FLUCTUATION OF ESTIMATES IN AN EM PROCEDURE FOR CATEGORICAL DATA

Seong Ho Kim and Sung-Ho Kim

Division of Applied Mathematics, Korea Advanced Institute of Science and Technology

Daejeon 305-701, South Korea

Abstract

Estimates from an EM algorithm are somewhat sensitive to the initial values for the estimates, and this sensitivity is likely to increase when the model becomes larger and more complicated. In this paper, we examined how the estimates fluctuate during an EM procedure for a recursive model of categorical variables. It is found that the fluctuation takes place mostly during the initial stage of the procedure and that it can be reduced by applying a Bayes method of estimation. Both real and simulated data are used for illustration.

Key Words: Bayes method; Calibrated initial values; Conditional positive association; Directed acyclic graph; Rank order of estimates.

1 INTRODUCTION

The EM method as proposed by Dempster et al. (1977) has been used widely for parameter estimation when data are incomplete with missing values for a set of random variables. It is easy to understand and the algorithm consists of two operations, expectation (E-step) for the missing variables and likelihood-maximization (M-step) using the estimates from the preceding operation. There is an abundance of literature on the EM method concerning the issues of applications, convergence rates, and a variety of improved versions of it (see, for example, van Dyk and Meng (1997)). We will confine our attention on a possible fluctuation of the estimates during the estimation process, and all the variables are assumed categorical. The models we consider in this paper are graphical log-linear models whose model structures are representable via directed acyclic graphs (Whittaker, 1990).

In educational testing, a task performance model is developed based on test data and prerequisite or causal relations among the cognitive features such as problem solving capabilities, computational skills, adaptability, multi-step thinking ability, memory of facts, meta knowledge, etc. The cognitive feature will be termed knowledge units. It is reasonable to assume that better knowledge states may yield a better performance. In the same context, a better state of a set of prerequisite knowledge units of a certain knowledge unit may yield a better understanding of the knowledge unit (Haertel and Wiley, 1992). This phenomenon is called in stochastic terms positive dependence among variables, namely conditional (positive) association (see Holland and Rosenbaum (1986) and Junker and Ellis (1997) for a detailed description.) In addition to educational testing and cognitive science, this conditional association is an important notion in the research fields such as systems reliability and biology also (Holland and Rosenbaum, 1986).

Under the assumption of conditional association, conditional probabilities are expected to be ordered according to the states of conditioning variables. For instance, suppose that X_3 is causally related with X_1 and X_2 and that these variables are all binary, taking values 0 or 1. We write, for simplicity, $P(X_3 = 1|(X_1, X_2) = (x_1, x_2)) = g(x_1, x_2)$. Then we expect that $g(0,0) \leq \min\{g(0,1), g(1,0)\} \leq \max\{g(0,1), g(1,0)\} \leq g(1,1)$. To express this in general terms, let $\mathbf{X} = (X_1, \cdots, X_r)$ be causally related to Y and assume that the variables are all binary. The order between two vectors $\mathbf{x}$ and $\mathbf{x}'$ is expressed as $\mathbf{x} \preceq \mathbf{x}'$ if $x_i \leq x_i'$ for $i = 1, \cdots, r$. Then we say that Y and $\mathbf{X}$ are conditionally positively associated if the following holds:

$$P(Y = 1|\mathbf{X} = \mathbf{x}) \leq P(Y = 1|\mathbf{X} = \mathbf{x}') \tag{1}$$

when $\mathbf{x} \preceq \mathbf{x}'$. We will call this the conditional positive association (CPA) condition.

When the CPA is assumed for a set of variables, the estimates for the variables are expected to satisfy the CPA condition. However, experience says that as the number of conditioning variables increases, it is more probable that the CPA is violated in the estimates of the conditional probability. What make things more difficult is that the estimates from an EM algorithm depend upon the initial values for the estimates (Wu, 1983), which becomes more serious as the model structure gets more complicated and more variables are involved in the model. This sensitivity of the estimates to the initial values can be reduced to some extent by applying the calibration method of Kim (2002).

Since our data involve unobservable variables, we applied an EM algorithm, and the model we considered is a recursive model (Wermuth and Lauritzen, 1983) of categorical variables. If the CPA is violated at some point of the EM procedure and the estimates remain about the same thereafter, there is no point in continuing the EM procedure. Thus, our main interest in this paper is in how the estimates fluctuate during the EM process and when we can expect the CPA condition to hold in the estimates.

There may be a number of local maxima when a model includes many missing values and unsuitably extreme values can be produced by maximum likelihood for the probabilities of the unobserved variables in the model (Thiesson, 1991). In an effort to avoid these phenomena, it is suggested in the literature that we penalize the likelihood for deviating from prior assessments (Green, 1990) and that we impose Dirichlet prior distributions when carrying out parameter estimation (Bishop, Fienberg, and Holland, 1975; Spiegelhalter, Dawid, Lauritzen, and Cowell, 1993).

Sufficient conditions are proposed in this paper under which the rank order of the estimates is preserved before and after an E-step of an EM procedure. Given a model structure represented in a directed acyclic graph, we can express the joint probability for the whole model as a product of conditional or marginal probabilities of the variables in the model, where the set of conditional variables for a variable is determined according to the direction of the edges in the graph. The estimates are for the conditional probabilities of individual variables in a model. Thus, the rank order of the estimates remains the same before and after an M-step. We will elaborate in later sections on how the rank order of the estimates changes during the EM process. Furthermore, we will show, through empirical results, how a Bayes method improves the estimation in the context of rank order.

This paper consists of 6 sections. In section 2, we introduce notation and terminologies and, in section 3, we describe the EM algorithm that is used in this paper. Section 4 then proposes conditions under which the rank order of estimates is preserved before and after an E-step, and it is shown that the rank order remains the same before and after an M-step. In section 5, we present empirical results using both real and simulated data and by applying both Bayes and non-Bayes method. The results indicate that the Bayes method is better than the other method in the sense of the CPA condition. Section 6 concludes the paper with summarizing

remarks.

2 NOTATION AND TERMINOLOGY

A contingency table is formed by classifying a number of objects according to a set of criteria and counting the number of objects in each classification. We express this formally by introducing a finite set V of *classification criteria* and for each $v \in V$, a finite set $\mathcal{I}_v$ of the possible criterion *levels* for v. We often refer to the criteria as *variables*. The *cells* of the table are the elements $i = (i_v)_{v \in V}$ of the product $\mathcal{I}$ of the level sets $i \in \mathcal{I} = \times_{v \in V} \mathcal{I}_v$. Data typically appear in two different forms: as a *list* of $|n|$ objects $(i^1, ..., i^{|n|})$, where each entry identifies which cell a given object belongs to, or as a *contingency table* of *counts* $n = \{n(i)\}_{i \in \mathcal{I}}$. Here $|n| = \sum_{i \in \mathcal{I}} n(i)$. We denote by $N(i)$ the random quantity of the cell count at cell $i \in \mathcal{I}$.

If we introduce the indicator functions

$$\mathcal{X}^i(j) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise,} \end{cases}$$

the counts are given as

$$n(i) = \sum_{\nu=1}^{|n|} \mathcal{X}^i(j^\nu)$$

The table has a *dimension* equal to the number $|V|$ of variables.

