Università degli Studi di Padova

Padua Research Archive - Institutional Repository

Recovering a probabilistic knowledge structure by constraining its parameter space

Original Citation:

Availability:
This version is available at: 11577/2452678 since:

Publisher:

Published version:
DOI: 10.1007/s11336-008-9095-7

Terms of use:
Open Access

This article is made available under terms and conditions applicable to Open Access Guidelines, as described at http://www.unipd.it/download/file/fid/55401 (Italian only)

(Article begins on next page)

PSYCHOMETRIKA—VOL. 74, NO. 1, 83–96
MARCH 2009
DOI: 10.1007/S11336-008-9095-7

RECOVERING A PROBABILISTIC KNOWLEDGE STRUCTURE BY CONSTRAINING ITS PARAMETER SPACE

LUCA STEFANUTTI AND EGIDIO ROBUSTO

UNIVERSITY OF PADUA

In the Basic Local Independence Model (BLIM) of Doignon and Falmagne (Knowledge Spaces, Springer, Berlin, 1999), the probabilistic relationship between the latent knowledge states and the observable response patterns is established by the introduction of a pair of parameters for each of the problems: a lucky guess probability and a careless error probability. In estimating the parameters of the BLIM with an empirical data set, it is desirable that such probabilities remain reasonably small. A special case of the BLIM is proposed where the parameter space of such probabilities is constrained. A simulation study shows that the constrained BLIM is more effective than the unconstrained one, in recovering a probabilistic knowledge structure.

Key words: probabilistic knowledge structures, basic local independence model, constrained parameter estimation.

1. Introduction

Given a collection Q of problems (or items) in some domain, in knowledge space theory (Albert, 1994; Albert & Lukas, 1999; Doignon & Falmagne, 1985, 1999; Falmagne, Doignon, Koppen, Villano, & Johannesen, 1990), the knowledge state of a student is the collection $K \subseteq Q$ of all problems that this student is capable of solving. A knowledge structure is a pair $(Q, \mathcal{K})$, where $\mathcal{K}$ is a collection of knowledge states, which contains at least the empty set and Q. Typically, not all subsets of the full set Q are knowledge states. Assumptions on a dependence relation among the items are usually made and these assumptions are restrictions that determine which subsets of Q are states and which are not (see, e.g., Falmagne et al., 1990). Only those subsets that are consistent with such assumptions belong to the knowledge structure $\mathcal{K}$.

A probabilistic knowledge structure (PKS) is a knowledge structure $(Q, \mathcal{K})$ equipped with a probability distribution π on the knowledge states (Falmagne & Doignon, 1988). It is essentially an unrestricted latent class model where the latent classes are the knowledge states. The probabilistic relationship between the latent knowledge states and the observable response patterns is established by the introduction of a pair of parameters for each of the dichotomous items: a lucky guess probability and a careless error probability.

In estimating a PKS with an empirical data set, it is desirable that such probabilities remain reasonably small. A lucky guess or a careless error equal or greater than, say, 0.5 would be rather difficult to interpret. A possible interpretation is that the data are highly noisy: 50% of the students who master a given problem fail it by careless error. Conversely, 50% of those students that are not capable of solving a given problem solve it by chance. In such a situation, the data set would be classified as too noisy, and thus discarded.

There is however another interpretation which is related to the fit of the model. Suppose the data have been generated by some "true" but unknown knowledge structure $\mathcal{K}$ with small lucky guess and careless error probabilities. What happens if an incorrect model $\mathcal{K}' \neq \mathcal{K}$ is fitted to these data? The simulation study described in Section 3 shows that an incorrectly specified

Requests for reprints should be sent to Luca Stefanutti, Dipartimento di Psicologia Generale, University of Padua, Padua, Italy. E-mail: luca.stefanutti@unipd.it

© 2008 The Psychometric Society

model has good chances to obtain an acceptable or even a good fit by an ad hoc inflation of the careless error and lucky guess probabilities. In a comparative sense, the correct model $\mathcal{K}$ and the incorrect model $\mathcal{K}'$ will not differ that much in terms of likelihood.

A simple and straightforward way to avoid such an inflation is to constrain the lucky guess and careless error probability estimates within a certain, reasonably small, interval. A constrained version of the BLIM is developed (C-BLIM), maximum likelihood parameter estimation for the C-BLIM is derived and the model is explored through a series of simulation studies.

The rest of this section is an overview of the basic local independence model (BLIM) for probabilistic knowledge structures (Falmagne & Doignon, 1988). Maximum likelihood estimation for the parameters of the C-BLIM is introduced in Section 2. Section 3 presents a simulation study in which the implications of the C-BLIM are explored and discussed.

1.1. The Basic Local Independence Model (BLIM)

This section is a brief overview of a probabilistic model for knowledge structures proposed by Falmagne and Doignon (1988) in the context of efficient knowledge assessment. The relationships between this model and other similar models in the literature on cognitive diagnosis are briefly discussed at the end of the section. Let Q be a nonempty finite set containing n distinct dichotomous items, and $\mathcal{K}$ be a knowledge structure on Q. Both Q and $\mathcal{K}$ are fixed throughout the section.

The binary response pattern (collection of binary responses to the items in Q) of a student randomly sampled from the population is represented by a discrete random variable $\mathbf{R}$ whose realizations are vectors $\mathbf{r} \in \{0, 1\}^n$. The kth element of $\mathbf{r}$ is equal to one if the student's response to the kth item on an n-item knowledge assessment exam is correct and the kth element of $\mathbf{r}$ is equal to zero otherwise ($k = 1, \ldots, n$). The unknown knowledge state of this student is represented by a discrete random variable $\mathbf{K}$ whose realizations are elements $K \in \mathcal{K}$. The probability of sampling a student whose response pattern is $\mathbf{r}$ is denoted by $P(\mathbf{R} = \mathbf{r})$, and the probability that the knowledge state of this student is $K \in \mathcal{K}$ is denoted by $P(\mathbf{K} = K)$.

The connection between the observable response patterns and the unobservable knowledge states is given in the BLIM by the following unrestricted latent class model (see, e.g., Goodman, 1974; Haberman, 1979)

$$P(\mathbf{R} = \mathbf{r}) = \sum_{K \in \mathcal{K}} P(\mathbf{R} = \mathbf{r} | \mathbf{K} = K) P(\mathbf{K} = K), \tag{1}$$

where $P(\mathbf{R} = \mathbf{r}|\mathbf{K} = K)$ is the conditional probability that the response pattern of a randomly sampled student is $\mathbf{r} \in \{0, 1\}^n$ given that his knowledge state is $K \in \mathcal{K}$.

