응용통계연구 제 7권 2호
1994년 9월 pp. 145-157

Estimation from Incomplete Data in Multivariate Distributions under Stochastic Ordering[1)]

Kwang Mo Jeong[2)]

Abstract

For multivariate distributions satisfying stochastic ordering, we suggest maximum likelihood estimation with incomplete data via an EM algorithm. In this paper we restrict our attention to the contingency tables with partially cross-classified observations. We may use the existing isotonic regression program to implement EM algorithm, and we illustrate the estimation process through an example.

1. Introduction

In univariate cases the assumption that one distribution is stochastically larger than another means that larger values are more likely from the former than the latter. Inference based on stochastically ordered univariate data has been studied extensively by many researchers. The theory of isotonic regression has played a key role in the estimation and testing for stochastically ordered populations. Among others we refer to Brunk, Franck, Hanson and Hogg(1966), Barlow, Bartholomew, Bremner and Brunk(1972), and Robertson and Wright(1981) for isotonic regression estimates.

The multivariate version of stochastic ordering also conveys the intuition that larger values are more likely for one population than another. For example we can consider two coherent systems where one system lasts longer than the other. We note that the life time of a coherent system is a nondecreasing function of the lifetimes of its component parts. Recently many authors have shown interests in stochastically ordered multivariate data. Sampson and Whitaker(1989) considered maximum likelihood estimates(MLEs) of multivariate distributions under stochastic ordering, and Lucas and Wright(1991) studied testing for stochastic ordering between multinomial populations.

On the other hand Dykstra(1982) studied MLEs for survival functions of censored data, and Lee(1987) suggested MLEs for the multinomial populations with fixed and random zeros. Frequently we encounter incomplete observations in multivariate data where some

1) This paper was supported(in part) by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1992.

2) Department of Statistics, Pusan National University, Kumjung Ku, Pusan, 609-735, KOREA.

cases are not observed completely, that is, they have missing values or they are partially cross-classified.

A general technique for treating incomplete data is the EM algorithm which takes expectation(E) steps and maximization(M) steps. For the concepts and some well-known properties we refer readers to Dempster, Laird, and Rubin(1977), and Little and Rubin(1987). The EM algorithm has been applied to various types of incomplete data. In contingency tables with missing data Fuchs(1982) used an EM method to obtain MLEs in view of model selections. Very recently Meng and Rubin(1993) proposed a modified version of EM algorithm, and called it ECM algorithm which takes advantages of the simplicity of complete data conditional maximum likelihood estimation by replacing a complicated M step of EM with several computationally simpler conditional maximization(CM) steps.

So far the problems about stochastic ordering relations and incomplete observations have been addressed separately. But in practice we encounter varieties of incomplete data from probability distributions under some restriction on the stochastic order relation among them.

The purpose of this paper is to obtain MLEs of probability distributions when they have stochastic order relation. This paper consists of 5 sections. In Section 2, we briefly review the estimation process for complete data under stochastic order constraints and we explain that the MLEs can be obtained by applying the existing isotonic regression algorithms. In Section 3, we consider the estimation from incomplete data under stochastic ordering. The EM algorithm is briefly outlined. We consider the EM method under some order relations, in particular for contingency tables with partially cross-classified observatins. We modify the Fortran program by Bril, Dykstra, Pillers, and Robertson(1984) to perform the EM method. In Section 4, the method proposed in Section 3 is explained through an example. Finally, in Section 5 we summarize and conclude the paper with comments on some problems.

2. Estimation for complete data

The MLEs for complete data from arbitrary distributions satisfying some stochastic order relations were obtained by Sampson and Whitaker(1989). In this section we deal with incomplete data from the probability distributions satisfying stochastic order relations, and we confine our attention to contingency tables.

To introduce the concept of stochastic ordering we let X and Y be two discrete random vectors with a common support on an $I \times J$ lattice. We may consider two ordinal contingency tables with I rows and J columns categories. We denote $\mathrm{p} = (p_{ij})$ and $\mathrm{q} = (q_{ij})$ to be cell probabilities of contingency tables, that is,

$$p_{ij} = \Pr(X=(i,j)), \quad q_{ij} = \Pr(Y=(i,j)).$$

To define a stochastic order let $u = (u_1, u_2)$ and $v = (v_1, v_2)$ be vectors in R^2, then U is an upper set in R^2 if $u \in U$ and $u_i \le v_i$, for $i = 1, 2$, imply that $v \in U$. A lower set is similarly defined in reverse ordering. We say that X is stochastically larger than Y, denoted by $X \geq^{st} Y$, if

$$\Pr(X \in U) \geq \Pr(Y \in U),$$

for every upper set U. For contingency tables this is equivalent to

$$\sum_{(i,j) \in U} p_{ij} \geq \sum_{(i,j) \in U} q_{ij}, \tag{2.1}$$

for every upper set U, where $\sum_{(i,j) \in U}$ means the sum over all $(i,j) \in U$.

The relation (2.1) can easily be extended to r dimensional contingency tables with $I_1, I_2, \cdots, I_r$ categories for each variable, respectively. Then one population is stochastically larger than the other iff

$$\sum_{i \in U} p_i \geq \sum_{i \in U} q_i,$$

for every upper set U, where $\sum_{i \in U}$ is taken over all $i = (i_1, \cdots, i_r)$ belonging to the upper set U.

2.1. One-sample problems

First, we consider a one-sample problem where one probability distribution is a known standard, and we also confine our attention to 2 dimensional contingency tables. Without loss of generality, let $q = (q_{ij})$ be known as $q^0 = (q^0_{ij})$. Let $x = (x_{ij})$ be observed counts for the table with cell probabilities $p = (p_{ij})$.