Basic manipulations with contingency tables are those of forming marginal tables and of forming slices. An *A-marginal table* is, for $A \subseteq V$, obtained by classifying the objects solely according to the criteria in A, i.e., by only considering the variables in A. It has *marginal cells* $i_A \in \mathcal{I}_A = \times_{v \in A} \mathcal{I}_v$. The marginal counts are the quantities $n(i_A)$. Again, if we let

$$\mathcal{X}^{i_A}(j) = \begin{cases} 1 & \text{if } j_A = i_A \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$n(i_A) = \sum_{\nu=1}^{|n|} \mathcal{X}^{i_A}(j^\nu) = \sum_{j \in \mathcal{I}} n(j) \mathcal{X}^{i_A}(j).$$

For the marginal corresponding to the empty set we get $n(i_\emptyset) = |n|$, the total number of observations.

We denote the vector of *expected cell counts* by $(m(i))_{i\in\mathcal{I}}$ and that of the *maximum likelihood estimate* of the mean vector by $(\widehat{m}(i))_{i\in\mathcal{I}}$.

The relationship among the set of classification variables that are involved in a recursive model can be represented by a directed acyclic graph (DAG). A DAG consists of nodes and arrows (or directed edges). $a \to b$ indicates that the state of b is influenced by the state of a. In this situation, we call node a a parent node of node b and call b a child node of a. We will denote by $pa(v)$ the set of the parent nodes of v and let $fa(v) = \{v\} \cup pa(v)$. A path from a to b is a sequence of nodes $a \to \cdots \to b$, where all the arrows are in the same direction. The descendants $de(a)$ of a are the nodes b such that there is a path from a to b. The expressions $pa(A)$ and $fa(A)$ denote the collection of parents and family of variables in $A \subseteq V$: $pa(A) = \cup_{a\in A} pa(a) \setminus A$ and $fa(A) = \cup_{a\in A} fa(a)$. A node which does not have any child node will be called a *terminal* node, and the node which does not have any parent node will be called a *root* node. Since the model structure of a graphical model is represented by a graph, we use random variable and node in the same sense.

3 EM ALGORITHM

Consider a recursive model of a set V of classification variables with $V = \Delta \cup \Lambda$ where Δ is a set of the observed variables and Λ a set of the unobserved (latent) variables. We assume that all the variables in Δ are terminal in the DAG of the recursive model. We will describe an EM procedure for the model.

An E-step consists in calculating $\widehat{m}(i_V)^{(r+1)} = E\left(N(i_V)|\{n(i_\Delta)\}_{i_\Delta\in\mathcal{I}_\Delta},\ \{\hat{m}(i_V)^{(r)}\}_{i_V\in\mathcal{I}}\right)$ based on the observation $\{n(i_\Delta)\}_{i_\Delta\in\mathcal{I}_\Delta}$ and the current estimates $\{\hat{m}(i_V)^{(r)}\}_{i_V\in\mathcal{I}}$ of $\{m(i_V)\}_{i_V\in\mathcal{I}}$. Thus

$$\widehat{m}(i_V)^{(r+1)} = n(i_\Delta)\frac{\widehat{m}(i_V)^{(r)}}{\widehat{m}(i_\Delta)^{(r)}}. \tag{2}$$

Once the E-step is carried out, the new estimates satisfy $\widehat{m}(i_\Delta)^{(r+1)} = n(i_\Delta)$. This means that the E-step calibrates the estimates to data.

The estimates, $\widehat{m}(i_V)^{(r+1)}$, resulting from an E-step are regarded as complete data for the following M-step, where new estimates, $\widehat{m}(i_V)^{(r+2)}$, are calculated through likelihood-

maximization. In other words, the M-step is implemented according to the expression

$$\widehat{m}(i_V)^{(r+2)} = |n| \prod_{v \in V} \frac{\widehat{m}(i_{fa(v)})^{(r+1)}}{\widehat{m}(i_{pa(v)})^{(r+1)}}. \tag{3}$$

For convenience's sake, we will denote $n(i_V)/|n|$ by $q(i_V)$, $n(i_A)/|n|$ by $q(i_A)$, $\widehat{m}(i_V)/|n|$ by $\widehat{p}(i_V)$, and $\widehat{m}(i_A)/|n|$ by $\widehat{p}(i_A)$, for $A \subseteq V$. Then,

$$\frac{\widehat{m}(i_{fa(v)})^{(r)}}{\widehat{m}(i_{pa(v)})^{(r)}} = \frac{\widehat{p}(i_{fa(v)})^{(r)}}{\widehat{p}(i_{pa(v)})^{(r)}} = \widehat{p}(i_v|i_{pa(v)})^{(r)}.$$

Replacing m with p and n with q in (2) and (3) yields, respectively,

$$\widehat{p}(i_V)^{(r+1)} = q(i_\Delta)\frac{\widehat{p}(i_V)^{(r)}}{\widehat{p}(i_\Delta)^{(r)}}.$$

and

$$\widehat{p}(i_V)^{(r+2)} = \prod_{v \in V} \widehat{p}(i_v|i_{pa(v)})^{(r+1)}.$$

The last expression says that the likelihood-maximization is implemented in the form of a product of the "maximum likelihood" estimates of the conditional probabilities, $p(i_v|i_{pa(v)})$, based on the results from the preceding E-step. The quotation marks are used since the estimates that are calculated before the convergence of an EM procedure are not maximum likelihood estimates in the real sense of the words.

An E-step and its subsequent M-step form a cycle of the EM algorithm and we repeat the cycle until the estimates converge. The convergence of the EM procedure is well established (Dempster et al. 1977, Wu 1983), and the EM procedure for categorical data is described in detail in Chapter 9 of Little and Rubin (1987).

4 RANK ORDER OF THE ESTIMATES

In this section, we investigate how the rank order of the estimates changes in the EM process. Since different methods are applied at the E- and M-steps, the estimates are compared at each of the steps.

4.1 After an E-step

We present two theorems that provide sufficient conditions for the rank order of the estimates of the conditional probabilities to remain the same before and after an E-step. In Theorem 1, we examine how the rank order of estimates varies at a terminal node which is observable, and we do the same for a non-terminal node in Theorem 2.

Theorem 1. *Let $v(\in \Delta)$ be terminal, v be binary, and $pa(v)(\neq \emptyset) \subseteq \Lambda$. Assume*

$$\frac{E\left[\frac{q(i_v,k_\delta)}{\widehat{p}(i_v,k_\delta)^{(r)}};\widehat{p}(k_\delta|i_{pa(v)})^{(r)}\right]}{E\left[\frac{q(j_v,k_\delta)}{\widehat{p}(j_v,k_\delta)^{(r)}};\widehat{p}(k_\delta|i_{pa(v)})^{(r)}\right]}\frac{E\left[\frac{q(j_v,k_\delta)}{\widehat{p}(j_v,k_\delta)^{(r)}};\widehat{p}(k_\delta|j_{pa(v)})^{(r)}\right]}{E\left[\frac{q(i_v,k_\delta)}{\widehat{p}(i_v,k_\delta)^{(r)}};\widehat{p}(k_\delta|j_{pa(v)})^{(r)}\right]} = 1+\eta, \tag{4}$$

where η is an arbitrary real value, $\delta = \Delta \setminus \{v\}$, and

$$E\left[\frac{q(i_v,k_\delta)}{\widehat{p}(i_v,k_\delta)^{(r)}};\widehat{p}(k_\delta|i_{pa(v)})^{(r)}\right] = \sum_{k_\delta \in \mathcal{I}_\delta}\frac{q(i_v,k_\delta)}{\widehat{p}(i_v,k_\delta)^{(r)}}\widehat{p}(k_\delta|i_{pa(v)})^{(r)}.$$