The BLIM is then characterized by three types of parameters: a parameter π_K specifying the probability $P(\mathbf{K} = K)$ of each knowledge state, a careless error parameter α_q and a lucky guess parameter β_q for every item $q \in Q$. The parameter α_q is interpreted as the probability $P(\mathbf{R}_q = 0 | q \in \mathbf{K})$ that a student will fail q given that this item is indeed solvable from his knowledge state. The parameter β_q specifies the probability $P(\mathbf{R}_q = 1 | q \notin \mathbf{K})$ that a student solves q given that this last is not in his knowledge state.

Let the response patterns in $\{0, 1\}^n$ be indexed by $i \in \{1, 2, \ldots, 2^n\}$ and the knowledge states in $\mathcal{K}$ be indexed by $j \in \{1, 2, \ldots, m := |\mathcal{K}|\}$. Assuming local independence among the responses, given the knowledge states, the conditional probability of a response pattern $\mathbf{r}_i$ given state $K_j \in \mathcal{K}$ takes on the form

$$P(\mathbf{R} = \mathbf{r}_i | \mathbf{K} = K_j) = \prod_{k=1}^{n} \left[\alpha_k^{1-r_{ik}} (1 - \alpha_k)^{r_{ik}}\right]^{w_{jk}} \left[\beta_k^{r_{ik}} (1 - \beta_k)^{1-r_{ik}}\right]^{1-w_{jk}}, \tag{2}$$

where $r_{ik} \in \{0, 1\}$ is the kth element of response pattern $\mathbf{r}_i$, and the membership indicator for item k and knowledge state K_j is defined such that: $w_{jk} = 1$ if k is an element of K_j and $w_{jk} = 0$ if k is not an element of K_j. Equations (1) and (2) are the two basic equations of the BLIM. As already stated, the model applies to dichotomous items. Concerning possible extensions of the BLIM to polytomous items, a good starting point could be the approach followed by Schrepp (1997). Another important issue that is not considered in the present version of the BLIM is that of missing data. Although an extension of the model to this case seems not difficult to obtain, it will not be considered in the present article.

There are close connections between the BLIM and other models often used in cognitive diagnosis. The more closely related one is the so-called DINA (Deterministic Inputs Noisy AND-gate) model (Haberman, 1979; Junker, 2001; Macready & Dayton, 1977); see also in this connection, Maris (1999). There are essentially two main differences between the DINA and the BLIM. In the first place, the DINA model assumes the existence of a latent set of skills (or discrete cognitive components) each of which can be either present or absent in an individual, and that are used in a conjunctive manner to solve the items. This assumption is not made in the BLIM model, as it is developed on the level of the items and on the possible knowledge structures on the set Q of items. Secondly, in the DINA, a knowledge state is defined as *any* subset of skills, while a knowledge structure in the sense of Doignon and Falmagne's theory is typically a strict subset of the powerset on Q (i.e., not all subsets of Q are knowledge states, in general). Further parallels could be found between the two types of models, but they will not be discussed here.

1.2. Estimation of the BLIM

The basic local independence model is essentially a latent class model where the latent classes are the knowledge states in $\mathcal{K}$. Therefore, maximum likelihood estimates of the three types of parameters α, β, and π of the model can be obtained by an application of the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977).

The observed data sample is a $s \times n$ binary matrix $\mathbf{X}$ whose each row is the response pattern of a single subject to the n different items. Every single entry in the matrix $\mathbf{X}$ is denoted by $x_{vk} \in \{0, 1\}$, where $x_{vk} = 1$ if and only if subject v solved item k.

The knowledge states of the subjects are obviously unknown, but if there was complete information, every single subject v would be represented by a pair $(\mathbf{x}_v, K_v)$ where $\mathbf{x}_v$ is a $1 \times n$ binary vector representing the response pattern of v and $K_v \in \mathcal{K}$ is the knowledge state of that subject. Indicating with $\mathbf{Y}$ the collection of all such pairs in an empirical sample, the complete data log-likelihood of the model is

$$\ell(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{v=1}^{s} \ln P(\mathbf{x}_v, K_v|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}), \tag{3}$$

where $P(\mathbf{x}_v, K_v|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$ is the joint probability of response pattern $\mathbf{x}_v$ and knowledge state K_v given the model parameter vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$. In the iteration $t + 1$ of the EM algorithm, the conditional expectation of the complete data log-likelihood $\ell(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$ is maximized, given the observed data $\mathbf{X}$, and the parameter values $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$, $\boldsymbol{\pi}_t$ obtained in a previous iteration of the algorithm.

Let $P(\mathbf{x}_v|K_j, \boldsymbol{\alpha}, \boldsymbol{\beta})$ be the conditional probability of response pattern $\mathbf{x}_v$ given knowledge state $K_j \in \mathcal{K}$ and parameter values $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The Bayesian posterior probability of knowledge state K_j given response pattern $\mathbf{x}_v$ and previous estimates $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is

$$p_{jv,t} := \frac{P(\mathbf{x}_v|K_j, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)\pi_{jt}}{\sum_{l=1}^{m} P(\mathbf{x}_v|K_l, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)\pi_{lt}},$$

where $m = |\mathcal{K}|$ is the total number of knowledge states in the model. Then the conditional expected log-likelihood of the complete data at iteration $t+1$ turns out to be

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) := \sum_{v=1}^{s}\sum_{j=1}^{m} \ln\big[P(\mathbf{x}_v|K_j, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)\big] p_{jv,t} + \sum_{v=1}^{s}\sum_{j=1}^{m} \ln(\pi_{jt}) p_{jv,t} \tag{4}$$

and the parameter values that in this iteration maximize $L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$ are

$$\alpha_{k,t+1} = \frac{\sum_{v=1}^{s}\sum_{j=1}^{m} p_{jv,t}(1 - x_{vk})w_{jk}}{\sum_{v=1}^{s}\sum_{j=1}^{m} p_{jv,t} w_{jk}}$$

and

$$\beta_{k,t+1} = \frac{\sum_{v=1}^{s}\sum_{j=1}^{m} p_{jv,t}(1 - x_{vk})(1 - w_{jk})}{\sum_{v=1}^{s}\sum_{j=1}^{m} p_{jv,t}(1 - w_{jk})}$$

for each item k and

$$\pi_{j,t+1} = \frac{1}{s}\sum_{v=1}^{s} p_{jv,t}$$

for each knowledge state $K_j \in \mathcal{K}$.

2. Parameter Estimation in the Constrained BLIM

In this section, maximum likelihood parameter estimation is proposed in which the log-likelihood of the BLIM model is maximized, subject to the constraint that the parameters α and β are less or equal to some constant $\lambda \in [0, 1]$. Since α and β are probability values, it is clear that with $\lambda = 1$ the constrained maximization problem reduces to an unconstrained one.

