The Korean Communications
in Statistics Vol. 7, No. 1, 2000
pp. 65-85

An Improvement on Estimation for Causal Models of Categorical Variables of Abilities and Task Performance

Sung-Ho Kim[1)]

ABSTRACT

The estimates from an EM, when it is applied to a large causal model of 10 or more categorical variables, are often subject to the initial values for the estimates. This phenomenon becomes more serious as the model structure becomes more complicated involving more variables. In this regard, Wu (1983) recommends among others that EMs are implemented several times with different sets of initial values to obtain more appropriate estimates. In this paper, a new approach for initial values is proposed. The main idea is that we use initials that are calibrated to data. A simulation result strongly indicates that the calibrated initials give rise to the estimates that are far closer to the true values than the initials that are not calibrated.

Key words: Bayes estimation, calibrated initial values, discrepancy, Dirichlet prior, EM, interval of imprecision, order distortion, randomly selected initial values, recursive models, subjective probability.

1. Introduction

EM (Dempster et al., 1977) is a most popular method to estimate parameters of a model which involves latent variables. It is easy to understand and the algorithm consists of two operations, expectation for the missing variables and likelihood-maximization. Literature abounds about the EM concerning the issues of applications, convergence rates, and a variety of improved versions of it (see, for example, van Dyk and Meng (1997)).

This paper confines attention to EM algorithms for recursive models (Lauritzen and Wermuth, 1983) of categorical variables only. Recursive models of categorical

1) The author wishes to acknowledge the financial support of the Korean Research Foundation made in the program year of 1998.

Division of Applied Mathematics, Korea Advanced Institute of Science and Technology, Daejon, South Korea.

variables pertain to an exponential family and so the maximum likelihood (ML) estimates for such models are easy to obtain when data are complete and so is the corresponding EM. But when the model is large and complicated, we often see that estimates from an EM are sensitive to the initial values of the estimates (see for example Wu (1983)). For this reason, we would often apply an EM several times with different sets of randomly selected
initials and select the best-looking estimates out of the collection of the EM outputs.

When the relationship among the variables involved in a model is causal, we usually call such a model a causal model. When the relationship is causal, the corresponding model structure is recursive. So I will use both terms in this paper and call a model recursive when its structure is stressed and causal when the relationship is stressed.

In this paper, I will propose a selection method of initial values by applying the notion of calibration, which will be described in detail in a later section, and show, by a simulation experiment, that the initial values by the method yield estimates that are far closer to the true values than randomly selected initial values. A main idea behind the selection method is that we select initial values so that their marginals on the observed variables are equal to data.

This paper consists of 8 sections. Section 2 presents graphical terminologies that will be used in the paper, and section 3 describes a basis of the selection method of initial values that will be proposed in this paper. Section 4 then builds the main result of the paper upon the basis of section 3. Sections 5 and 6 present simulation results, calibrated initials being used in the former section and randomly selected initials in the latter. The results strongly recommend using calibrated initials. Section 7 turns our attention to Bayes estimation which may help make the final estimates more appropriate for a particular variable, and section 8 concludes the paper with some discussions.

2. Directed Acyclic Graphs and Terminologies

The relationship among the set of variables that are involved in a recursive model can be represented by a directed acyclic graph (DAG). An example of DAG is displayed in Figure 1. The circles and boxes in the figure have a particular meaning and it will be described in a later section. As illustrated in the figure, a DAG consists of nodes and arrows (or directed edges). $a \rightarrow b$ stands for 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 $fa(v) = v \cup pa(v)$. The 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.

We will denote by V the index set of the variables that are involved in a given model. We will denote by i_A the cell entry of the contingency table of the variables indexed in A, by I_A the collection of all the possible i_A's, by i the cell entry of the contingency table of all the variables involved in a given model, by $m_A(i_A)$ the cell mean at the cell entry i_A of the contingency table of the variables indexed in A, and we will use n instead of m to represent the observed frequency. We will denote by δ the index set of the observed variables. Estimates will be hatted on the corresponding parameters. *Cell entry* will also be called *configuration.*

3. Initial Values and Subjective Probabilities

We may use a model where all the variables involved are independent and generate random numbers to obtain initial values for an EM no matter what the level of complexity of a given recursive model is. But in practice, such a random number approach would take a longer time until convergence and yield estimates that look inappropriate. A rule of thumb of the selection method to be proposed in this paper is that we pick the initial values for the ML estimates so that the intrinsic relations among the variables may be consistently incorporated with observed data. By consistency I mean that the observed frequencies for the observed variables and experts' opinions on the relations among the variables that are expressed in terms of subjective (conditional) probabilities (Savage, 1972; Kyberg and Smokler, 1980)are combined into a system of equations as in

$$r(X=x) = \sum_{i \in I_{pa(X)}} \psi(x|pa(X)=i)\,\psi(pa(X)=i), \tag{1}$$

where $r(X=x)$ denotes the relative frequency that the observable X takes on the value x, $\psi(x|pa(X)=i)$ denotes the subjective probability that $X=x$ given that $pa(X)=i$, $\psi(pa(X)=i)$ denotes the subjective probability that $pa(X)=i$, and the summation goes over all the possible configurations of $pa(X)$.

In practice, $\psi(x|pa(X)=i)$ is determined relatively more easily than $\psi(pa(X)=i)$. For instance, the conditional probability that a student gives a correct answer to a

test item conditional on that the student has a certain state of knowledge may be relatively easier to guess than the marginal probability that the student has a certain state of knowledge. In the next section we will see how we can use the initial subjective conditional probabilities and data and obtain initial values for the marginal probabilities for the root nodes of a model.

