Fluctuation of estimates in an EM procedure

Seong-Ho Kim and Sung-Ho Kim[1]

ABSTRACT

Estimates from an EM algorithm are somewhat sensitive to the initial values for the estimates, and it is more likely when the model becomes larger and more complicated. In this article, 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 first half of the procedure and that it can be subdued by applying the Bayesian method of estimation. Both simulation data and real data are used for illustration.

Keywords: Bayesian method; Calibrated initial values; Directed acyclic graph; Dirichlet prior

1. Introduction

The EM method as proposed by Dempster et al. (1977) has been widely used 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 for the missing variables and likelihood-maximization. We will confine our attention on a possible fluctuation of the estimates during the estimation process, and the variables are all categorical.

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 in 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. 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 on this topic.) Under the assumption of conditional association, conditional probabilities are expected to be ordered according to the states of conditioning variables. Experience says that as the number of conditioning variables increase, it is more probable that the conditional positive association (CPA) is violated in the estimates of the conditional probability. What make things more difficult is that

[1]Division of Applied Mathematics, Korea Advanced Institute of Scicence and Technology, Daejeon, 305-701, South Korea. e-mail: mathan@mail.kaist.ac.kr and shkim@mail.kaist.ac.kr

the estimates from the EM algorithm depends 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 subdued to some level by applying the calibration method of Kim (2002).

This work was motivated by the need that the estimates satisfy the CPA. 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. Our main interest in this work is in how the estimates fluctuate during the estimation process and when we can see if the CPA obtains in the estimates.

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 possible *levels* of these. 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 $|n| = \sum_{i \in \mathcal{I}} n(i)$. 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.

An *A-marginal table* is for $A \subseteq V$ obtained by only classifying the objects according to the criteria in A, i.e., by only considering the variables in A. It has *marginal cells* $i_A \in \mathcal{I}_A = \times_{v \in A} I_v$. The marginal counts are the quantities $n(i_A)$. Again, if we let

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

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). $a \rightarrow b$ 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 . The expression

$pa(A)$ denotes the collection of parents in $A \subseteq V$: $pa(A) = \cup_{a \in A} pa(a) \setminus A$. 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.

3. EM algorithm

Suppose $V = \Delta \cup \Lambda$ is a set of classification variables to the complete unobserved data, where all the variables in V are binary and Δ is a set of the observed variables and Λ that of those to the unobserved (latent) variables. Furthermore, all the variables of Δ are terminal.

An E-step is implemented according to the expression $\widehat{m}(i_V)^{(r+1)} = n(i_\Delta)\frac{\widehat{m}(i_V)^{(r)}}{\widehat{m}(i_\Delta)^{(r)}}$, where r is the iteration count of the E-and M-steps. Once the E-step is carried out, the new estimates satisfy $\widehat{m}(i_\Delta)^{(r+1)} = n(i_\Delta)$. An M-step is implemented according to the expression $\widehat{m}(i_V)^{(r+1)} = |n| \prod_{v \in V} \frac{\widehat{m}(i_{fa(v)})^{(r)}}{\widehat{m}(i_{pa(v)})^{(r)}}$.

4. Order of magnitude among the estimates

4.1. After an E-step

In this section, we present two theorems that that provide sufficient conditions for the order of magnitude among the estimates of the conditional probabilities remain the same before and after an E-step. Theorem 4.1 is concerned with a terminal node which is observable, and Theorem 4.2 with the node of a latent variable.