Since maximization of the likelihood corresponds to minimization of the negative log-likelihood, the general setup of the problem is to minimize a nonlinear function $f(x)$ subject to the inequality constraints $g_l(x) \geq 0$, $l = 1, 2, \ldots, I$. An optimization method which solves this problem is any method which provides a solution satisfying the Karush–Kuhn–Tucker conditions (see, e.g., Wright, 1997). One such method is the so-called *log-barrier*, in which the original constrained minimization problem is converted to an unconstrained one, and the function to be minimized takes on the general form

$$h(x, \mu) := f(x) - \mu \sum_{l=1}^{I} \ln\big[g_l(x)\big],$$

where $\mu \geq 0$ is a *penalty parameter*. It is easily seen that as $g_l(x)$ tends to zero, $h(x, \mu)$ approaches $+\infty$, thus providing a "barrier" to crossing the boundary.[1]

Given an initial and sufficiently large value μ_0 of the penalization parameter, the minimization procedure takes place in a finite number of steps. In each new iteration $t+1$, the penalization parameter μ is gradually decreased by some amount (say, $\mu_{t+1} = c\mu_t$, for $0 < c < 1$) and an unconstrained maximization of the function $h(x, \mu_{t+1})$ is done.

If the initial guesses of the parameter estimates belong to the feasible region (i.e., all inequality constraints are satisfied at the outset) then: (a) if some local minimizer of the function f is an interior point of the feasible region, the barrier algorithm will reach such point otherwise

[1] A similar approach, but in a different context, is that of Houseman, Coull, and Betensky (2006).

(b) if such a minimizer lies outside the feasible region then a point belonging to the boundary of the region will be reached.

In our specific application, it happens that the constrained parameter space is a convex subset X of the whole parameter space. It should be noted that in this case, if also the function f to be minimized is convex, any point in the interior of the convex region X which is a local minimum of f is also a global minimum in X (see Proposition 1.22, p. 67 in Bertsekas, 1996).

Thus, given suitable upper bounds $\alpha_k^*, \beta_k^* \in [0, 1]$, in the problem at hand, there are four types of inequality constraints for each item k:

(i) $\alpha_k \geq 0$
(ii) $\alpha_k \leq \alpha_k^*$
(iii) $\beta_k \geq 0$
(iv) $\beta_k \leq \beta_k^*$

and the problem itself consists of maximizing the conditional expected log-likelihood in (4), under the constraints (i) to (iv). It should further be noted that between the two left-side terms of (4), only the first one depends on the two parameters α_k and β_k. Thus, the part of L which concretely undergoes a constrained maximization is just

$$L'(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \sum_{v=1}^{s} \sum_{j=1}^{m} \ln\big[P(\mathbf{x}_v | K_j, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)\big] p_{jv,t}.$$

By an application of the barrier method, such a constrained maximization corresponds to an unconstrained minimization of the function

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) := -\left\{ L'(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \mu \sum_{k=1}^{n} \ln\big[\alpha_k(\alpha_k^* - \alpha_k)\beta_k(\beta_k^* - \beta_k)\big] \right\},$$

where μ is the penalty parameter introduced above.

The function $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is minimized by setting to zero its first partial derivatives with respect to the parameters α_k and β_k. The first partial derivative of $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to the parameter α_k turns out to be

$$\frac{\partial Q(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \alpha_k} = \frac{(\alpha_k^* - \alpha_k)[a_{k,1}\alpha_k - (\mu + a_{k,0})(1 - \alpha_k)] + \mu\alpha_k(1 - \alpha_k)}{\alpha_k(1 - \alpha_k)(\alpha_k^* - \alpha_k)}, \tag{5}$$

where $a_{k,0}$ and $a_{k,1}$ are defined in the following way:

$$a_{k,0} := \sum_{v} \sum_{j} p_{jv,t}(1 - x_{vk}) w_{jk}$$

and

$$a_{k,1} := \sum_{i} \sum_{j} p_{jv,t} x_{vk} w_{jk}.$$

By setting the numerator of the right-hand term of (5) equal to zero, a second degree equation is obtained. Between the two roots of this equation, the one which actually satisfies the constraint $0 \leq \alpha_k \leq \alpha_k^*$ is given by the formula (let $a_k := a_{k,0} + a_{k,1}$):

$$\alpha_{k,t+1} = \frac{(a_{k,0} + 2\mu) + \alpha_k^*(a_k + \mu)}{2(a_k + 2\mu)} - \left\{ \frac{[(a_{k,0} + 2\mu) + \alpha_k^*(a_k + \mu)]^2 - \alpha_k^*(a_{k,0} + \mu)}{2(a_k + 2\mu)} \right\}^{\frac{1}{2}}.$$

Following a similar development for the parameters β_k, and defining the two quantities

$$b_k := \sum_v \sum_j p_{jv,t}(1 - w_{jk}),$$

and

$$b_{k,1} := \sum_v \sum_j p_{jv,t} x_{vk}(1 - w_{jk}),$$

one obtains that

$$\beta_{k,t+1} = \frac{(b_{k,1} + 2\mu) + \beta_k^*(b_k + \mu)}{2(b_k + 2\mu)} - \left\{ \frac{[(b_{k,1} + 2\mu) + \beta_k^*(b_k + \mu)]^2 - \beta_k^*(b_{k,1} + \mu)}{2(b_k + 2\mu)} \right\}^{\frac{1}{2}}$$

is the value of β_k which maximizes the expected log-likelihood at iteration $t + 1$ of the EM algorithm, under the constraint $0 \leq \beta_k \leq \beta_k^*$.

3. A Simulation Study

The two models BLIM and C-BLIM were assessed in a simulation study with respect to goodness-of-fit and goodness-of-recovery. In general, goodness-of-recovery tells how well the true model parameters are recovered by a specific estimation method. By goodness-of-recovery, we mean here also whether the knowledge structure that has generated the data is correctly recovered or identified. When the data are very noisy, i.e., when the careless errors and lucky guesses tend to be rather high for many items, this point may become problematic, because there could be no way at all to recover the true knowledge structure from the data. A question is thus under which conditions the knowledge structure that generated a given data set is recoverable and how well its parameters are recovered in the two models BLIM and C-BLIM.

3.1. Simulation of the Data Sets

The random data sets were generated according to the BLIM model, a fixed number of 20 items, and a fixed randomly generated knowledge structure $\mathcal{K}_0$ containing 200 knowledge states. The number of response patterns in each of the data sets (the sample size) was set to a fixed number of 1,000. Four distinct models, all based on the same knowledge structure $\mathcal{K}_0$, were used to generate the data. The main difference between these four models was in the choice of the upper bound (henceforth denoted by λ_{true}) of the interval of the uniform distribution that was used to generate the true careless error and lucky guess parameters (the lower bound was fixed at 0 for all four models). In the first model, this upper bound was set to 1.0; in the second model, it was set to 0.5; in the third model, it was 0.25, and in the last one, the upper bound was 0.1. For each of the four models a total number of 100 random data sets were generated.