If

$$\widehat{p}(i_v|i_{pa(v)})^{(r)} > \widehat{p}(i_v|j_{pa(v)})^{(r)}, \tag{5}$$

then we have the following result:

(i) If $\widehat{p}(i_v|j_{pa(v)})^{(r)} = 0$, then

$$\widehat{p}(i_v|i_{pa(v)})^{(r+1)} > \widehat{p}(i_v|j_{pa(v)})^{(r+1)} = 0 \tag{6}$$

(ii) otherwise,

$$\frac{\widehat{p}(i_v|i_{pa(v)})^{(r+1)}}{\widehat{p}(i_v|j_{pa(v)})^{(r+1)}} > 1+\eta\frac{\rho}{1+\rho}, \tag{7}$$

where

$$\rho = \frac{\widehat{p}(j_v|i_{pa(v)})^{(r)}E\left[\frac{q(j_v,k_\delta)}{\widehat{p}(j_v,k_\delta)^{(r)}};\widehat{p}(k_\delta|i_{pa(v)})^{(r)}\right]}{\widehat{p}(i_v|i_{pa(v)})^{(r)}E\left[\frac{q(i_v,k_\delta)}{\widehat{p}(i_v,k_\delta)^{(r)}};\widehat{p}(k_\delta|i_{pa(v)})^{(r)}\right]}.$$

Proof.

$$\begin{aligned}
\widehat{p}(i_v|i_{pa(v)})^{(r+1)} &= \frac{\widehat{p}(i_v, i_{pa(v)})^{(r+1)}}{\widehat{p}(i_{pa(v)})^{(r+1)}} \\
&= \sum_{k_{\Delta\setminus\{v\}}} \sum_{k_{\Lambda\setminus pa(v)}} \frac{\widehat{p}(i_v, i_{pa(v)}, k_{\Delta\setminus\{v\}}, k_{\Lambda\setminus pa(v)})^{(r)}}{\widehat{p}(i_v, k_{\Delta\setminus\{v\}})^{(r)}}\ q(i_v, k_{\Delta\setminus\{v\}}) \\
&\quad \Big/ \sum_{k_v} \sum_{k_{\Delta\setminus\{v\}}} \sum_{k_{\Lambda\setminus pa(v)}} \frac{\widehat{p}(k_v, i_{pa(v)}, k_{\Delta\setminus\{v\}}, k_{\Lambda\setminus pa(v)})^{(r)}}{\widehat{p}(k_v, k_{\Delta\setminus\{v\}})^{(r)}}\ q(k_v, k_{\Delta\setminus\{v\}}) \qquad (8)
\end{aligned}$$

For simplicity of expression, we set $\delta = \Delta \setminus \{v\}$,

$$p_{i|j} = \widehat{p}(i_v|j_{pa(v)})^{(r)},$$

$$\widehat{g}_r(i_v, k_\delta) = \frac{q(i_v, k_{\Delta\setminus\{v\}})}{\widehat{p}(i_v, k_{\Delta\setminus\{v\}})^{(r)}},$$

and

$$E_{i|j} = E\Big[\widehat{g}_r(i_v, k_\delta); \widehat{p}(k_\delta|j_{pa(v)})^{(r)}\Big].$$

Then (8) may be written

$$\begin{aligned}
&\sum_{k_\delta} \widehat{p}(i_v, i_{pa(v)}, k_\delta)^{(r)} \widehat{g}_r(i_v, k_\delta) \\
&\Big/ \sum_{k_v} \sum_{k_\delta} \widehat{p}(k_v, i_{pa(v)}, k_\delta)^{(r)} \widehat{g}_r(k_v, k_\delta) \\
= \ &\widehat{p}(i_v, i_{pa(v)})^{(r)} \sum_{k_\delta} \widehat{g}_r(i_v, k_\delta) \widehat{p}(k_\delta|i_{pa(v)})^{(r)} \\
&\Big/ \sum_{k_v} \widehat{p}(k_v, i_{pa(v)})^{(r)} \sum_{k_\delta} \widehat{g}_r(k_v, k_\delta) \widehat{p}(k_\delta|i_{pa(v)})^{(r)} \\
&\qquad \text{(since } v \text{ is independent of } \delta \text{ given that } pa(v) \text{ is known)} \\
= \ &p_{i|i} E_{i|i} \big/ \big[p_{i|i} E_{i|i} + p_{j|i} E_{j|i}\big] \qquad \text{(since } v \text{ is binary).} \qquad (9)
\end{aligned}$$

By applying the same argument, we have

$$\widehat{p}(i_v|j_{pa(v)})^{(r+1)} = p_{i|j} E_{i|j} \Big/ \big[p_{i|j} E_{i|j} + p_{j|j} E_{j|j}\big]. \qquad (10)$$

From equations (9) and (10) follows that

$$
\begin{aligned}
& \widehat{p}(i_v|i_{pa(v)})^{(r+1)} \big/ \widehat{p}(i_v|j_{pa(v)})^{(r+1)} \\
= \;& \big[p_{i|i}p_{i|j}E_{i|i}E_{i|j} + p_{i|i}p_{j|j}E_{i|i}E_{j|j}\big] \big/ \big[p_{i|i}p_{i|j}E_{i|i}E_{i|j} + p_{i|j}p_{j|i}E_{i|j}E_{j|i}\big] \\
= \;& \big[1 + p_{i|i}p_{j|j}E_{i|i}E_{j|j} \big/ \big(p_{i|i}p_{i|j}E_{i|i}E_{i|j}\big)\big] \big/ \big[1 + p_{i|j}p_{j|i}E_{i|j}E_{j|i} \big/ \big(p_{i|i}p_{i|j}E_{i|i}E_{i|j}\big)\big] \\
= \;& \big[1 + p_{j|j}E_{j|j} \big/ \big(p_{i|j}E_{i|j}\big)\big] \big/ \big[1 + p_{j|i}E_{j|i} \big/ \big(p_{i|i}E_{i|i}\big)\big] \\
= \;& \big[1 + (1+\eta)p_{j|j}E_{j|i} \big/ \big(p_{i|j}E_{i|i}\big)\big] \big/ \big[1 + p_{j|i}E_{j|i}/\big(p_{i|i}E_{i|i}\big)\big] \qquad \text{(by condition (4))} \\
> \;& \big[1 + (1+\eta)p_{j|i}E_{j|i} \big/ \big(p_{i|i}E_{i|i}\big)\big] \big/ \big[1 + p_{j|i}E_{j|i} \big/ \big(p_{i|i}E_{i|i}\big)\big] \\
& \qquad\qquad \text{(since } p_{j|j}/p_{i|j} > p_{j|i}/p_{i|i} \text{ by condition (5))} \\
= \;& 1 + \eta\big[p_{j|i}E_{j|i}/\big(p_{i|i}E_{i|i}\big)\big] \big/ \big[1 + p_{j|i}E_{j|i}/\big(p_{i|i}E_{i|i}\big)\big].
\end{aligned}
$$