It is reasonable to assume that a better state of knowledge corresponds to the probability of a correct response to a test item which is at least as high as that corresponding to a poorer state of knowledge. We will say that an *order distortion* (OD for short) occurs when the estimate of the probability of a correct response to a certain item is higher for a poorer state of knowledge than for a better state of knowledge. When the initial values were selected at random, we would often see the OD phenomenon in the ML estimates.

Jeong et al. (1998) showed, using simple model structures, an empirical result that when we use randomly selected initial values, we may see the undesirable OD phenomenon far more often than when we select the initial values incorporating subjective conditional probabilities and data in a consistent manner. While Jeong et al. considered simple model structures, we will not put a limit on the structure of a causal model in this paper and derive a general approach for selecting initial values that incorporates subjective conditional probabilities and data in a consistent manner.

4. The Notion of Calibrated Initial Values

We will begin this section by looking into the geometry of EM as applied to recursive models of categorical variables. The E-step is implemented through the expression

$$\hat{m}_V(i)^{(r+1)} = n_\delta(i_\delta)\frac{\hat{m}_V(i)^{(r)}}{\hat{m}_\delta(i_\delta)^{(r)}}$$

Once carried out, the new estimates $\hat{m}_V^{(r+1)}$ satisfy that

$$\hat{m}_\delta(i_\delta)^{(r+1)} = n_\delta(i_\delta). \tag{2}$$

That is, when the new estimates are marginalized on δ, the marginals are the same as n_δ. Assuming that V is of K variables, we can see that (2) means geometrically that the points $(i, \hat{m}(i)^{(r+1)})$ lie in the hyperplane $\mathcal{H}_1$ given by

$\mathcal{H}_1 = \{(i, m(i)); m_\delta(i_\delta) = n_\delta(i_\delta)$ for all possible configurations $i_\delta\}$.

On the other hand, the M-step is implemented through the expression

$$\hat{m}(i)^{(r+1)} = n \prod_{v \in V} \hat{p}(i_v | i_{pa(v)})^{(r)}, \tag{3}$$

where

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

The right-hand side of (3) represents the model structure for V, which is also representable in terms of conditional independence (Dawid, 1979; Pearl, 1988; Whittaker, 1990). A set of independence relationships among V defines a hyperplane for $(i, m(i))$. A good example is given in section 2.7 of Bishop, Fienberg, and Holland (1975), where a hyperplane of independence of two binary variables is displayed.

We will denote the hyperplane defined by a given model structure by $\mathcal{H}_2$. Then the points $(i, \hat{m}(i)^{(r+1)})$ obtained by (3) must lie in $\mathcal{H}_2$. Therefore, the final estimates from an EM must be found in $\mathcal{H}_1 \cap \mathcal{H}_2$.

The EM problem is an optimization problem for the likelihood function where the domain of the likelihood function is confined to $\mathcal{H}_1 \cap \mathcal{H}_2$. When $\mathcal{H}_1 \cap \mathcal{H}_2$ is of dimension 4 or higher, it is hard to see if a set of final estimates from an EM corresponds to the global maximum point of the likelihood function or a local maximum point. The key idea behind the selection method to be proposed in this paper is that the initial values, $\{\hat{m}(i)^{(0)}\}_{i \in I_V}$, be selected so that the point $(\hat{m}(i)^{(0)}, i \in I_V)$ may be contained at least in $\mathcal{H}_1$.

The initial estimates, whether they are obtained from experts' opinions or not, may not be contained in $\mathcal{H}_1$. When the initial estimates are from a group of experts and not contained in $\mathcal{H}_1$, it is like we begin an EM process with a set of initial values that may be meaningless to the experts. In this respect, it is desirable that the initial values, $\{\hat{m}(i)^{(0)}\}_{i \in I_V}$, look reasonable to experts and that

$$(\hat{m}(i)^{(0)}, i \in I_V) \in \mathcal{H}_1 \tag{4}$$

After an E-step, $\hat{m}^{(r)}$ will be in $\mathcal{H}_1$, and after an M-step, it will lie in $\mathcal{H}_2$ but may not be in $\mathcal{H}_1 \cap \mathcal{H}_2$. The estimates may stay in $\mathcal{H}_1 \cap \mathcal{H}_2$ after some number of iterations of the E and M-steps.

Once $\hat{m}^{(r)} \in \mathcal{H}_1 \cap \mathcal{H}_2$, we have

$$\hat{m}^{(r+i)} = \hat{m}^{(r)}, \quad i = 1, 2, \cdots.$$

As a matter of fact, we stop an EM process when $\hat{m}^{(r+2)}$ from an M-step falls

within some neighborhood of $\hat{m}^{(r)}$ from the preceeding M-step. This implies that the final estimate $\hat{m}^{(r+2)}$ from an EM, unless $\hat{m}^{(r+2)} = \hat{m}^{(r)}$, is not in $\mathcal{H}_1 \cap \mathcal{H}_2$ but located at a point in $\mathcal{H}_2$ which lies close to $\mathcal{H}_1 \cap \mathcal{H}_2$. This observation suggests (4) as a desirable property of initial values, and the simulation results of section 5 are strongly in favor of (4).

Expression (4) is achieved by the process described below, where we assume that only the variables of the terminal nodes are observable:

(a) Determine $\phi(v|i_{pa(v)})$ for every non-root node v.

(b) Assign arbitrary values in $(0,1)$ for the marginal probabilities of the root nodes of the model structure.

(c) For each root node v', obtain the mean of the conditional probabilities $\{p(v'|y)\}_{y \in \mathcal{Y}}$, where $\mathcal{Y}$ is the set of observed vector-values y for the vector of the terminal nodes of the model structure.

(d) Replace the arbitrary values assigned in (b) for the root nodes with the corresponding means as obtained in (c). This replacement along with the subjective probabilities given in (a) yields a set of initial values of the estimates that satisfy (4), as will be proved in Theorem 1.