3.2. Estimation of the Model Parameters

For each of the $100 \times 4 = 400$ random data sets, three alternative models were estimated. In the first model, henceforth called the *correct model*, the knowledge structure was exactly the one used to generate the data, namely $\mathcal{K}_0$. In the second and third models, two randomly generated knowledge structures, different from the correct one, were used. The knowledge structures used in these two additional models contained, as the correct one, 200 knowledge states. These two models will be henceforth referred to as the *incorrect models* 1 and 2, and the corresponding knowledge structures will be denoted, respectively, by $\mathcal{K}_1$ and $\mathcal{K}_2$.

To have a measure of how much the three knowledge structures differed one another, a particular discrepancy index for knowledge structures was computed (see the Appendix). The discrepancy between $\mathcal{K}_1$ and $\mathcal{K}_0$ was $D(\mathcal{K}_1, \mathcal{K}_0) = 4.01$ ($SD = 0.90$) and that between $\mathcal{K}_2$ and $\mathcal{K}_0$ was $D(\mathcal{K}_2, \mathcal{K}_0) = 4.02$ ($SD = 0.91$). This means that on the average, a knowledge state in $\mathcal{K}_1$ (resp. $\mathcal{K}_2$) differs for at least 4.01 (resp. 4.02) items from the knowledge states in $\mathcal{K}_0$. The distance between the two structures $\mathcal{K}_1$ and $\mathcal{K}_2$ was also computed and it was $D(\mathcal{K}_1, \mathcal{K}_2) = 3.94$ in a direction and $D(\mathcal{K}_2, \mathcal{K}_1) = 4.01$ in the other.

For every random data set, each of the three alternative models was estimated four different times by C-BLIM, each time with a different value of the upper bound λ_{est} of the α and β parameter estimates (equal for all α and β parameters). The following values of λ_{est} were used: $\lambda_{\text{est}} = 1$, $\lambda_{\text{est}} = 0.5$, $\lambda_{\text{est}} = 0.25$, $\lambda_{\text{est}} = 0.1$. This means that for each of the data sets there were a total of $3 \times 4 = 12$ distinct models to estimate.

3.3. Testing the Models

As a goodness-of-fit index of the different estimated models, the standard likelihood ratio Chi-square statistic was used. It is well known that for large and sparse data matrices the approximation to the asymptotic distribution of this statistic lacks of accuracy and cannot be used in practice. This is the case of the present simulation study, because with 20 items the theoretical number of distinct binary response patterns is huge (2^{20}) and a data set of 1,000 response patterns is definitely too small. However, for the purpose of comparing the goodness-of-fit of alternative models, likelihood ratio Chi-square can still be appropriate. Given two models i and j, parametric bootstrap (see, e.g., Langeheine, Pannekoek, and van de Pol, 1996; von Davier, 1997) can be used to estimate the proportion $P(\chi_i^2 < \chi_j^2)$ of data sets in which χ_i^2 happens to be less than χ_j^2.

To test goodness-of-fit of the correct knowledge structure against each of the two incorrect ones, the proportion $P(\chi_0^2 < \chi_1^2, \chi_2^2)$ was computed of data sets in which the Chi-square of the model incorporating the correct knowledge structure (χ_0^2) was less than the Chi-square of both incorrect models (χ_1^2 and χ_2^2). For short, in the sequel this proportion will be denoted by p_0. Along with this proportion, an average Chi-square was also computed for each of the estimated models by

$$\overline{\chi}_{jk}^2 = \frac{1}{100} \sum_{i=1}^{100} \chi_{ijk}^2,$$

where χ_{ijk}^2 is the value of the Chi-square obtained for data set i, model $j \in \{0, 1, 2\}$, and $\lambda_{\text{est}} = k \in \{0.1, 0.25, 0.5, 1.0\}$.

3.4. Results

Figure 1 compares goodness-of-fit of the model incorporating the correct knowledge structure with the other two models, for each of the four values of λ_{true} and each of the four values of λ_{est}. The values of λ_{true} are along the horizontal axis (we recall that only the four values 1.00, 0.50, 0.25, 0.10 were used in the simulations). The proportion p_0 is measured along the vertical axis. Each of the four curves in the diagram corresponds to a different value of the upper bound λ_{est}.

Some comments on the diagram in Figure 1 are in order. A first thing to notice is that regardless of the value of λ_{est}, the lowest value of the proportion p_0 is always obtained when the upper bound λ_{true} is 1.0. It should be observed that with such an upper bound the true α and β probabilities were higher than 0.5 for more than half of the items, with a maximum value of 0.93. Therefore, the noise in the generated data was pretty high. When parameter estimation is not constrained (i.e., $\lambda_{\text{est}} = 1.0$), the proportion p_0 is 0.31. That is, the correct model obtained

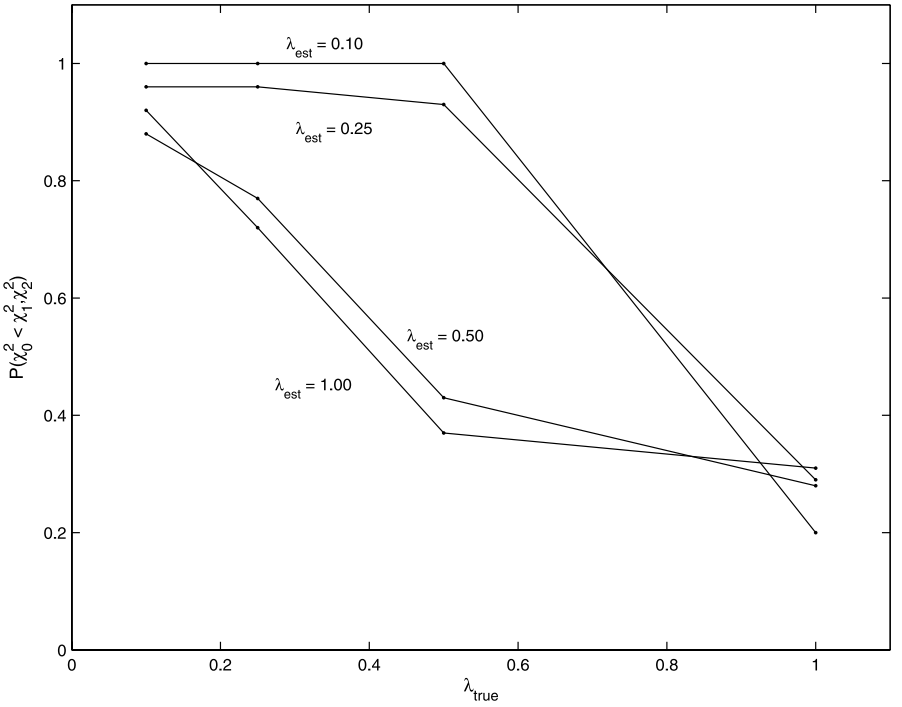

FIGURE 1.
Proportion of simulated data sets in which χ_0^2 (correct knowledge structure's Chi-square) turned out to be smaller than both χ_1^2 and χ_2^2 (incorrect knowledge structures' Chi-square). The upper bound λ_{true} is on the horizontal axis. Each curve corresponds to a different value of λ_{est}.

a Chi-square that is less than those of the incorrect models in only 31 data sets out of 100. The situation seems not improving with smaller values of λ_{est}. This result concerning $\lambda_{\text{true}} = 1.0$ indicates a problem in separating the correct model from the incorrect ones, that is a problem of recovering the true underlying knowledge structure when the data are too noisy. Further details on this point will be given again later.