This completes the proof. □

Theorem 2. *Let $v \in \Lambda$, v be binary, and $pa(v) \cap \Delta = \emptyset$. Assume*

$$
\frac{E\left[\frac{q(k_\Delta)}{\widehat{p}(k_\Delta)^{(r)}}; \widehat{p}(k_\Delta|i_v, i_{pa(v)})^{(r)}\right]}{E\left[\frac{q(k_\Delta)}{\widehat{p}(k_\Delta)^{(r)}}; \widehat{p}(k_\Delta|j_v, i_{pa(v)})^{(r)}\right]} \frac{E\left[\frac{q(k_\Delta)}{\widehat{p}(k_\Delta)^{(r)}}; \widehat{p}(k_\Delta|j_v, j_{pa(v)})^{(r)}\right]}{E\left[\frac{q(k_\Delta)}{\widehat{p}(k_\Delta)^{(r)}}; \widehat{p}(k_\Delta|i_v, j_{pa(v)})^{(r)}\right]} = 1 + \eta, \tag{11}
$$

where η is an arbitrary real value. If

$$
\widehat{p}(i_v|i_{pa(v)})^{(r)} > \widehat{p}(i_v|j_{pa(v)})^{(r)}, \tag{12}
$$

then we have the following result:

(i) If $\widehat{p}(i_v|j_{pa(v)})^{(r)} = 0$, then

$$
\widehat{p}(i_v|i_{pa(v)})^{(r+1)} > \widehat{p}(i_v|j_{pa(v)})^{(r+1)} = 0 \tag{13}
$$

(ii) otherwise,

$$
\frac{\widehat{p}(i_v|i_{pa(v)})^{(r+1)}}{\widehat{p}(i_v|j_{pa(v)})^{(r+1)}} > 1 + \eta\frac{\rho}{1+\rho}, \tag{14}
$$

where

$$
\rho = \frac{\widehat{p}(j_v|i_{pa(v)})^{(r)} E\left[\frac{q(k_\Delta)}{\widehat{p}(k_\Delta)^{(r)}}; \widehat{p}(k_\Delta|j_v, i_{pa(v)})^{(r)}\right]}{\widehat{p}(i_v|i_{pa(v)})^{(r)} E\left[\frac{q(k_\Delta)}{\widehat{p}(k_\Delta)^{(r)}}; \widehat{p}(k_\Delta|i_v, i_{pa(v)})^{(r)}\right]}.
$$

Proof. Let $\Lambda_v = \Lambda \setminus fa(v)$.

$$
\begin{aligned}
&\widehat{p}(i_v|i_{pa(v)})^{(r+1)} \\
&= \frac{\widehat{p}(i_v, i_{pa(v)})^{(r+1)}}{\widehat{p}(i_{pa(v)})^{(r+1)}} \\
&= \sum_{k_\Delta}\sum_{k_{\Lambda_v}} \frac{\widehat{p}(i_v, i_{pa(v)}, k_\Delta, k_{\Lambda_v})^{(r)}}{\widehat{p}(k_\Delta)^{(r)}}\ q(k_\Delta) \Big/ \left[\sum_{k_v}\sum_{k_\Delta}\sum_{k_{\Lambda_v}} \frac{\widehat{p}(k_v, i_{pa(v)}, k_\Delta, k_{\Lambda_v})^{(r)}}{\widehat{p}(k_\Delta)^{(r)}}\ q(k_\Delta)\right] \\
&= \sum_{k_\Delta} \widehat{p}(i_v, i_{pa(v)}, k_\Delta)^{(r)} \widehat{g}_r(k_\Delta) \Big/ \left[\sum_{k_v}\sum_{k_\Delta} \widehat{p}(k_v, i_{pa(v)}, k_\Delta)^{(r)} \widehat{g}_r(k_\Delta)\right] \\
&= \widehat{p}(i_v, i_{pa(v)})^{(r)} \sum_{k_\Delta} \widehat{g}_r(k_\Delta) \widehat{p}(k_\Delta|i_v, i_{pa(v)})^{(r)} \Big/ \left[\sum_{k_v} \widehat{p}(k_v, i_{pa(v)})^{(r)} \sum_{k_\Delta} \widehat{g}_r(k_\Delta) \widehat{p}(k_\Delta|k_v, i_{pa(v)})^{(r)}\right] \\
&= p_{i|i}E_{ii} \big/ \big[p_{i|i}E_{ii} + p_{j|i}E_{ji}\big] \qquad \text{(since } v \text{ is binary)}, \qquad (15)
\end{aligned}
$$

where

$$\widehat{g}_r(k_\Delta) = \frac{q(k_\Delta)}{\widehat{p}(k_\Delta)^{(r)}}$$

and

$$E_{ij} = E\left[\frac{q(k_\Delta)}{\widehat{p}(k_\Delta)^{(r)}}; \widehat{p}(k_\Delta|i_v, j_{pa(v)})^{(r)}\right].$$

Applying the same argument yields

$$\widehat{p}(i_v|j_{pa(v)})^{(r+1)} = p_{i|j}E_{ij} \big/ \big[p_{i|j}E_{ij} + p_{j|j}E_{jj}\big]. \qquad (16)$$

From equations (15) and (16),

$$
\begin{aligned}
&\widehat{p}(i_v|i_{pa(v)})^{(r+1)} \big/ \widehat{p}(i_v|j_{pa(v)})^{(r+1)} \\
&= \big[1 + p_{j|j}E_{jj}/(p_{i|j}E_{ij})\big] \big/ \big[1 + p_{j|i}E_{ji}/(p_{i|i}E_{ii})\big] \\
&= \big[1 + (1+\eta)p_{j|j}E_{ji}/(p_{i|j}E_{ii})\big] \big/ \big[1 + p_{j|i}E_{ji}/(p_{i|i}E_{ii})\big] \qquad \text{(by condition (11))} \\
&> \big[1 + (1+\eta)p_{j|i}E_{ji}/(p_{i|i}E_{ii})\big] \big/ \big[1 + p_{j|i}E_{ji}/(p_{i|i}E_{ii})\big] \\
&\qquad \text{(since } p_{j|j}/p_{i|j} > p_{j|i}/p_{i|i} \text{ by condition (12))} \\
&= 1 + \eta\big[p_{j|i}E_{ji}/(p_{i|i}E_{ii})\big] \big/ \big[1 + p_{j|i}E_{ji}/(p_{i|i}E_{ii})\big]
\end{aligned}
$$

This completes the proof. □

It is worthwhile to note in the above two theorems that once the estimate of a probability is 0 or 1, this estimate remains the same throughout the subsequent EM procedure, which becomes apparent by Theorem 3 below.

4.2 After an M-step

At an M-step, we maximize the likelihood of a given model based on the estimates from the preceding E-step, and the resulting likelihood is given by

$$\prod_{v \in V} \hat{p}(i_v | i_{pa(v)})^{(r)},$$

where r is the iteration count at the preceding E-step. For a DAG of V, let $V_{(v)} = V \setminus de(v)$ for $v \in V$.

Lemma 1. *For a DAG of V*

$$\sum_{i_{de(v)}} \prod_{u \in V} \widehat{p}(i_u | i_{pa(u)}) = \prod_{u \in V_{(v)}} \widehat{p}(i_u | i_{pa(u)}). \tag{17}$$

Proof. See the Proposition 3.22 in Lauritzen (1996). □

Theorem 3. *For $v \in V$,*

$$\widehat{p}(i_v | i_{pa(v)})^{(r+1)} = \widehat{p}(i_v | i_{pa(v)})^{(r)}, \tag{18}$$

where the left-hand side is the estimate from an M-step and the right-hand side from the preceding E-step.