Step (d) completes the process for generating the initial values, $\{\hat{m}(i)^{(0)}\}_{i \in I_V}$. We will now see why these values satisfy (4). Consider a simple model M of two categorical variables A and X which are dependent upon each other. Suppose that A is latent and X is observable, and denote the observations by $x_1, x_2, \cdots, x_N$. We may assume that we have a set of subjective probabilities of X given A and choose arbitrary values in $(0,1)$ for the marginals of A. For observation x_n, we can obtain, using the assumed subjective probabilities and the arbitrary values for the marginals, the conditional probability $P(A=i_A|X=x_n)$, for $i_A \in I_A$. Denoting the mean of the conditional probabilities by $\dot{r}(A=i_A)$ for each i_A, we have

$$\dot{r}(A=i_A) = \sum_{i_X \in I_X} P(A=i_A|X=i_X) r(X=i_X), \tag{5}$$

where $r(X=i_X)$ is the relative frequency that $X=i_X$ out of the N observations for each $i_X \in I_X$. The following result is an immediate application of (5).

Theorem 1 *Consider two categorical variables* A *and* X, *where* A *is latent and* X *is observable. Suppose that the conditional probabilities of* X *given* A *are known and that the data,* $x_1, x_2, \cdots, x_N$, *are given for* X. *Then, for a set of arbitrary values in* $(0,1)$ *for the marginals of* A, *there exists a marginal of* A *(denoted by* $\hat{r}(A=i_A)$*) as given by (5) that satisfies*

$$\sum_{i_A \in I_A} \hat{r}(A=i_A) P(X=i_X|A=i_A) = r(X=i_X), \text{ for } i_X \in I_X. \tag{6}$$

Proof For a particular configuration i_X of X, we have

$$\begin{aligned}
\sum_{i_A} \hat{r}(A=i_A)\ & P(X=i_X|A=i_A) \\
&= \sum_{i_A} \Big(\sum_{i_X^* \in I_X} P(A=i_A|X=i_X^*) r(X=i_X^*) \Big) P(X=i_X|A=i_A) \\
&= \sum_{i_X^* \in I_X} \Big(\sum_{i_A} P(X=i_X|A=i_A) P(A=i_A|X=i_X^*) \Big) r(X=i_X^*) \\
&= \sum_{i_X^* \in I_X} E(P(X=i_X|A)|X=i_X^*) r(X=i_X^*) \\
&= r(X=i_X).
\end{aligned}$$

The first equality in the proof follows from (5) and the last equality holds since

$$E(P(X=i_X|A)|X=i_X^*) = \begin{cases} 0 & \text{when } i_X \neq i_X^* \\ 1 & \text{otherwise.} \end{cases}$$

□

When the probabilities $\hat{r}(A=i_A)$ as in (6) are used for A, we will say that the initial values, $\{\hat{m}(i)^{(0)}\}_{i \in I_V}$, are *calibrated to data* (or *calibrated* for short).

We can extend the theorem to a recursive model of K latent variables, $A_1, \cdots, A_K$, and J observable variables, $X_1, \cdots, X_J$, where the observables are all terminal in the corresponding directed acyclic graph $\mathcal{G}$. Let

$$\mathbf{A} = (A_1, \cdots, A_K) \text{ and } \mathbf{X} = (X_1, \cdots, X_J)'.$$

The result (6) remains valid when A and X therein are replaced with the vectors $\mathbf{A}$ and $\mathbf{X}$.

Since the model is recursive, so is the structure of $\mathbf{A}$. Hence, we may further assume that as for the latent variables which are not root nodes in $\mathcal{G}$ we can determine subjective conditional probabilities for each of the non-root latent variables.

Arbitrary values in $(0,1)$ may now be assigned for the root latent variables. Suppose that there are L $(L \leq K)$ root variables and for convenience' sake, let the first L latent variables are root nodes. Once a set of subjective conditional probabilities are assigned to all the non-root nodes in $\mathcal{G}$, we can obtain the conditional probabilities of the random vector $\mathbf{X}$ given values of $\mathbf{A}_{ini} = (A_1, \cdots, A_L)$. Then by applying a vector version of Theorem 1 with A and X therein replaced with $\mathbf{A}_{ini}$ and $\mathbf{X}$, respectively, we can find a set of joint probabilities of $\mathbf{A}_{ini}$, $\dot{r}(\mathbf{A}_{ini} = i)_{i \in I_{A_{ini}}}$ as appearing in (6). Therefore, for any recursive model of categorical variables, we can obtain initial values, $\{\hat{m}(i)^{(0)}\}_{i \in I_V}$ that satisfy (1), i.e., that are calibrated to data.

Although not directly related to the issue addressed in this section, the following quote from Spiegelhalter et al. (1993) is worth noting at this point because it looks as if they share a common concern with us:

> "However, they (i.e., Spiegelhalter and Cowell (1992)) also show that withsystematic missing data on intermediate nodes such as in the CHILD network, the estimation procedure may be inconsistent and strongly reliant on the prior distribution. Therefore considerable care is required when specifying priors for nodes that are not observed, and it may be preferable to marginalise over nodes that are not to be observed and learn on this collapsed graph." (p. 243)

5. Simulation Results Using Calibrated Initial Values

In this section we will show, using simulated data, how the calibrated initial values in which there is no OD work for EM, and then, in the subsequent section, we will compare the result with that by randomly selected initial values. The model to use consists of 6 latent variables and 7 observables. All the variables are assumed to be binary in the model, and the latents are denoted by A's (think of 'abilities') and the observables by X's (think of 'item scores'). The model structure is given in Figure 1, where the circles represent latent variables and the boxes observables.