Still concerning the curve of unconstrained model in Figure 1 ($\lambda_{\text{est}} = 1.0$), we see that the proportion p_0 increases as λ_{true} decreases and, for $\lambda_{\text{true}} = 0.1$, it reaches a value of 0.92. This just suggests that when the noise in the data is reduced enough, the true knowledge structure becomes uncoverable.

Results are rather similar for the curve of $\lambda_{\text{est}} = 0.5$. Instead, a quite interesting thing happens when the constraint λ_{est} is less or equal to 0.25. What we observe in these cases is that the two curves for $\lambda_{\text{est}} = 0.25$ and $\lambda_{\text{est}} = 0.1$ start decreasing much later than the other two. This suggests that unless the data are so noisy to make the true knowledge structure uncoverable, it is much likely that with the same amount of noise in the data, the true knowledge structure is correctly uncovered when the upper bound λ_{est} takes on small values. In particular, with $\lambda_{\text{est}} = 0.1$ and for $\lambda_{\text{true}} \leq 0.5$, the proportion of simulated data sets in which the correct model wins against the two incorrect ones is 1.0, that is, 100% of the data sets. Curiously enough, this happens in spite of the fact that such a small value of λ_{est} will certainly give rise to biased estimates for all those α and β parameters whose true value lies above λ_{est}. This point will be discussed later in more detail.

The four diagrams in Figure 2 show the average chi-square $\overline{\chi}_i^2$ obtained for each of the three alternative estimated models. The upper left diagram shows $\overline{\chi}_i^2$ for the case $\lambda_{\text{est}} = 1.0$

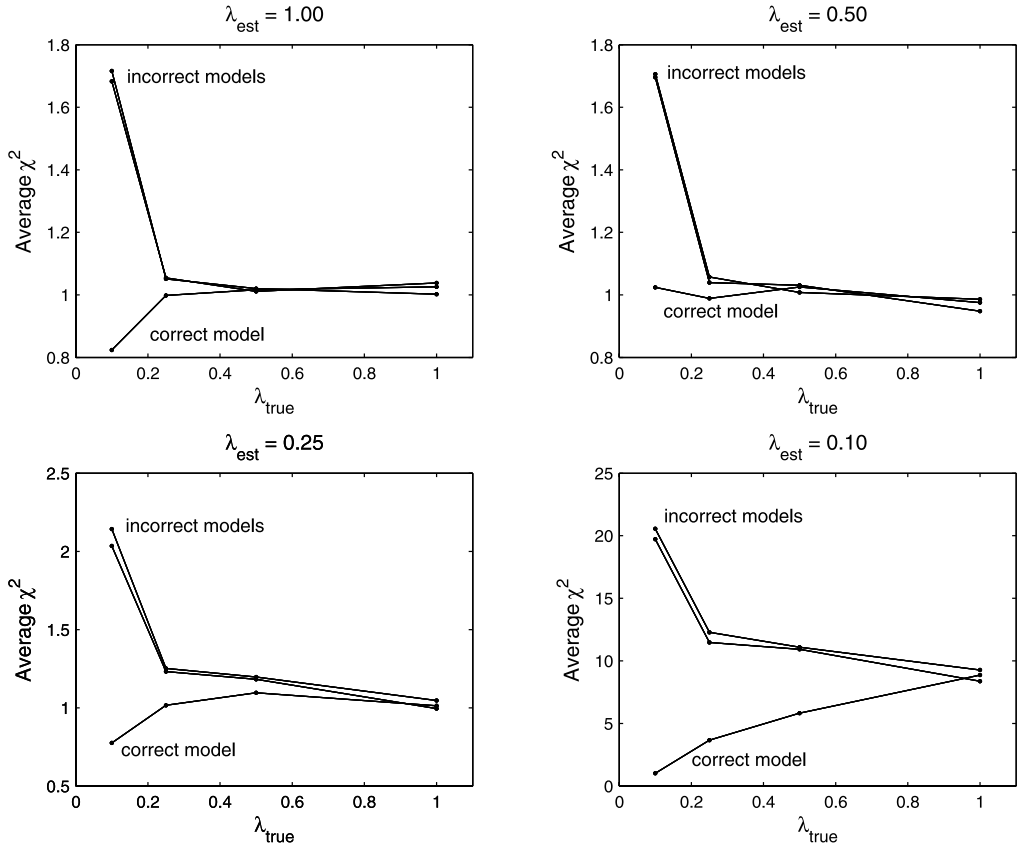

FIGURE 2.
Average Chi-square of the three estimated models with different values of the upper bound λ_{est}. The values along the vertical scale must be multiplied by 10^6.

(unconstrained model). The already mentioned problem of separating the correct model from the incorrect ones, when too much noise is present, appears once again in this diagram. For $\lambda_{\text{true}} = 1.0$, the average Chi-square of the three models is almost the same, meaning that on the average all of them fit equally well the data. Similar conclusions are drawn for the two cases $\lambda_{\text{true}} = 0.5$ and $\lambda_{\text{true}} = 0.25$. Only for $\lambda_{\text{true}} = 0.1$, the correct model is clearly separated from the other two. For $\lambda_{\text{est}} = 0.5$ (upper right diagram), things are quite similar. Finally, confirming what already observed in Figure 1, when $\lambda_{\text{est}} \leq 0.25$ separation between the correct model and the incorrect ones starts earlier (i.e., for higher values of λ_{true}).

To summarize, it seems important to recognize that for all values of λ_{est} used in this simulation study, the correct model cannot be uncovered when the data are too noisy ($\lambda_{\text{true}} = 1.0$). However, provided that λ_{true} is at most 0.5, the smaller the value of λ_{est} the better the separation between the correct model and the incorrect ones.

The results discussed so far only concern the problem of separating the correct knowledge structure from other incorrect ones. However, nothing has been stated yet about how well the parameters of the models incorporating such knowledge structures are recovered.