Proof.

$$
\begin{aligned}
&\widehat{p}(i_v|i_{pa(v)})^{(r+1)} \\
&= \frac{\widehat{p}(i_v, i_{pa(v)})^{(r+1)}}{\widehat{p}(i_{pa(v)})^{(r+1)}} \\
&= \Big[\sum_{i_{V\setminus fa(v)}} \prod_{u\in V} \widehat{p}(i_u|i_{pa(u)})^{(r)}\Big] \Big/ \Big[\sum_{i_{V\setminus pa(v)}} \prod_{u\in V} \widehat{p}(i_u|i_{pa(u)})^{(r)}\Big] \\
&= \Big[\sum_{i_{V_{(v)}\setminus fa(v)}} \Big(\sum_{i_{de(v)}} \prod_{u\in V} \widehat{p}(i_u|i_{pa(u)})^{(r)}\Big)\Big] \Big/ \Big[\sum_{i_{V_{(v)}\setminus pa(v)}} \Big(\sum_{i_{de(v)}} \prod_{u\in V} \widehat{p}(i_u|i_{pa(u)})^{(r)}\Big)\Big] \\
&= \Big[\sum_{i_{V_{(v)}\setminus fa(v)}} \prod_{u\in V_{(v)}} \widehat{p}(i_u|i_{pa(u)})^{(r)}\Big] \Big/ \Big[\sum_{i_{V_{(v)}\setminus pa(v)}} \prod_{u\in V_{(v)}} \widehat{p}(i_u|i_{pa(u)})^{(r)}\Big] \qquad \text{(by Lemma 1)} \\
&= \Big[\widehat{p}(i_v|i_{pa(v)})^{(r)} \sum_{i_{(V_{(v)}\setminus\{v\})\setminus pa(v)}} \prod_{u\in V_v\setminus\{v\}} \widehat{p}(i_u|i_{pa(u)})^{(r)}\Big] \Big/ \Big[\sum_{i_{(V_{(v)}\setminus\{v\})\setminus pa(v)}} \prod_{u\in V_v\setminus\{v\}} \widehat{p}(i_u|i_{pa(u)})^{(r)}\Big] \\
&\qquad\qquad \text{(since } v \text{ is terminal in } V_{(v)}\text{)} \\
&= \widehat{p}(i_v|i_{pa(v)})^{(r)}.
\end{aligned}
$$

□

An immediate corollary of this theorem is

Corollary 1. *If, for $v \in V$,*

$$\widehat{p}(i_v|i_{pa(v)})^{(r)} > \widehat{p}(j_v|j_{pa(v)})^{(r)},$$

where the estimates are from an E-step, $i_v, j_v \in \mathcal{I}_v$, and $i_{pa(v)}, j_{pa(v)} \in \mathcal{I}_{pa(v)}$, then

$$\widehat{p}(i_v|i_{pa(v)})^{(r+1)} > \widehat{p}(j_v|j_{pa(v)})^{(r+1)}, \tag{19}$$

where the estimates are from the subsequent M-step.

According to Theorems 1 and 2, we can see that the rank order of the estimates of $P(i_v|i_{pa(v)})$ for a node v becomes stable after a certain number (N) of EM iterations, where N may depend upon the size and complexity of the model. As is shown in Theorem 3, the M-step preserves the current rank order of the estimates. So, it is the E-step that affects the rank order.

As the estimates in expressions (4) and (11) converge, the value of η will approach zero as the iteration count r increases. This means that, for each node v, the rank order of the

estimates of $p(i_v|i_{pa(v)})$ becomes less likely to change as the EM procedure proceeds. This is a general phenomenon for every node in a model, whether the node is terminal (see Theorem 1) or non-terminal (see Theorem 2).

If the rank order of the estimates for a node were different from that of the initial estimates, it most probably means that the rank order has changed at least once at an early stage of the EM procedure. This is illustrated in next section using both real and simulated data. Thus it is recommended that the EM procedure is monitored periodically and check for any possible distortion of the rank order. If any distortion is detected, then it is desirable that we try a new set of initial values for the EM.

The fluctuation of estimates can be controlled to some level by applying a Bayes method (Chapter 12 of Bishop, Fienberg, and Holland (1975)) at the M-step. Since the variables are all categorical, we impose, for a node $v \in V$, a Dirichlet prior on the $p(i_v|i_{pa(v)})$'s for each $i_{pa(v)}$. For example, if X_v is of l levels, we can think of a Dirichlet prior with parameters $\alpha_1, \cdots, \alpha_l$, and the resulting posterior distribution of $(p(i_v|i_{pa(v)}))_{i_v \in \mathcal{I}_v}$ is a Dirichlet distribution with parameters $n\hat{p}(X_v = 1, i_{pa(v)})^{(r)} + \alpha_1, \cdots, n\hat{p}(X_v = l, i_{pa(v)})^{(r)} + \alpha_l$, where we set $\mathcal{I}_v = \{1, 2, \cdots, l\}$. That is , the posterior means of $p(i_v|i_{pa(v)})$ are given by

$$E(p(i_v|i_{pa(v)})|n(i),\ i \in \mathcal{I}) = \frac{n(i_{fa(v)}) + \alpha_{i_v}}{n(i_{pa(v)}) + \sum_{j \in \mathcal{I}_v} \alpha_j}. \tag{20}$$

Of course, the Dirichlet priors could vary across the variables of a model. The estimates that are obtained at an M-step are calibrated to data at an E-step and the Bayes method has nothing to do with the E-step.

5 EMPIRICAL RESULTS

We analyzed a data set of 7 multiple choice items of the Mathematics section of the Korean SAT that was administered in 1999. After investigating the 7 items, we ended up with eight knowledge units or cognitive features that are relevant to the 7 test items. This means we have 7 observed binary variables for item scores (0 for an incorrect answer and 1 for a correct answer) and 7 unobservable binary variables for the states of the knowledge units (0 for a poor state of knowledge and 1 for a good enough state).

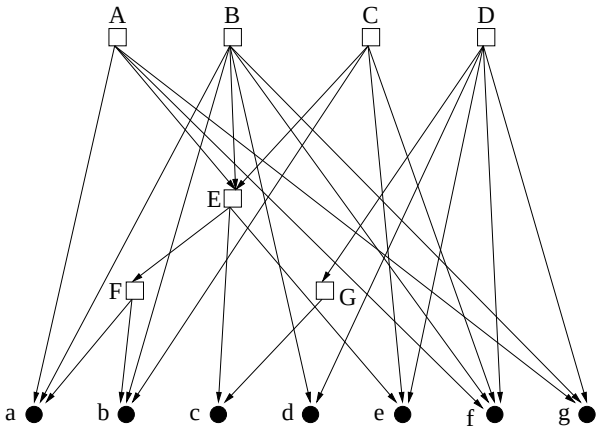

Figure 1: A DAG for real data. Bullets are for the item score variables and boxes for the knowledge states.

The 14 variables are related as in Figure 1, where an arrow from a box to a bullet stands for a causal relation between the corresponding knowledge unit and the test item and an arrow from a box to a box mostly stands for a prerequisite relationship between the corresponding pair of knowledge units (Mislevy, 1994). If an item can be solved when a test-taker possesses a good knowledge of certain knowledge units, then the item-score variable is said to be causally related to the knowledge units and arrows run from the corresponding knowledge-state variables to the item-score variable. The structure of the relationship among the item-score variables and the knowledge-state variables is based on the opinions of a group of experts of the test subject. The knowledge units are listed in Table 1.