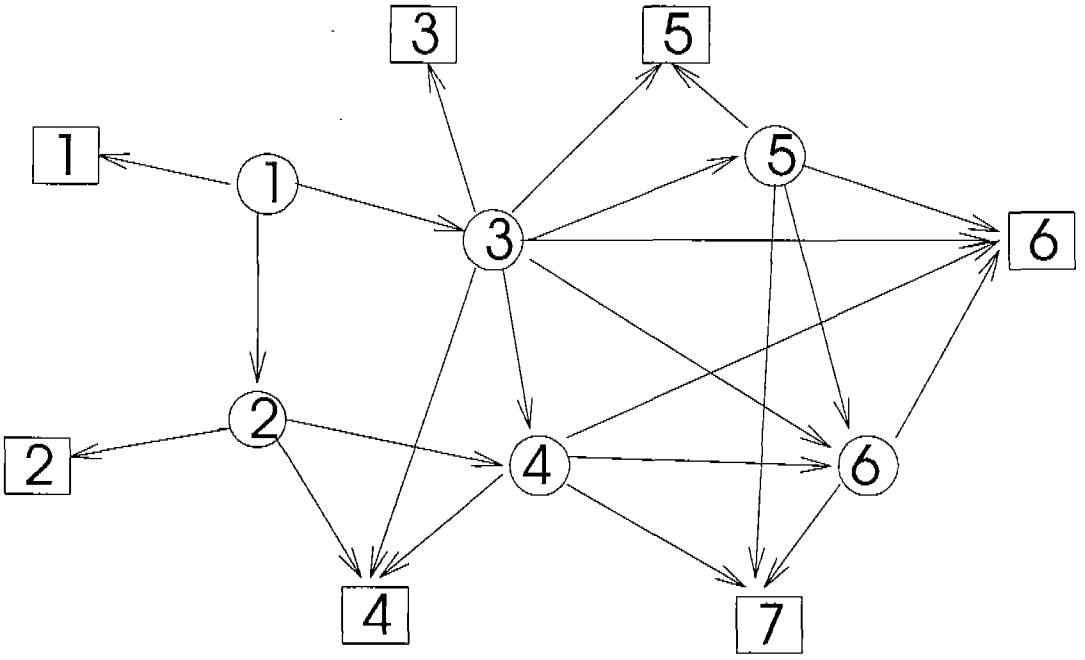

Figure 1: A DAG of 6 latent variables (circle nodes) and 7 observables (in boxes)

The model structure in the figure is an artifact from educatinal testing. Suppose that a circle represents a variable which indicates the states of a certain ability (or knowledge) and a box represents an item score variable. The directed edge between a pair of circles represents that the state of the ability at the head of the edge is influenced by the state of the ability at the tail of the edge. The directed edge from a circle node to a box node represents that the item score of the item at the box is influenced by the state of the ability corresponding to the circle.

The 6 ability nodes may be interpreted as being labelled according to some hierarchy among themselves. For instance, abilities 2 and 3 are prerequisite to ability 4. As for the relationship between ability and item score, some item nodes have multiple parent nodes while some others have single parents. We may interpret the relationship as that test takers may need as many abilities as the number of the parent nodes of a certain item score node to solve the corresponding item. But this does not necessarily imply that the probability of giving a correct answer to a certain item when all the parent-node abilities are not in good states is zero. It could be a lucky guessing, appropriate reasoning, or test wisdom that makes the probability positive.

In the simulation study, we consider binary variables only, each taking on 0 or 1. A data set was generated from a model as given in Table 1 in Appendix. We took

the data size large enough (1,000,000) in comparison with the total number of the cells ($2^{13}=8,192$) of the contingency table for the graph in Figure 1 so that, firstly, the actual probabilities may be as close as to the values I assigned and, secondly, we may avoid as much as possible chancy fluctuations in the estimates due to a small size of data.

Wu(1983) noted earlier that estimates from an EM are often subject to the initial values. When models are relatively simple such that each node has at most one parent node, randomly selected initial values for an EM may work well for recursive models. However, as the model structure becomes complicated, with more nodes involved and with more nodes having multiple parent nodes, randomly selected initial values may end up with inappropriate-looking or order-distorted (see section 3) estimates. We will see empirically in this section that if we used calibrated initial values, we might expect that the final estimates from an EM look more appropriate than those whose initial values are not calibrated.

For notational convenience, we will denote the circle node labelled k by A_k and the box node labelled i by X_i. When we used model structures that are simpler than that in Figure 1 where each node has at most two parent nodes, the initial values did not matter much, that is, whether the initials were calibrated or not we had estimates that were close to the actual probabilities. So we added more arrows to get the graph in Figure 1 where nodes A_6, X_4, X_6, and X_7 have 3 or 4 parent nodes. 8 or 16 conditional probabilities are to be estimated for each of these nodes, and the possibility of OD becomes relatively higher.

We tried three different sets of calibrated initials as in Table 1. As displayed in the table, the initials in set 1 are relatively smaller than the actual probabilities, those in set 3 relatively larger than the actual probabilities, and those in set 2 are the same as those in set 3 except for the probilities where the conditional variables are all equal to 1.

The estimates are in the last three columns of Table 1. Out of the three calibrated initials, sets 2 and 3 are preferable. Only one very minor OD is found at node X_6 for set 3 and two very minor ODs are found at the same node for set 2, while 5 ODs are found at nodes A_6, X_4, and X_6 for set 1. The goodness-of-fit levels were more or less the same for the three sets with the P-value 0.09 (the Pearson or LR statistic values were about the same around 82 with 66 degrees of freedom.)

6. A Comparison of Global Discrepancy of Estimates between the Set of Calibrated Initials and the Set of Randomly Selected Initials

The only difference in the simulation set-up between the preceeding and the current sections is that the initial values are calibrated in the former while they are not in the latter. We use the same model structure and the same true probabilities as in the preceeding section.