A scatter plot of the estimated parameters against generating parameters in both correct model and incorrect model 1 is depicted in Figure 3, for $\lambda_{\text{true}} = \lambda_{\text{est}} = 1.0$. Upper diagrams refer to the correct model and left diagrams refer to α and β parameters. It can be clearly seen that estimation is very bad for both models. The elevated amount of noise in the data prevents

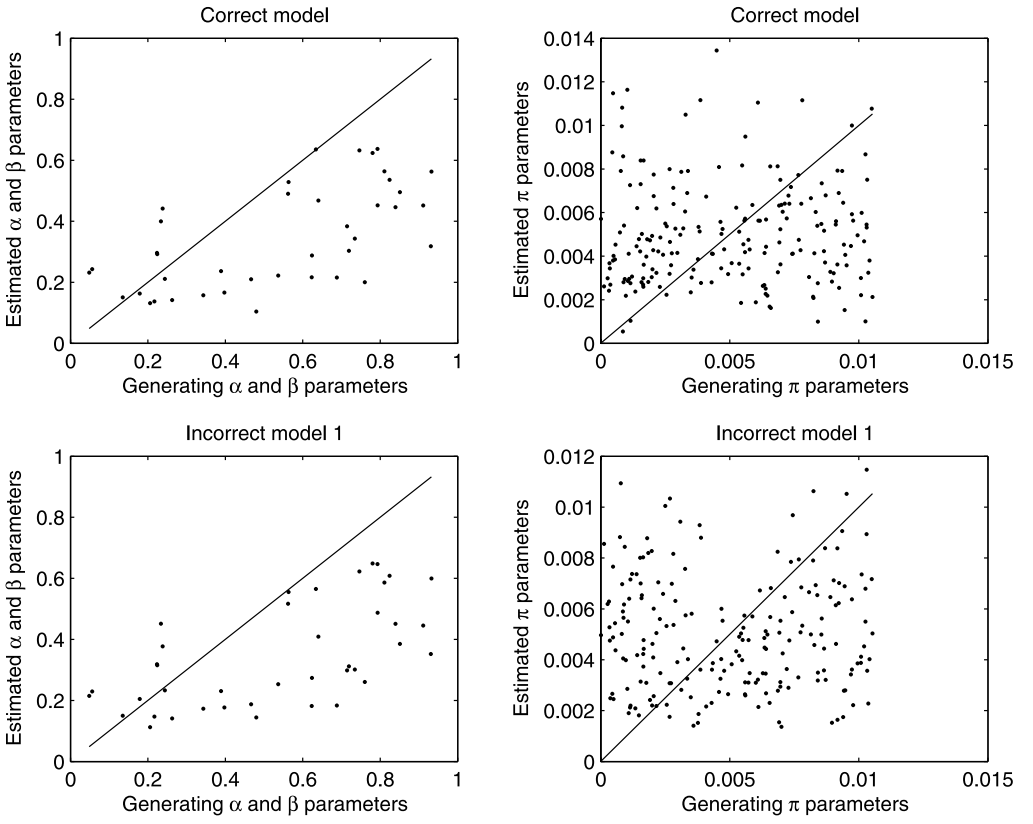

FIGURE 3.
Generating parameters (*x axis*) versus estimated parameters (*y axis*) in both correct model (*upper diagrams*) and incorrect model 1 (*lower diagrams*) for $\lambda_{\text{true}} = \lambda_{\text{est}} = 1.0$. The *straight line* $x = y$ is added for reference.

any reasonable estimate of the underlying parameters even when the estimated model is indeed correctly specified. It should further be observed that in both models the α and β parameters are almost all underestimated.

Figure 4 shows the case of $\lambda_{\text{true}} = 0.5$ and $\lambda_{\text{est}} = 1.0$. The bias of the correct model is negligible for all parameters. Concerning the α and β parameters, it is rather interesting to notice that in the incorrect model they are almost invariably overestimated. Recalling that for the condition $\lambda_{\text{true}} = 0.5$, $\lambda_{\text{est}} = 1.0$, correct and incorrect models fit equally well the data, this result is now quite well understood. The price that the incorrect models have to pay for reaching a fit which is as good as that of the correct one is an ad-hoc inflation of the careless error and lucky guess parameters.

The condition $\lambda_{\text{true}} = 0.5$, $\lambda_{\text{est}} = 0.25$ is shown in Figure 5. A comparison between the α and β estimates in the correct model and those in the incorrect one sheds light to a nice property of the C-BLIM. What it is seen in the top left diagram of this figure can be regarded as a "partial recovery" of the true model parameters. In fact, bias is small for only those parameters that lie below the upper bound λ_{est}. For the remaining parameters, all the estimates lie exactly on the boundary of the feasible region. This does not happen however with the incorrect model. For many of the α and β parameters that are smaller than λ_{est}, their estimates lie on the boundary anyway. This just reflects the tendency of the incorrect model to increase the α and β parameter estimates until the model likelihood reaches its maximum value. By imposing an upper bound to such parameter estimates, the C-BLIM simply prevents such an undesirable behavior.