As a whole, we need to deal with 14 binary variables or $2^{14} = 16,384$ cell probabilities for parameter estimation of the model as in Figure 1. The parameters are given in the form of marginal or conditional probability. As for the item-score variable e in the figure, knowledge units C, D, and E are relevant. In other words, $pa(e) = \{C, D, E\}$. So we need to estimate $P(X_e = 1|X_{C,D,E} = i_{pa(e)})$. We can expect that the conditional probabilities are all positive since the choice of a correct answer is possible even by a lucky guess on an item when the item is too much for a test-taker. The sample size of the data we used is 150,799 and the stopping threshold for the EM method is 0.01 for this data.

To show how the estimates fluctuate during the EM process, we display the values of $\widehat{P}(X_b = 1|i_{pa(b)})^{(r)}$ and $\widehat{P}(X_g = 1|i_{pa(g)})^{(r)}$ in Figure 2. For convenience' sake, we denote by

Table 1: The list of the knowledge units involved in the model in Figure 1

Code	Contents
A	DK about sets
B	DK about numbers and equations
C	DK about plane geometry
D	PK for inference
E	DK about one-variable functions
F	DK about trigonometrical functions
G	PK for problem recognition

NOTE: DK is an acronym of "declarative knowledge" and PK of "procedural knowledge."

Table 2: The probabilities, $P(X_v = 1|i_{pa(v)})$, used for the simulated data of Section 5.

$\|pa(v)\|$	$\|i_{pa(v)}\|$				
	0	1	2	3	4
0*	0.55				
1	0.15	0.85			
2	0.10	0.15	0.85		
3	0.10	0.15	0.25	0.85	
4	0.10	0.15	0.25	0.35	0.85

(*): This corresponds to a root node.

$p0, p1, \cdots, p7$, respectively, the values of $\widehat{P}(X_b = 1|i_{pa(b)})$, in the order of the configurations of $i_{pa(b)}$ from (0,0,0) to (1,1,1), in Figure 2, and analogously for $\widehat{P}(X_g = 1|i_{pa(g)})^{(r)}$. The left column of the figure is for X_b and the right column for X_g.

We can see in the figure that the estimates from the real data (see the first row of the figure) fluctuate much more than those from simulated data (the second row), where the simulated data are generated as follows. The graph in Figure 1 is the structure of a recursive model and so the simulated data can be generated by generating values for individual variables following the direction of the arrows in the graph using the (conditional) probabilities as listed in Table 2. For instance, we use the probability values in the row of $|pa(v)| = 2$ for generating a value of X_c. The values are generated in the order of $X_A, X_B, X_C, X_D, X_E, X_F, X_G, X_a, X_b, \cdots, X_g$. One observation consists of the 14 values of these variables. We repeat this process to obtain as many observations as we want. We may use the probability values other than those in Table 2 provided that the CPA condition is satisfied.

Table 3: Estimates of $P(X_b = 1|i_{pa(b)})$ with the sample size 150,799. The stopping threshold of the EM is 0.01. The α and β in the table are parameters of a beta prior.

	i_B	i_C	i_F	real	simulated	Bayes method using real data with $\alpha+\beta=$			
				data	data	10	30	50	100
p0	0	0	0	0.1023	0.0008	0.0788	0.0301	0.0314	0.1071
p1	0	0	1	1.0000	0.7854	0.2383	0.2774	0.2631	0.2657
p2	0	1	0	0.1665	0.2458	0.2457	0.2156	0.1950	0.3418
p3	0	1	1	0.5885	1.0000	0.0091	0.0156	0.0270	0.6173
p4	1	0	0	0.2090	0.1637	0.0430	0.1893	0.1918	0.4496
p5	1	0	1	1.0000	0.1435	0.9820	0.9165	0.9115	0.9686
p6	1	1	0	1.0000	0.1735	0.1359	0.2428	0.1809	0.7399
p7	1	1	1	0.9082	0.8544	0.9252	0.9720	0.9652	0.9122
OD^*				4	4	6	4	5	1

(*): OD denotes the number of the pair of the estimates for which condition (1) is not satisfied. The maximum value of OD is 19 for $P(X_b = 1|i_{pa(b)})$.

In the first two rows in the figure, we can see a wild fluctuation of estimates at an early stage of the EM process. In particular, the estimates $p2$ and $p6$ fluctuate over a wide range of values. This fluctuation may be influenced by the initial values or by the model structure we choose.

While we see a wild fluctuation of estimates when they are based on real or simulated data, the fluctuation is reduced when we apply a Bayes method (see the last row of Figure 2). Since all the variables in the model in Figure 1 are binary, we imposed a Beta prior with its parameters α and β, $\alpha + \beta = 100$, on every variable in the model. After a little bit of fluctuation at the beginning of the EM process, the estimates change gradually and the rank order of the estimates remains almost the same throughout the EM process.

The final estimates of $P(X_b = 1|i_{pa(b)})$ are given in Table 3 and those of $P(X_g = 1|i_{pa(g)})$ in Table 4. In the last row of each of the tables are the numbers of pairs of the estimates for which condition (1) is not satisfied. We denote the number for each variable by OD (think of *O*rder *D*istortion). The maximum value of the OD for a given variable depends upon the number of configurations of its conditional variables. The maximum value can be found simply by counting the pair of configurations that are ordered in the sense of "$\preceq$" which is defined