An advantage of a simulation study is that we know the true model. Making use of this advantage, we will compute the discrepancy between the set of the cell estimates ($\widehat{m_i}$) and the set of the true cell frequencies (n_i), where the discrepancy is represented in three different formulae:

For the index set V of the 13 variables in the model structure as in Figure 1,

$$D_1 = \sum_{i \in I_V} \frac{(n_i - \widehat{m}_i)^2}{\widehat{m}_i}$$

$$D_2 = \sum_{i \in I_V} \frac{(n_i - \widehat{m}_i)^2}{n_i}$$

$$D_3 = \left(\sum_{i \in I_V} (n_i - \widehat{m}_i)^2 \right) / 10^8.$$

D_1, D_2 and D_3 are Pearson Chi-square statistic, Neyman Chi-square statistic, and a sum of squared-errors, respectively. The D_1 and D_2 may not have Chi-square distributions since the estimates underwent a numerical restriction by the E-step of the EM algorithm. However, we don't have to worry about their distributions, since our aim is to compare the performances of calibrated and randomly selected initials on the same scale. The 10^8 in the denominator of D_3 keeps the value of D_3 within some range; otherwise, its value becomes enormously large.

I generated 65 sets of randomly selected initial values and obtained the D values. When compared with the D values from the 3 sets of calibrated initials as used in the preceeding section, the D values from these 65 sets were apparently larger than those from the calibrated initials as displayed in Figure 2. Since the D_1 and D_2 values range over a multiple of 10^5 to a multiple of 10^{18} and D_3 over a multiple of 100 to a multiple of 10^5, they are transformed as follows so that they may shrink into a much smaller range:

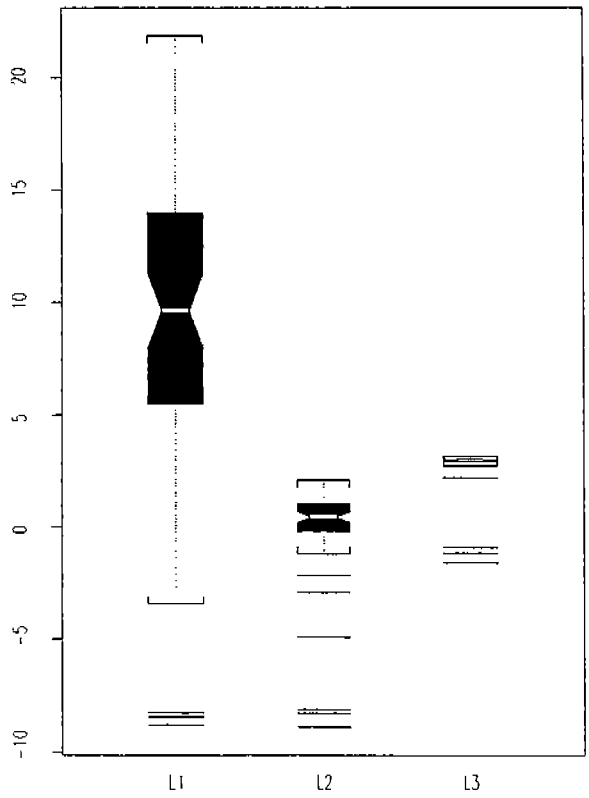

Figure 2: Box plots of the three L values as in (7). The whiskers are 1.5 times the inter-quartile range apart from each other.

$$\begin{aligned} L_i &= \log(D_i/10^{8}), \text{ for } i=1,2. \\ L_3 &= \log D_3. \end{aligned} \tag{7}$$

In the figure, the three bars at the bottom of the box plots are respectively the discrepancy (L) values of the estimates from the three sets of the calibrated initial values. The whiskers in the box plot are one and half times the inter-quartile range apart from each other. Although only 3 sets of calibrated initials were used in the simulation study, the effect of initial value calibration is believed to have been shown clearly.

Overall, the estimates from randomly selected initials were not reliable and, in particular, when there were OD phenomena in the initials, we could often see serious OD phenomena in the estimates. I will refrain from including a table of such estimates, since the table may not make any sense other than Figure 2 does.

7. Bayes Estimation

As the model becomes larger and more complicated, it becomes more likely that the OD takes place even when there is no OD in the initials and the initials are calibrated. When there is a minor OD in the estimates from calibrated initials with no OD, a reasonable treatment for that then is a Bayes estimation with an appropriate prior for the order-distorted node.

Spiegelhalter and Lauritzen (1990) introduce the notions of global independence and local independence on the prior. The former is the assumption that the prior for the whole model is given as a product of the priors on the parameters of the individual nodes of the model, and the latter being the assumption that the priors on a node are independent each other among all the possible states or values of the parents of the node. These notions are instrumental in Bayesian statistical reasoning and their application appears in Spiegelhalter et al. (1993) and Thiesson (1996) among others.

Priors are usually given in the form of a distribution, but we may allow them in the form of point estimates along with ranges to reflect an imprecision of each of the estimates. Some examples of the point estimates are found in Spiegelhalter et al. (1993) and Thiesson (1996). In this paper, all the (conditional) probabilities are multinomial, and so Dirichlet priors are used at a node. When priors are given in terms of point values along with ranges of imprecision, we can then convert them into Dirichlet distributions under a certain condition that the range of imprecision represents a one standard error interval (see Spiegelhalter et al. (1991).) This simplified prior was used in fixing the minor OD in the estimates whose initial values are set 3 in Table 1.

Table 2 in Appendix shows the result of a Bayes estimation using the simplified prior. In the table, the initial values in sets 3 and 4 are the same except that a couple of priors are imposed on conditional probabilities, $P(A_6=1|(A_3,A_4,A_5)=(0,0,0))$ and $P(X_6=1|(A_3,A_4,A_5,A_6)=(1,0,0,0))$. There is a minor OD in the estimates from set 3 at node X_6. The estimates for the conditional probability with the configurations of the conditional variables, (0,0,0,0) and (1,0,0,0), were 0.086 and 0.064, respectively. While there was only one OD in the estimates for set 3, we need to assign prior on two conditional probabilities because of a chain reaction phenomenon. When a prior was imposed on a conditional probability of X_6, the imposition affected the estimation for the conditional probability of A_6.

