigma_i, \theta_{kj})}{P(u_{kji} | \varsigma_i, \theta_{kj})}.$$

Keeping in mind that $P(u_{kji} | \varsigma_i, \theta_{kj}) = P^{u_{kji}}_{kji} Q^{1-u_{kji}}_{kji}$ and using the development under (5), we have that

$$\frac{\partial \log P(u_{kj.} | \theta_{kj}, \varsigma)}{\partial \theta_{kj}} = D \sum_{i=1}^{n} a_i (1 - c_i)(u_{kji} - P_{kji}) W_{kji} \qquad (11)$$

As we have adopted the prior distribution Normal (μ_k, σ_k^2) for θ_{kj}, the second portion of (10) is

$$\frac{\partial \log g(\theta_{kj}|\eta_k)}{\partial \theta_{kj}} = -\frac{(\theta_{kj} - \mu_k)}{\sigma_k^2}$$

By (11) and (12), we have that the estimation equation for θ_{kj} is

$$D\sum a_i(1-c_i)(u_{kji} - P_{kji})W_{kji} - \frac{(\theta_{kj} - \mu_k)}{\sigma_k^2} = 0.$$

As this equation does not have an explicit solution, some iterative method can be used to solve it. To do that it is necessary the second derivative of $\log g_{kj}^*(\theta_{kj})$ with relation to θ_{kj}, whose expression is

$$H(\theta_{kj}) = \sum_{i=1}^{n} (u_{kji} - P_{kji})W_{kji} \times$$

$$\left[H_{kji} - (u_{kji} - P_{kji})W_{kji}h_{kji}^2\right] - \frac{1}{\sigma_k^2},$$

where h_{kji} and H_{kji} are given by (7) and (8), respectively.

Estimation by EAP

The Bayes expected a posteriori (EAP) consists in obtaining the expectation of the posterior distribution, that can be written as

$$g(\theta|u_{kj.}, \varsigma, \eta_k) = \frac{P(u_{kj.}|\theta, \varsigma)g(\theta|\eta_k)}{P(u_{kj.}|\varsigma, \eta_k)}$$

It follows that the estimator is given by

$$\hat{\theta}_{kj} \equiv E\left[\theta|u_{kj.}, \varsigma, \eta_k\right] = \frac{\int_{\Re} \theta g(\theta|\eta_k)P(u_{kj.}|\theta, \varsigma)d\theta}{\int_{\Re} g(\theta|\eta_k)P(u_{kj.}|\theta, \varsigma)d\theta}$$

This form of estimation has the advantage of being calculated directly, not being necessary the application of iterative methods.

Simulation Results

In this section we present one application of the proposed methodology in simulated data. The data were generated based on N = 5000 individuals and to $n = 5$ genes. The total simulation consisted of 1000 replications. The known gene parameters are presented below. All the calculations were done via a computer program developed by the authors using the computer language *Ox* (see Doornik, 1998) using the BFGS routine for maximization.

The Genes Parameters

In order to generate the data it was assumed that the genes parameters are those presented in Table 1. It was adopted the 4 parameter logistic model with D = 1.7. The values for parameter *a* (discrimination) varied from 0.8 (low discrimination) to 1.2 (high discrimination) and the values for parameters *b* (predisposition) varied from -0.5 to 3.0. For the parameters *c* it was considered only one value (0.20) and for the γ, only 0.9. It was adopted the 4 parameter logistic model with $D = 1.7$.

Table 1 - Genes parameters.

	1	2	3	4	5
a_i	0.8	0.8	1.0	1.2	1.2
b_i	-0.5	1.0	1.5	0.5	0.0
c_i	0.2	0.2	0.2	0.2	0.2
γ_i	0.9	0.9	0.9	0.9	0.9

From Table 2 we see that the average estimates obtained from 1000 replicates are very accurate for all genes. We see that the estimations procedure works very well, still when we have a relatively small number of genes. Results were obtained with a larger number of genes and the results were still very good. However, we hope that estimation problems just appear when the number of genes is too small.

The Table 3 presents the standard deviations obtained from 1000 estimates. The largest values are associated with the parameters *a* and *b*. With exception of the gene 2, the values associated with parameter *a* are larger than those associated to *b*.

Concluding remarks

We have introduced a new proposal for genes and person diagnostic via Item Response Models. From a simulation study, it was shown that the models provide good estimates for several genes configurations. However, other studies and models should be proposed to allow, for example, different levels of activities of the genes. Longitudinal models, following the lines of Tavares and Andrade (2004) and Andrade and Tavares (2004), should also be considered.

Table 2 - Deviations from the average estimates with relation to the true gene parameters.

	1	2	3	4	5
a_i	0.0193	0.0333	0.0233	0.0254	0.0117
b_i	0.0021	-0.0103	0.0472	-0.0062	0.0038
c_i	-0.0038	-0.0038	-0.0036	0.0001	-0.0006
γ_i	0.0042	0.0031	0.0195	-0.0031	0.0020

Table 3 - Standard deviations for the 1000 estimates.

	1	2	3	4	5
a_i	0.1576	0.1738	0.2562	0.1293	0.1164
b_i	0.1080	0.2055	0.2019	0.0934	0.0804
c_i	0.0346	0.0377	0.0244	0.0242	0.0232
γ_i	0.0411	0.0726	0.0798	0.0336	0.0289

Acknowledgments

This work was partially supported by grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenadoria para o Aperfeiçoamento de Pessoal de Nível Superior (CAPES), National Program of Excellency (PRONEX) contract No 76.97.1081.00, Thematic Project of Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) contract No 96/01741-7, Integrated Program for Teaching, Research and Extension (PROINT) of the UFPA, and PARD. It was partially developed when the first author was a doctoral student at the Department of Statistics of the University of São Paulo.

References

Andrade DF and Tavares HR (2004) Item response theory for longitudinal data: Population parameter estimation. To appear in Journal of Multivariate Analysis.

Andrade DF, Tavares HR and Valle RC (2000) Item Response Theory: Concepts and applications. Associação Brasileira de Estatística, São Paulo (in Portuguese).

Bock RD and Zimowski MF (1997) Multiple Group IRT. In: van der Linder WJ and Hambleton RK (eds) Handbook of Modern Item Response Theory. Spring-Verlag, New York.

Chow YS and Teicher H (1978) Probability Theory: Independence, Interchangeability, Martingales. Springer-Verlag, New York.

Doornik JA (1998) Object-Oriented Matrix Programming using Ox 2.0. Timberlake Consultants Ltd and Oxford, London. www.nuff.ox.ac.uk/Users/Doornik.

Hambleton RK, Swaminathan H and Rogers HJ (1991) Fundamentals of Item Response Theory. Sage Publications, Newburg Park.

Lord FM (1980) Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, Inc., Hillsdale.

Paas LJ (1998) Mokken scaling characteristic sets and acquisition patterns of durable and financial products. Journal of Economic Psychology 19:353-376.

Tavares HR and Andrade DF (2004) Item response theory for longitudinal data: Item and population ability parameters estimation (to appear in Test).

Sanathanan L and Blumenthal N (1978) The logistic model and estimation of latent structure. Journal of the American Statistical Association 73:794-798.