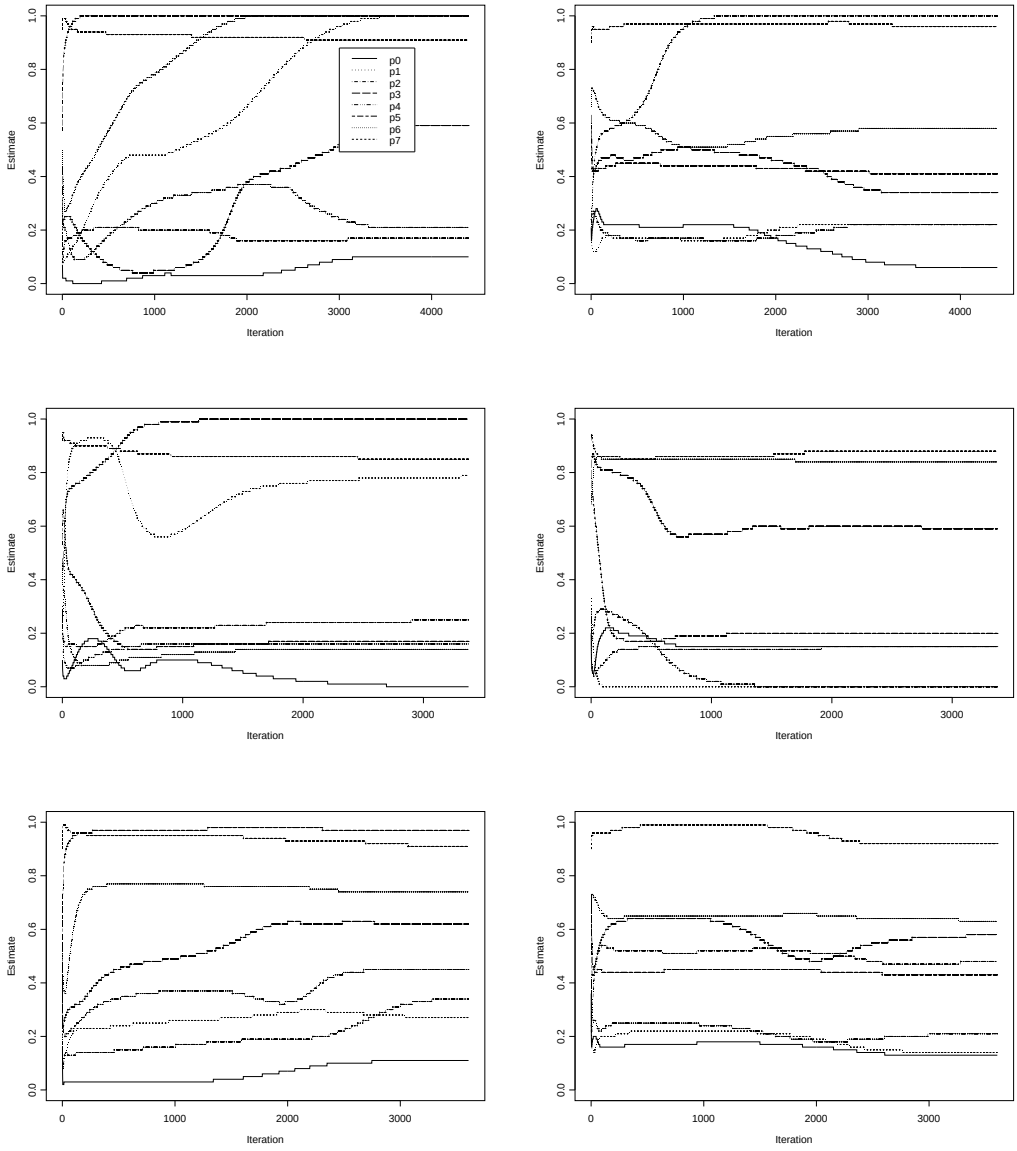

Figure 2: A time-series-type display of $\widehat{P}(X_b = 1|i_{pa(b)})^{(r)}$ and $\widehat{P}(X_g = 1|i_{pa(g)})^{(r)}$. $\widehat{P}(X_b = 1|i_{pa(b)})^{(r)}$-values are displayed in the first column and $\widehat{P}(X_g = 1|i_{pa(g)})^{(r)}$-values in the second column, where $p0, p1, \cdots, p7$ are short labels of the $\hat{P}$ values in ascending order of the conditional random vector starting from (0,0,0) through (1,1,1). The estimates from an EM based on real data are displayed in the 1st row, the estimates from an EM based on simulated data in the second row, and the estimates from a Bayesian EM based on real data in the last row.

in the paragraph containing expression (1). For example, if a variable has two binary, parent variables, then the maximum of OD is 5, and if three binary, parent variables are involved,

Table 4: Estimates of $P(X_g = 1|i_{pa(g)})$ with the sample size 150,799. The stopping threshold, OD, and (α, β) are explained in Table 3.

	i_A	i_B	i_D	real	simulated	Bayes method using real data with $\alpha+\beta=$			
				data	data	10	30	50	100
p0	0	0	0	0.0557	0.1459	0.1873	0.0647	0.1032	0.1321
p1	0	0	1	0.2225	0.0000	0.0397	0.1735	0.1458	0.1446
p2	0	1	0	1.0000	0.0016	0.8752	0.6309	0.5676	0.4766
p3	0	1	1	0.3378	0.5856	0.4328	0.5128	0.4823	0.5773
p4	1	0	0	0.2228	0.1462	0.1573	0.1624	0.1599	0.2108
p5	1	0	1	0.5837	0.8442	0.4389	0.4235	0.4191	0.4279
p6	1	1	0	0.4118	0.1992	0.5625	0.5442	0.5534	0.6349
p7	1	1	1	0.9585	0.8808	0.9078	0.9079	0.8934	0.9180
OD				3	2	4	2	2	0

then the maximum is 19.

The estimates based on the real data are in the 5th column of Table 3, the estimates based on simulated data in the 6th column, and the estimates from a Bayes method using the real data are in the last four columns of the table. If we look at the OD values, we can see that they decrease as $\alpha+\beta$ increases with the Bayes method. Although some OD values from the Bayes method are larger than the OD values from the non-Bayes method, some of the estimates are either 0 or 1 while there are no such extreme values in the estimates from the Bayes method. The same phenomena are observed in Table 4.

The values of Pearson's chi-squared statistic for the model in Figure 1 are all large as shown in Table 5 whether they are based on the real data or the simulated data. This is mostly due to the large sample size (=150,799). The corresponding p-values of the goodness-of-fit tests are all near zero. Note that the p-value is so small (0.00017) even for the simulated data. The influence of the sample size upon the goodness-of-fit test is explored in Marsh et al. (1988) and Hu and Bentler (1999). They compared several alternatives of the Pearson's chi-squared statistic (X^2) for goodness-of-fit testing and suggested that we use these alternatives along with X^2 for a sound conclusion under the influence of the sample-size.

To see how the goodness-of-fit level compares between the Bayes method and non-Bayes methods, we subsampled a sample of size 1,000 from the original data of size 150,799. The estimation results are summarized in Table 6. Note that the p-value of a goodness-of-fit test is

Table 5: Values of the Pearson's chi-squared statistic. The p-values of the goodness-of-fit test for the values in the table are all zero except for the simulated data for which the p-value is 0.00017. The α and β in the table are parameters of a beta prior.

d.f.(73)	real	simulated	Bayes method using real data with $\alpha+\beta=$			
	data	data	10	30	50	100
χ^2	641.693	121.241	821.802	897.783	1180.95	1236.85

NOTE: Sample size = 150,799; stopping threshold of the EM= 0.01.

Table 6: Values of the Pearson's chi-squared statistic. The sample size is 1,000 and the stopping threshold of the EM is 0.001. The α and β in the table are the parameters of a beta prior.

d.f.(73)	real	simulated	Bayes method using real data with $\alpha+\beta=$			
	data	data	1	2	3	5
χ^2	63.143	51.692	73.701	86.153	94.227	100.845
P-value	0.7881	0.9723	0.4548	0.1392	0.0480	0.0171

a measure of appropriateness of an interested model, a larger p-value indicating a higher level of appropriateness. The p-values indicate that the model in Figure 1 is appropriate and that the p-value decreases as $\alpha+\beta$ of the beta prior increases. This is a natural consequence since, as indicated in expression (20), a large value of $\alpha+\beta$ tends to pull the estimates towards the value that the Beta prior calls for.

Although the p-values of the goodness-of-fit test are larger for the estimates by a non-Bayes method than the p-values from a Bayes method, we can see, in Tables 7 and 8, that some of the estimates from the non-Bayes method are either 0 or 1, while there is no such value in the estimates from the Bayes method. For our data set, the extreme values, 0 and 1, are not appropriate as estimates because lucky guessing and making a mistake for some unknown reason are always possible in taking a test which consists of multiple choice problems.