When an OD takes place, it is on us which parameter to impose prior on. In this simulation study, I imposed Beta priors in the simplified form as in the fifth column of Table 2. In the column, p and sd in "$p(sd)$" mean the ean and the standard deviation of the imposed Beta prior, the values are selected in the same spirit as Spiegelhalter et al. (1991). For notational convenience, we will write $\alpha(1,0,0)$ for $P(A_6=1|(A_3,A_4,A_5)=(1,0,0))$. When a prior was imposed on X_6, another OD took place in the conditional probability of A_6between $\alpha(0,0,0)$ and $\alpha(1,0,0)$. Their estimates were 0.111 and 0.091, respectively. For this pair, I determined to assign a prior upon $\alpha(0,0,0)$.

As shown in the last column of Table 2, the quality of the estimates are more or less the same between sets 3 and 4 except that there is now no OD in the estimates from set 4.

In Table 2, the Beta priors for A_6 and X_6 are assigned with a view to remove the OD phenomena in the estimates set set 3. As mentioned above, the OD took place between $\alpha(0,0,0)$ and $\alpha(1,0,0)$. I decided to impose a prior upon $\alpha(0,0,0)$ because the true value is equal to 0.05. But in practice, we do not know the true value, and so where to impose a prior is a matter of personal experience and trouble-shooting. If the true value were not known, one may impose a prior upon $\alpha(1,0,0)$ arguing that the estimates for $\alpha(0,0,1)$ and $\alpha(0,1,0)$ are 0.23 and 0.19, respectively and so that the true value of $\alpha(1,0,0)$ should be around these values. Of course, if we regard this viewpoint as acceptable, we may give it a try.

Back to the table, once we decided to impose upon $\alpha(0,0,0)$, we need to specify a Beta distribution. According to Spiegelhalter et al. (1993) and Thiesson (1996), we may select the values for the center and the length of the interval that we want the corresponding estimate be found in. In the table, I selected 0.06 and 0.03 respectively as the values for the center and half of the length of the desired interval in the hope that the corresponding estimate is located near 0.06 and not larger than 0.09 which is actually the estimate of $\alpha(1,0,0)$. We may do a similar thing as above for X_6.

It is desirable that the issue of which probability of an OD pair to put a prior on is dealt with based on the current estimates, the nature of the relation between the corresponding node and the set of its parent nodes, and experts' opinion on the OD among others.

8. Further Discussion and Concluding Remarks

The notion of calibration were often used in evaluating probability predictors which later was developed into comparing them by incorporating the concept of refinement (see Fienberg and Kim (1998) for an overview). We say that a probability predictor for a certain event is calibrated if, for each value p of the probabilities that the predictor uses for prediction, the relative frequency of the event out of the total number of the cases that the prediction value is p is equal to p. So this is an issue of linking a subjective probability to a relative frequency. This perspective per se carries over into calibrating initial values. Initial values for an EM, when applied to a complicated model structure, had better be selected in such a manner that the model structure and the data are well incorporated and experts' opinions, if any, are consistently reflected therein. This point is formulated substantially in section 4.

Another thing to note in regard to initial values for a model of abilities and test performance is the OD phenomenon. In reality, we do not know the true values of the parameters. So an OD in the estimates is a warning signal about the quality of the estimates. When OD is found at several variables, it is desirable that a new set of initial values is obtained provided the model structure is well chosen. But when there are one or two ODs, a Bayes estimation is recommendable. Since the distribution for our model is multinomial, we may use Dirichlet priors either in a distribution form or in a simplified form with an imprecision interval. We used priors given in the latter form in this paper. When an OD is being fixed at a node, another OD may occur at a neighboring node as a result of chain reaction in estimation, which was pointed out in the preceeding section.

EM is one of the most popular statistical estimation methods for a model that contains latent variables. It is easy to understand and use. Because of these merits, EM is expected to maintain its popularity notwithstanding its relatively slow convergence rate (see Van Dyk and Meng (1997) for the issue of convergence rate). When we deal with a causal model of categorical variables involving 10 or less variables, we may have no problem in applying an EM algorithm to obtain appropriate estimates even with randomly selected initial variables. But as the number of variables increases and the model structure becomes more complicated, randomly selected initial values may often end up with inappropriate estimates. But if we select initial values carefully so that they may be calibrated and the model structure and experts' opinions may be well incorporated in them, then we may have much more appropriate estimates that are far closer to the true values than the estimates from

randomly selected initials, as is illustrated in this paper.

In assigning values for the conditional probabilities $\phi(x|pa(X)=i)$ to obtain initial values, one may not worry about consulting experts. When experts are not available, we may assign values in such a way that the OD phenomenon may not occur and the initials may be calibrated. Of course, if experts' opinions were at hand, I would suggest they be respected provided the OD phenomenon does not occur.

References

[1] Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of Royal Statistical Society B*, **41**, 1, 1-31.

[2] Fienberg, S. E. and Kim, S.-H. (1998). Calibration and refinement fpr classification trees. *Journal of Statistical Planning and Inference*, **70**, 241-254.

[3] Jeong, M. S., Kim, S.-H., and Jeong, K. M. (1998). Initial value selection in applying an EM algorithm for recursive models of categorical variables. *Journal of the Korean Statistical Society*, **27**, 1, 25-55.

[4] Kyberg, Jr., H. E. and Smokler, H. E. (1980). *Studies in Subjective Probability* (edited), Huntington, New York: Robert E. Krieger Publishing Company.