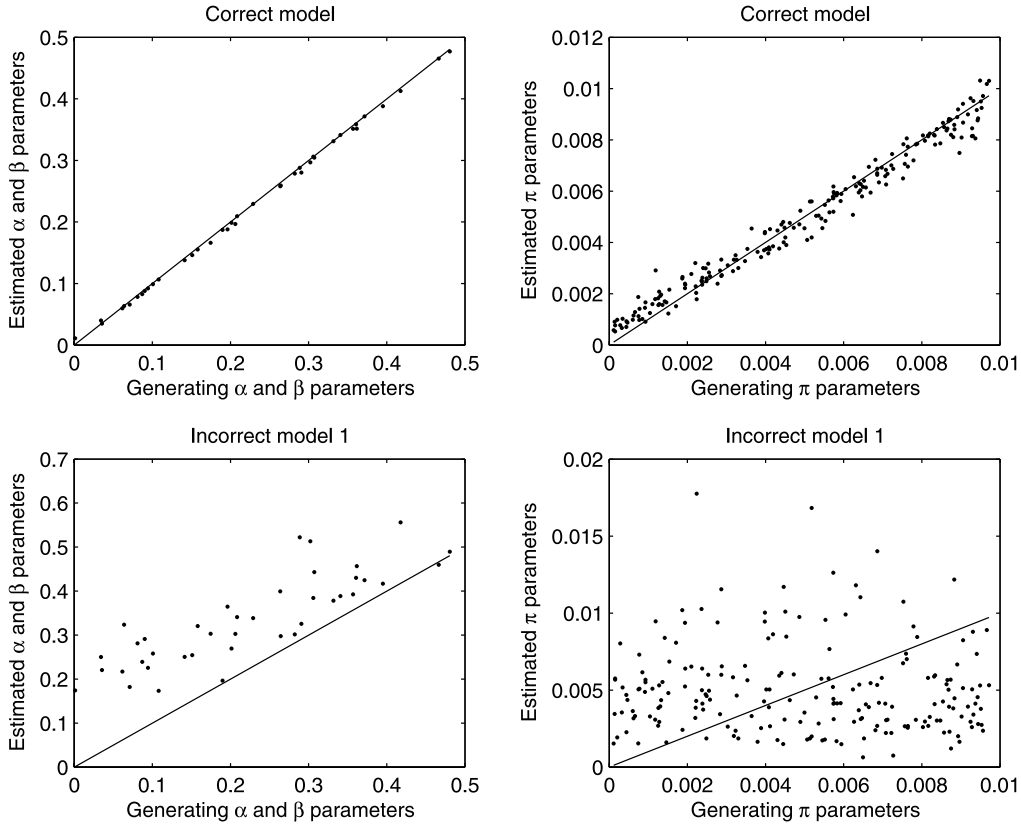

FIGURE 4.
Generating parameters (*x axis*) versus estimated parameters (*y axis*) in both correct model (*upper diagrams*) and incorrect model 1 (*lower diagrams*) for $\lambda_{\text{true}} = 0.5$ and $\lambda_{\text{est}} = 1.0$. The *straight line* $x = y$ is added for reference.

Finally, in Figure 6, the condition $\lambda_{\text{true}} = 0.1$, $\lambda_{\text{est}} = 0.25$ is considered. This time the bias of the estimates is negligible for all parameters of the correct model. However, once again most parts of the α and β estimates in the incorrect model end up to the boundary of the feasible region, confirming what already observed in the previous cases.

4. Final Remarks

The constrained BLIM and the method of analysis that can be derived from it seem especially useful when there is not much theory about the knowledge structure on a given set of problems. In such cases, more than one knowledge structure could be plausible in theory and the problem is which of them better fits the data. The simulation study suggests that in case of high noise in the data there is no way to separate a correctly specified model from others (all models will have approximately the same likelihood). However, if noise in the data is sufficiently small (say, less than 0.5), then the C-BLIM becomes helpful in the problem of recovering the knowledge structure underlying the data. In particular, if one of the alternative knowledge structures is correctly specified, then the introduction of an upper bound λ_{est} of at most 0.5 to the error parameters α and β allows to separate this particular knowledge structure from other incorrect ones. In fact, simulations show that as this upper bound gets lower, the likelihood of the incorrect models decreases much faster than that of the correct one.

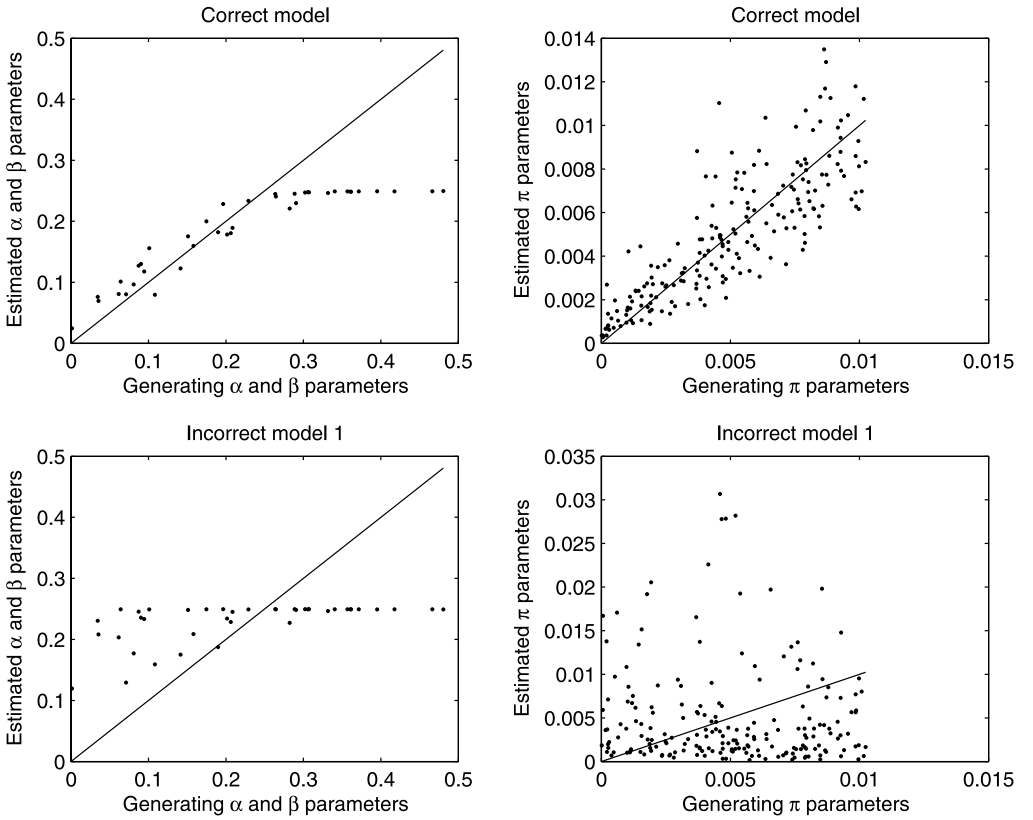

FIGURE 5.
Generating parameters (*x axis*) versus estimated parameters (*y axis*) in both correct model (*upper diagrams*) and incorrect model 1 (*lower diagrams*) for $\lambda_{\text{true}} = 0.5$ and $\lambda_{\text{est}} = 0.25$. The *straight line* $x = y$ is added for reference.

Once a knowledge structure has been separated from others, it is possible that for this particular model some of the alpha and beta parameters lie exactly on the upper bound. In this case, the upper bound itself could be elevated (up to, say, 0.5) and those parameters reestimated until they belong to the interior of the constrained parameter space. If at the end of this process, there are still some item parameters on the upper bound, this could be a sign of too much noise in the data concerning those items with high alpha or beta.

Alternative ways of constraining the error parameters of the model have been proposed. For instance, Junker and Sijtsma (2001) introduce a monotonicity constraint on such parameters. Given an item i, the monotonicity constraint requires that $\alpha_i \leq 1 - \beta_i$. Actually, this constraint does not prevent the possibility that the alpha and beta parameters get higher than a reasonable value anyway. That is, an ad hoc inflation of such parameters is still possible. As discussed, above the danger is that an incorrect model could have the same likelihood as a correctly specified one.

In the simulation studies discussed in Section 3, the assumption was made that the correct knowledge structure belongs to the collection of models that undergo a goodness-of-fit test. This situation could not hold in practice. In this case, the aim could be to establish which of the models at hand is a better approximation of the "true" knowledge structure underlying the data. The simulation study could be extended to such cases by considering some distance of each of these models from the "true" model (see, e.g., the Appendix) and to study how the model likelihood varies as a function of the imposed constraints and the distance between tested and "true" model.