In addition to that, the estimation results indicate that the OD values are comparable between the non-Bayes method and the Bayes method. In particular, when the sample size is 150,799, the OD value is either 0 or 1 with $\alpha+\beta=100$ (see Tables 3 and 4), and when the sample size is 1,000, it is either 0 or 2 with $\alpha+\beta=2$ (see Tables 7 and 8). Note that the p-value of the goodness-of-fit test is 0.14 (see Table 6) when $\alpha+\beta=2$. Considering the

Table 7: Estimates of $P(X_b = 1|i_{pa(b)})$ based on data of size 1,000. The stopping threshold is 0.001, and OD and (α, β) are explained in Table 3.

	i_B	i_C	i_F	real	simulated	Bayes method using real data with $\alpha+\beta=$			
				data	data	1	2	3	5
p0	0	0	0	0.0000	0.0000	0.0657	0.0769	0.1086	0.1079
p1	0	0	1	0.0000	0.2871	0.1352	0.1627	0.1781	0.1683
p2	0	1	0	0.1170	0.7648	0.0471	0.1214	0.1349	0.1302
p3	0	1	1	0.5534	0.3396	0.4140	0.3511	0.2912	0.2600
p4	1	0	0	0.0000	0.0584	0.2713	0.2073	0.1752	0.1721
p5	1	0	1	0.6950	0.3250	0.3855	0.2934	0.2508	0.2584
p6	1	1	0	0.0410	1.0000	0.0987	0.1194	0.1323	0.1417
p7	1	1	1	1.0000	0.3214	0.9695	0.9611	0.9563	0.9471
OD				1	4	2	2	1	1

Table 8: Estimates of $P(X_g = 1|i_{pa(g)})$ based on data of size 1,000. The stopping threshold of the EM is 0.001, and OD and (α, β) are explained in Table 3.

	i_A	i_B	i_D	real	simulated	Bayes method using real data with $\alpha+\beta=$			
				data	data	1	2	3	5
p0	0	0	0	0.1724	0.0000	0.0643	0.0679	0.0763	0.0787
p1	0	0	1	0.0000	0.5221	0.0776	0.0920	0.1055	0.1072
p2	0	1	0	0.2515	0.4378	0.2173	0.1787	0.1368	0.1327
p3	0	1	1	0.3592	0.2285	0.2242	0.2793	0.3077	0.2818
p4	1	0	0	0.2697	0.0000	0.2212	0.1721	0.1755	0.1390
p5	1	0	1	0.6556	0.3208	0.5095	0.5145	0.5303	0.5216
p6	1	1	0	0.4976	0.4964	0.4554	0.4239	0.4148	0.4008
p7	1	1	1	0.9575	0.5967	0.9120	0.9049	0.9013	0.8919
OD				1	3	0	0	0	0

ratio of $\alpha + \beta = 2$ to the sample size 1,000, $\alpha + \beta$ may be as large as 300 when the sample size is 150,799. It is important to note that the OD is worthwhile to look at provided that no extreme values such as 0 or 1 are found in the estimates. For instance, we see, in Table 7, that the OD for the real data by the non-Bayes EM and that by the Bayes EM with $\alpha + \beta = 2$ are 1 and 2, respectively. But the estimates which are calculated by the non-Bayes EM contain 4 extreme values while there are no such extreme values in the estimates which are calculated by the Bayes EM. The OD in the former set of estimates is thus of little use.

In applying the Bayes EM method, we assigned values to α and β in accordance with

Table 9: Ratios $\alpha/(\alpha+\beta)$ of Beta(α,β) priors on $P(X_v = 1|i_{pa(v)})$ that are used in the Bayes method. $|i_{pa(v)}|$ is the number of 1's in $i_{pa(v)}$.

$\|pa(v)\|$	$\|i_{pa(v)}\|$	$\alpha/(\alpha+\beta)$
1	0	0.1
	1	0.85
2	0	0.1
	1	0.2
	2	0.85
3	0	0.1
	1	0.15
	2	0.25
	3	0.85

suggestions of a group of psychometricians. Let i be a vector of 1's and 0's and $|i|$ be the number of 1's in i. Assuming all the variables are binary, the Beta priors on $P(X_v = 1|i_{pa(v)})$ for a variable X_v were as in Table 9. The values in the table reflect the facts that a lucky guess can lead to a correct answer to a multiple-choice problem and that a good enough knowledge for a problem does not always lead to a correct answer. We may try different values of α and β under the condition that we use larger values of $\alpha/(\alpha+\beta)$ for larger values of $|i_{pa(v)}|$, which is similar to the CPA condition in the sense that $\alpha/(\alpha+\beta)$ is the mean of the Beta prior (with parameters α and β) of $P(X_v = 1|i_{pa(v)})$. Appropriate α and β values can be found by trying several different values of them and choosing the values for which the p-value of the goodness-of-fit test is not smaller than a specified value (e.g., 0.05 or 0.1) and the OD values are near zero for all the variables provided that the estimates contain no or few extreme values.

The estimation results based on the real and the simulated data indicate that the Bayes EM is more likely to produce estimates that satisfy condition (1) than non-Bayes EM methods and that there is no fluctuation in the estimates during the Bayes EM process except during the first few iterations.

6 CONCLUDING REMARKS

The estimates in the EM procedure were unstable during an early period of the procedure, when we did not use a Bayes method, and we saw a wild fluctuation of estimates during that

period, more serious with real data and less with simulated data. The fluctuation seems to be manageable by applying a Bayes method with appropriate priors. This means that, when the CPA (i.e., condition (1)) is assumed for the probability model, the quality of the estimates of the probabilities is determined at an early stage of the EM procedure. Therefore, it is effective to check, at an early stage of the EM process, if condition (1) is violated. If such is the case, another set of initial values may have to be used or, if a Bayes method is applied, another set of parameter values may have to be used on a given prior distribution.

It is shown in section 4 that a violation of the CPA condition takes place, if at all, at an E-step where the rank order of the new estimates for a variable can possibly change from that of the previous estimates. The theorems in section 4 concerning an E-step are rooted in the fact that the EM process converges. Under the CPA assumption, however, the convergence of estimates does not make sense if the CPA condition was already violated at an early stage of the EM process and the violation remained throughout the process. Once the CPA condition is violated at an early stage of the EM, there is no knowing where the estimates will lead to. The empirical results suggest that, when all the variables involved in a given model are binary, the Bayes EM with appropriate values of α and β for the Beta prior would produce estimates that satisfy the CPA condition with an acceptable level of the goodness-of-fit test. We can anticipate an analogous result for any categorical data with the Beta prior replaced with a Dirichlet prior.

References

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B*, **39**, 1-38.

Van Dyk, D. and Meng, X. L. (1997). On the ordering and groupings of conditional maximizations within ECM-type algorithms. *J. Comput. Graph. Statist.*, **6**(2), 202-223.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation, *J. R.*

Statist. Soc. B, **52**(3), 443-452.

Haertel, E. H. and Wiley, D. E. (1992). Representations of ability structures: Implications for testing. In N. Fredericksen, R. J. Mislevy, and I. I. Bejar (Eds.) *Test theory for a new generation of tests* (pp. 359-384). Hillsdale, NJ.: Erlbaum.

Holland, P.W. and Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, **14**(4), 1523-1543

Hu, L.-T. and Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modelling*, **6** (1), 1-55.

Junker, B. W. and Ellis, J.L. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics*, **25**(3), 1327-1343.

Kim, S. -H. (2002). Calibrated initials for an EM applied to recursive models of categorical variables, *Computational Statistics and Data Analysis*, **40**, 97-110.

Lauritzen, S. L. (1996). *Graphical Models*, Clarendon Press, Oxford.

Little, R. J. A. and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Marsh, H.W., Balla, J.R., and McDonald, R.P. (1988), Goodness of fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, **103**, 391-410.

Mislevy, R. J. (1994). Evidence and inference in educational assessment, *Psychometrika*, **59**, 439-483.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in export systems, *Statistical science*, **8**(3), 219-283.

Thiesson, B. (1991). (G)EM algorithms for maximum likelihood in recursive graphical association models, Master's thesis, Dept. Mathematics and Computer Science, Aalborg Univ.

Wermuth, N. and Lauritzen, S.L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, **70**(3) 537-552.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York, NY.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**(1), 95-103.