[5] Pearl, J. (1988). *Probabilistic Reasoning in Intelligence Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers, Inc.

[6] Savage, L. J. (1972). *The Foundations of Statistics*, Second Revised Edition, New York: Dover Publications, Inc.

[7] Spiegelhalter, D. J. and Cowell, R. G. (1992). Learning in probabilistic expert systems. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 447-466. Clarendon Press, Oxford.

[8] Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems (with discussion). *Statistical Science*, **8**, 3, 219-283.

[9] Spiegelhalter, D. J., Harris, N. L., Bull, K., and Franklin, R. C. G. (1991). Empirical evaluation of prior beliefs about frequencies: methodology and a case study in congenital heart disease. *Technical Report 91-4*. MRC Biostatistics Unit, Cambridge.

[10] Spiegelhalter, D. J. and Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579-605.

[11] Thiesson, B. (1996) Score and information for recursive exponential models with incomplete data. Technical report R-96-2024. Department of Mathematics and Computer Science, Aalborg University, Denmark.

[12] Van Dyk, D. and Meng, X. L. (1997). On the ordering and groupings of conditional maximizations within ECM-type algorithms. *Journal of Computational and Graphical Statistics*, **6**, 2, 202-223.

[13] Whittaker, J. (1990). *Graphical models in applied multivariate statistics*, New York, NY: Wiley

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

Appendix

Table 1: True probabilities and three sets of calibrated initials for the model structure in Figure 1. The second column lists the configurations of the conditional variables for the corresponding conditional probability. The third column lists the actual probabilities assigned to the nodes. The initial values for the node probabilities are all calibrated. The P-values of the goodness-of-fit statistics (both Pearson and likelihood-ratio) are about 0.09.

list of prob's	config's	prob. values						
		true	initial values			estimates		
			set 1	set 2	set 3	set 1	set 2	set 3
$P(A_1=1)$		0.740	0.904	0.868	0.755	0.74	0.74	0.74
$P(A_1=1\|A_1)$	0	0.130	0.10	0.10	0.15	0.15	0.15	0.15
	1	0.750	0.70	0.70	0.85	0.76	0.76	0.76
$P(A_3=1\|A_1)$	0	0.221	0.10	0.10	0.15	0.22	0.22	0.22
	1	0.830	0.70	0.70	0.85	0.83	0.83	0.83
$P(A_4=1\|A_2,A_3)$	00	0.220	0.10	0.10	0.10	0.16	0.13	0.16
	01	0.220	0.15	0.30	0.30	0.29	0.25	0.27
	10	0.220	0.15	0.30	0.30	0.19	0.16	0.19
	11	0.840	0.60	0.65	0.85	0.86	0.85	0.85
$P(A_5=1\|A_3)$	0	0.151	0.10	0.15	0.15	0.15	0.12	0.14
	1	0.850	0.60	0.65	0.85	0.84	0.78	0.83
$P(A_6=1\|A_3,A_4,A_5)$	000	0.050	0.05	0.10	0.10	0.07	0.10	0.11
	001	0.100	0.10	0.20	0.20	0.14	0.20	0.23
	010	0.100	0.10	0.20	0.20	0.15	0.14	0.19
	011	0.197	0.15	0.40	0.40	0.41	0.41	0.47
	100	0.102	0.10	0.20	0.20	0.04	0.20	0.11
	101	0.199	0.15	0.40	0.40	0.16	0.35	0.35
	110	0.200	0.15	0.40	0.40	0.14	0.42	0.36
	111	0.749	0.65	0.65	0.90	0.89	0.79	0.91
$P(X_1=1\|A_1)$	0	0.141	0.10	0.20	0.20	0.14	0.14	0.14
	1	0.810	0.65	0.65	0.85	0.81	0.81	0.81
$P(X_2=1\|A_2)$	0	0.350	0.10	0.20	0.20	0.34	0.34	0.34
	1	0.950	0.65	0.70	0.90	0.95	0.95	0.95
$P(X_3=1\|A_3)$	0	0.100	0.10	0.20	0.20	0.10	0.10	0.10
	1	0.730	0.65	0.75	0.85	0.73	0.73	0.73

(to be continued)

list of prob's	config's	prob. values						
		true	initial values			estimates		
			set 1	set 2	set 3	set 1	set 2	set 3
$P(X_4=1\|A_2,A_3,A_4)$	000	0.100	0.05	0.10	0.10	0.10	0.10	0.10
	001	0.148	0.10	0.20	0.20	0.16	0.16	0.16
	010	0.149	0.10	0.20	0.20	0.15	0.15	0.15
	011	0.203	0.15	0.40	0.40	0.17	0.18	0.17
	100	0.150	0.10	0.20	0.20	0.13	0.13	0.13
	101	0.200	0.15	0.40	0.40	0.30	0.32	0.31
	110	0.200	0.15	0.40	0.40	0.14	0.18	0.18
	111	0.830	0.70	0.70	0.90	0.82	0.82	0.82
$P(X_5=1\|A_3,A_5)$	00	0.100	0.10	0.20	0.20	0.09	0.09	0.09
	01	0.150	0.20	0.40	0.40	0.21	0.24	0.22
	10	0.152	0.20	0.40	0.40	0.22	0.23	0.20
	11	0.680	0.60	0.70	0.90	0.67	0.71	0.68
$P(X_6=1\|A_3,A_4,A_5,A_6)$	0000	0.100	0.05	0.10	0.10	0.09	0.09	0.09
	0001	0.100	0.10	0.20	0.20	0.21	0.17	0.17
	0010	0.100	0.10	0.20	0.20	0.11	0.11	0.10
	0011	0.203	0.15	0.30	0.30	0.19	0.16	0.19
	0100	0.100	0.10	0.20	0.20	0.11	0.13	0.11
	0101	0.200	0.15	0.30	0.30	0.28	0.34	0.29
	0110	0.200	0.15	0.30	0.30	0.20	0.23	0.17
	0111	0.605	0.20	0.40	0.40	0.50	0.43	0.43
	1000	0.100	0.10	0.20	0.20	0.07	0.07	0.06
	1001	0.199	0.15	0.30	0.30	0.51	0.42	0.38
	1010	0.200	0.15	0.30	0.30	0.21	0.15	0.17
	1011	0.596	0.20	0.40	0.40	0.37	0.47	0.38
	1100	0.196	0.15	0.30	0.30	0.18	0.14	0.12
	1101	0.601	0.20	0.40	0.40	0.52	0.71	0.58
	1110	0.600	0.20	0.40	0.40	0.52	0.62	0.45
	1111	0.821	0.70	0.70	0.90	0.80	0.82	0.80
$P(X_7=1\|A_4,A_5,A_6)$	000	0.100	0.05	0.10	0.10	0.11	0.11	0.11
	001	0.150	0.10	0.20	0.20	0.29	0.30	0.23
	010	0.150	0.10	0.20	0.20	0.17	0.13	0.14
	011	0.700	0.15	0.40	0.40	0.31	0.43	0.34
	100	0.150	0.10	0.20	0.20	0.16	0.14	0.12
	101	0.700	0.15	0.40	0.40	0.42	0.72	0.52
	110	0.702	0.15	0.40	0.40	0.70	0.83	0.72
	111	0.929	0.70	0.70	0.90	0.90	0.90	0.90