Theorem 4.1. *Suppose $v(\in \Delta)$ is terminal and $pa(v)(\neq \emptyset) \subseteq \Lambda$ and assume*

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

where $\delta = \Delta \setminus \{v\}$ and

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

Then, $\widehat{p}(i_v|i_{pa(v)})^{(r)} > \widehat{p}(i_v|j_{pa(v)})^{(r)}$, it follows that $\widehat{p}(i_v|i_{pa(v)})^{(r+1)} > \widehat{p}(i_v|j_{pa(v)})^{(r+1)}$, where $i_v, j_v \in \mathcal{I}_v$ with $i_v \neq j_v$, $i_{pa(v)}, j_{pa(v)} \in \mathcal{I}_{pa(v)}$ with $i_{pa(v)} \neq j_{pa(v)}$, and the rth and the $(r+1)$th estimates are from an M-step and from the subsequent E-step, respectively.

Theorem 4.2. *Suppose $v \in \Lambda$ and $pa(v) \cap \Delta = \emptyset$ and assume*

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

where

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

Then $\widehat{p}(i_v|i_{pa(v)})^{(r)} > \widehat{p}(i_v|j_{pa(v)})^{(r)}$, *it follows that* $\widehat{p}(i_v|i_{pa(v)})^{(r+1)} > \widehat{p}(i_v|j_{pa(v)})^{(r+1)}$, *where* $i_v, j_v \in \mathcal{I}_v$ *with* $i_v \neq j_v$, $i_{pa(v)}, j_{pa(v)} \in \mathcal{I}_{pa(v)}$ *with* $i_{pa(v)} \neq j_{pa(v)}$, *and the* r*th and the* $(r+1)$*th estimates are from an M-step and from the subsequent E-step, respectively.*

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. Thus we have the theorem below.

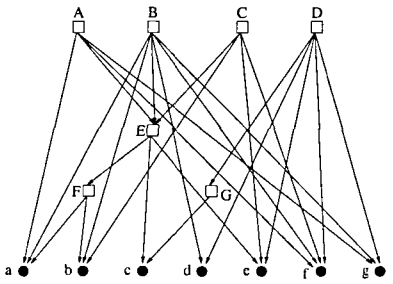
Theorem 4.3. *For* $v \in V$, $\widehat{p}(i_v|i_{pa(v)})^{(r+1)} = \widehat{p}(i_v|i_{pa(v)})^{(r)}$, *where the left-hand side is the estimate from an M-step and the right-hand side from the preceding E-step.*

In a nutshell, the order of magnitude of the estimates remains the same before and after every M-step, but this is not necessarily the case as for the E-step. Sufficient conditions are provided under which the order is maintained before and after an E-step. So, if the order of magnitude among the estimates is far from its initial status, the order distortion must have taken place at an E-step. According to Theorems 4.1 and 4.2, we can see that as the estimates get closer to limits, the order of magnitude in the estimates is more likely to keep its preceding status. Thus the order is more likely to be distorted at a relatively early stage of E-steps, as will be illustrated shortly. As a remedy for this, we propose applying a Bayesian method to the EM algorithm by imposing a Dirichlet prior on every variable of a given model (Bishop, Fienberg, and Holland 1975, Section 12.2).

5. Illustration

We analyzed a data set of 7 multiple choice items of the Mathematics section of the Korean SAT that was administered in 1999. We have 7 observed binary variables for item scores (0 for incorrect answer and 1 for 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).

The 14 variables are related as in the left side of Figure 5.1, 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.



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

Figure 5.1: (i) The left figure is a DAG for real data. Bullets are for the item score variables and boxes for the knowledge states. (ii) The right table is the list of the knowledge units involved in the model. DK is an acronym of "declarative knowledge" and PK of "procedural knowledge."

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 the right side of Figure5.1.