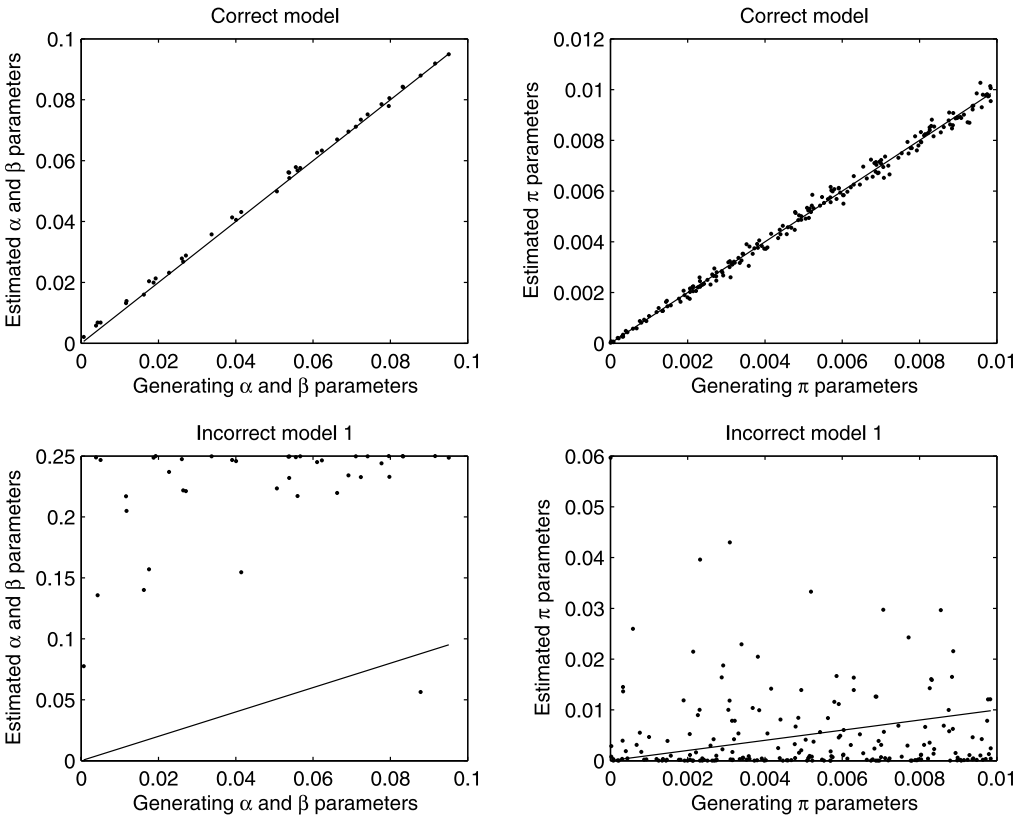

FIGURE 6.
Generating parameters (*x axis*) versus estimated parameters (*y axis*) in both correct model (*upper diagrams*) and incorrect model 1 (*lower diagrams*) for $\lambda_{\text{true}} = 0.1$ and $\lambda_{\text{est}} = 0.25$. The *straight line* $x = y$ is added for reference.

Acknowledgement

We are grateful to two anonymous referees for their useful comments on an earlier version of the article.

Appendix: Distance between Two Knowledge Structures

A measure of the distance between two knowledge structure can be obtained in the following way (for details, see Doignon & Falmagne, 1999). Given two knowledge structures $\mathcal{K}$ and $\mathcal{K}'$ and two knowledge states $K \in \mathcal{K}$ and $K' \in \mathcal{K}'$, the symmetric difference between K and K' is defined to be

$$K \Delta K' := \big|(K \setminus K') \cup (K' \setminus K)\big|.$$

The subset $K \Delta K'$ contains all elements in the union of the two sets not belonging to their intersection. The distance of a state $K \in \mathcal{K}$ from the knowledge structure $\mathcal{K}'$ is then computed as

$$d(K, \mathcal{K}') := \min\{K \Delta K' : K' \in \mathcal{K}'\}.$$

Then the discrepancy between $\mathcal{K}$ and $\mathcal{K}'$ is defined to be the mean of the minimum distances $d(K, \mathcal{K}')$ of the states $K \in \mathcal{K}$ from $\mathcal{K}'$:

$$D(\mathcal{K}, \mathcal{K}') = \frac{1}{|\mathcal{K}|} \sum_{K \in \mathcal{K}} d(K, \mathcal{K}').$$

It is clear that $D(\mathcal{K}, \mathcal{K}) = 0$. It should also be observed that D is not commutative, i.e., $D(\mathcal{K}, \mathcal{K}') \neq D(\mathcal{K}', \mathcal{K})$, in general. Nonetheless, $D(\mathcal{K}, \mathcal{K}')$ can be regarded as a measure of how well a knowledge structure $\mathcal{K}$ approximates another knowledge structure $\mathcal{K}'$. Along with $D(\mathcal{K}, \mathcal{K}')$, a standard deviation $SD(\mathcal{K}, \mathcal{K}')$ can also be computed.

References

Albert, D. (Ed.) (1994). *Knowledge structures*. New York: Springer.

Albert, D., & Lukas, J. (Eds.) (1999). *Knowledge spaces: theories, empirical research, applications*. Mahwah: Lawrence Erlbaum.

Bertsekas, D. (1996). *Constrained optimization and Lagrange multiplier methods*. Belmont: Athena Scientific.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

Doignon, J.-P., & Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, *23*, 175–196.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin: Springer.

Falmagne, J.-C., & Doignon, J.-P. (1988). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, *41*, 1–23.

Falmagne, J.C., Doignon, J.P., Koppen, M., Villano, M., & Johannesen, L. (1990). Introduction to knowledge spaces: how to build, search and test them. *Psychological Review*, *97*(2), 201–224.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231.

Haberman, S. (1979). *Qualitative data analysis* (Vols. 1, 2). New-York: Academic.

Houseman, E., Coull, B., & Betensky, R. (2006). Feature-specific penalized latent class analysis for genomic data. *Biometrics*, *62*(4), 1062–1070.

Junker, B.W. (2001). On the interplay between nonparametric and parametric irt, with some thoughts about the future. In A. Boomsma, M.A.J.V. Duijn, & T.A.B. Snijders (Eds.), *Essays on item response theory* (pp. 274–276). New York: Springer.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.

Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, *24*, 492–516.

Macready, G.B., & Dayton, C.M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*(2), 99–120.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212.

Schrepp, M. (1997). A generalization of knowledge space theory to problems with more than two answer alternatives. *Journal of Mathematical Psychology*, *41*(3), 237–243.

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: results of a Monte Carlo study. *Methods of Psychological Research*, *2*(2), 29–48.

Wright, S. (1997). *Primal-dual interior-point methods*. Philadelphia: SIAM.

Manuscript Received: 3 NOV 2006
Final Version Received: 14 OCT 2008
Published Online Date: 8 JAN 2009