Table 2: True probabilities and two sets of calibrated initials for the model structure in Figure 1. The second column lists the configurations of the conditional variables for the corresponding conditional probability. The third column lists the actual probabilities assigned to the nodes. The initial values in set 4 are the same as set 3 except that priors are imposed on two nodes, A_6 and X_6. The P-values of the goodness-of-fit statistics (both Pearson and likelihood-ratio) are about 0.09.

list of prob's	config's	prob. values				
		true	initial values		estimates	
			set 3	set 4	set 3	set 4
$P(A_1=1)$		0.740	0.755		0.74	0.74
$P(A_1=1\|A_1)$	0	0.130	0.15		0.15	0.15
	1	0.750	0.85		0.76	0.76
$P(A_3=1\|A_1)$	0	0.221	0.15		0.22	0.22
	1	0.830	0.85		0.83	0.83
$P(A_4=1\|A_2,A_3)$	00	0.220	0.10		0.16	0.17
	01	0.220	0.30		0.27	0.28
	10	0.220	0.30		0.19	0.19
	11	0.840	0.85		0.85	0.85
$P(A_5=1\|A_3)$	0	0.151	0.15		0.14	0.14
	1	0.850	0.85		0.83	0.83
$P(A_6=1\|A_3,A_4,A_5)$	000	0.050	0.10	0.06(0.03)*	0.11	0.06
	001	0.100	0.20		0.23	0.24
	010	0.100	0.20		0.19	0.20
	011	0.197	0.40		0.47	0.48
	100	0.102	0.20		0.11	0.06
	101	0.199	0.40		0.35	0.35
	110	0.200	0.40		0.36	0.37
	111	0.749	0.90		0.91	0.91
$P(X_1=1\|A_1)$	0	0.141	0.20		0.14	0.14
	1	0.810	0.85		0.81	0.81
$P(X_2=1\|A_2)$	0	0.350	0.20		0.34	0.34
	1	0.950	0.90		0.95	0.95
$P(X_3=1\|A_3)$	0	0.100	0.20		0.10	0.10
	1	0.730	0.85		0.73	0.73

(to be continued)

list of prob's	config's	prob. values				
		true	initial values		estimates	
			set 3	set 4	set 3	set 4
$P(X_4=1\|A_2,A_3,A_4)$	000	0.100	0.10		0.10	0.10
	001	0.148	0.20		0.16	0.15
	010	0.149	0.20		0.15	0.15
	011	0.203	0.40		0.17	0.17
	100	0.150	0.20		0.13	0.13
	101	0.200	0.40		0.31	0.30
	110	0.200	0.40		0.18	0.18
	111	0.830	0.90		0.82	0.82
$P(X_5=1\|A_3,A_5)$	00	0.100	0.20		0.09	0.09
	01	0.150	0.40		0.22	0.21
	10	0.152	0.40		0.20	0.20
	11	0.680	0.90		0.68	0.68
$P(X_6=1\|A_3,A_4,A_5,A_6)$	0000	0.100	0.10		0.09	0.09
	0001	0.100	0.20		0.17	0.15
	0010	0.100	0.20		0.10	0.11
	0011	0.203	0.30		0.19	0.18
	0100	0.100	0.20		0.11	0.11
	0101	0.200	0.30		0.29	0.32
	0110	0.200	0.30		0.17	0.17
	0111	0.605	0.40		0.43	0.40
	1000	0.100	0.20	0.095(0.007)	0.06	0.09
	1001	0.199	0.30		0.38	0.19
	1010	0.200	0.30		0.17	0.17
	1011	0.596	0.40		0.38	0.38
	1100	0.196	0.30		0.12	0.12
	1101	0.601	0.40		0.58	0.60
	1110	0.600	0.40		0.45	0.45
	1111	0.821	0.90		0.80	0.80
$P(X_7=1\|A_4,A_5,A_6)$	000	0.100	0.10		0.11	0.11
	001	0.150	0.20		0.23	0.22
	010	0.150	0.20		0.14	0.14
	011	0.700	0.40		0.34	0.33
	100	0.150	0.20		0.12	0.12
	101	0.700	0.40		0.52	0.55
	110	0.702	0.40		0.72	0.72
	111	0.929	0.90		0.90	0.90

* In $p(sd)$, p and sd are the mean and the standard deviation of a Beta prior, respectively.