To show how the estimates fluctuate during the EM process, we display the values of $\widehat{P}(X_b = 1|X_{B,C,F})^{(r)}$ and $\widehat{P}(X_g = 1|X_{A,B,D})^{(r)}$ in Figure 5.2. For convenience' sake, we denote by $p0, p1, \cdots, p7$, respectively, the values of $\widehat{P}(X_b = 1|X_{(B,C,F)} = x_{(B,C,F)})$, in the order of the configurations of $X_{(B,C,F)}$ from $(0,0,0)$ to $(1,1,1)$, in Figure 5.2, and analogously for $\widehat{P}(X_g = 1|X_{A,B,D})^{(r)}$. We can see in the figure that the estimates fluctuate much more with real data (see the first column of the figure) than with simulated data (the second column). The order of magnitude in the estimates based on the real data does not look stabilized until the end of the first half of the process. Notice that the estimates $p2$ and $p6$ fluctuate over a wide range of values during the first half of the procedure. Such a wild fluctuation is also seen in the estimates based on the simulated data. This fluctuation may be influenced by the initial values or by the model structure we choose. When we applied a Bayesian method by imposing a Beta prior (with its parameters α and β, $\alpha + \beta = 100$) on every variable in the model, we could have the estimates stabilized from the early stage of the process as is shown in the last column of Figure 5.2.

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

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. Ser. B 39, 1-38.

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

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

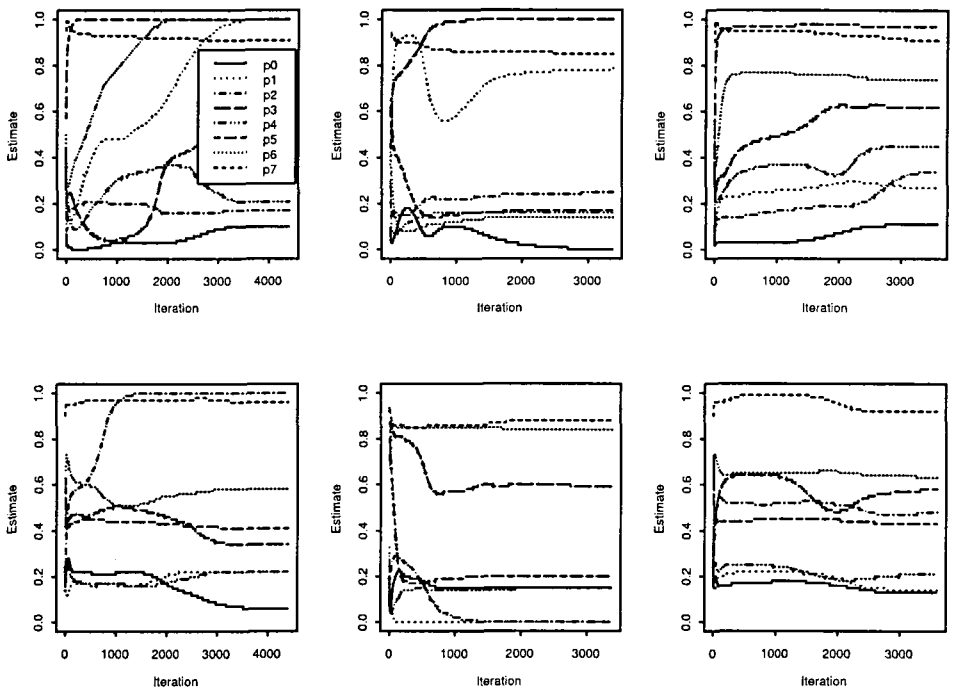


Figure 5.2: A time-series-type display of $\widehat{P}(X_b = 1|X_{B,C,F})^{(r)}$ and $\widehat{P}(X_g = 1|X_{A,B,D})^{(r)}$. $\widehat{P}(X_b = 1|X_{B,C,F})^{(r)}$-values are displayed in the first row and $\widehat{P}(X_g = 1|X_{B,C,F})^{(r)}$-values in the second row, where $p0, p1, \cdots, p7$ are explained in the text. The estimates from an EM based on real data are displayed in the first column, the estimates from an EM based on simulated data in the second column, and the estimates from a Bayesian EM based on real data in the last column.

Kim, S. -H. (2000), " Calibrated initials for an EM applied to recursive models of categorical variables," *Computational Statistics and Data Analysis*, 40, 91-110.

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

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

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