Then the likelihood function is given as

$$L(p) \propto \prod_i \prod_j p_{ij}^{x_{ij}}. \tag{2.2}$$

We estimate p_{ij} maximizing $L(\mathrm{p})$ subject to the constraints,

$$\sum_{(i,j)\in U} p_{ij} \geq \sum_{(i,j)\in U} q^{0}_{ij}. \tag{2.3}$$

Some authors call these as *restricted maximum likelihood estimates.*

For the complete data Robertson and Wright (1974) gave a theoretical derivation and consistency arguments in a general setting where the underlying distributions are not necessarily discrete. Sampson and Whitaker(1989) discussed the MLEs and its computational aspects for multivariate distributions under stochastic ordering. Research in this area has been found on the complete data so far.

According to Sampson and Whitaker(1989) the MLE of p subject to (2.3) is represented by

$$\hat{p}_{ij} = x_{ij} \min_{L} \max_{U} \{ \sum q^{0}_{kl} / \sum x_{kl} \}, \tag{2.4}$$

where the sum is taken over all $(k,l) \in L \cap U$ for every upper set U and lower set L.

The estimate in (2.4) is based on the isotonic regression technique and it can be computed using existing isotonic regression algorithms. At this point we will briefly outline the concept of isotonic regression estimates. Let $S = \{(i,j) \mid i=1,2,\cdots,I;\ j=1,2,\cdots,J\}$ be the set of category combinations, and we define the partial order $\ll$ on S by $(i,j) \ll (k,l)$ iff $i-k \leq 0$ and $j-l \leq 0$. Least squares regression estimates f_{ij} is the solution of

$$\min_{f_{ij}} \sum_{i,j} (g_{ij} - f_{ij})^2 w_{ij}$$

with f_{ij} subject to

$$f_{ij} \leq f_{kl}, \quad \text{if } (i,j) \ll (k,l), \tag{2.5}$$

where g_{ij} are given and w_{ij} positive constants.

Robertson and Wright(1981) obtained the MLEs of multinomial probabilities under stochastic ordering regarding them as the least squares regression estimates. For more detailed discussion of isotonic regression see Barlow et al.(1972).

As a natural extension to contingency tables we obtain the MLE of p_{ij} subject to the constraints in expression (2.3) as follows:

$$\hat{p}_{ij} = \bar{p}_{ij}\, E_{\bar{\mathrm{p}}}\left(\frac{\mathrm{q}^{0}}{\bar{\mathrm{p}}} \mid \mathrm{f} \right)_{ij}, \tag{2.6}$$

where $\bar{\mathrm{p}} = (\bar{p}_{ij})$ denotes sample proportions and $E_{\mathrm{w}}(\mathrm{g} \mid \mathrm{f})_{ij}$ denotes the least

squares isotonic regression estimates of $g = (g_{ij})$ over the set $f = (f_{ij})$, where f_{ij} satisfy the relation in (2.5). The MLE under stochastic ordering can be represented as the sample proportion multiplied by the isotonic regression estimates of $q^0 / \bar{p}$ with weight $\bar{p}$. For complete data the isotonic regression estimates can be obtained using the existing algorithm.

Dykstra and Robertson(1982) proposed an efficient iterative algorithm for solving the isotonic regression on r dimensional subset with respect to component-wise partial ordering. Bril et al.(1984) provided a Fortran program implementing the Dykstra and Robertson algorithm for 2 dimensional data.

2.2. Two-sample problems

Let p and q be two unknown probability distributions satisfying the stochastic order relations (2.1). Let $x = (x_{ij})$ and $y = (y_{ij})$ be complete data from p and q, respectively. We estimate p and q maximizing the likelihood function

$$L(p, q) \propto \prod_i \prod_j p_{ij}^{x_{ij}} q_{ij}^{y_{ij}},$$

subject to the constraint (2.1).

The estimation process can be simplified using the results of one-sample problems. This approach has been used successfully in the univariate setting by Barlow and Brunk(1972) and Dykstra(1982), and it has been extended to the multivariate problem by Sampson and Whitaker(1989). By fixing q we apply the one-sample problem results and also similarly by fixing p we use the results for reversed stochastic ordering. This two-sample problem can be reformulated as solving two one-sample problems.

First, we take a pooled estimator of p and q based on the combined data x and y. Let $\bar{r}_{pool} = (\bar{r}_{ij})$ be the pooled estimator defined by

$$\bar{r}_{ij} = \frac{x_{ij} + y_{ij}}{m + n}$$

under the assumption $p = q$, where m and n are the sample sizes of the two contingency tables.

Second, we solve the following two one-sample problems:

$$\max_{\mathrm{p}} \prod_{ij} p_{ij}^{x_{ij}} \quad \text{subject to} \quad \mathrm{p} \geq^{st} \bar{\mathrm{r}}_{pool} ,$$

$$\max_{\mathrm{q}} \prod_{ij} q_{ij}^{y_{ij}} \quad \text{subject to} \quad \mathrm{q} \leq^{st} \bar{\mathrm{r}}_{pool} .$$

3. Estimation for incomplete data

3.1. EM algorithm

For incompletely observed contingency tables we let $\{z_{ij}^{AB}\}$ be the fully observed counts, $\{z_i^A\}$ and $\{z_j^B\}$ be partially observed counts of row variable A and column variable B, respectively. The tables, $\{z_i^A\}, \{z_j^B\}$, for partially observed counts are called *supplemental tables.*

Let $\mathrm{x} = (x_{ij})$ represent the unobserved frequencies that would have been obtained if all the data records were complete. Here we assume that the missing observations are *missing at random,* which implies that given particular values of the obseerved variables, the values missing on the other variables are missing at random. Under the assumptions for an observation from a supplemental table, for example, observed only on the variable A, with $A = i$, we have

$$\Pr(\mathrm{B} = j \mid \mathrm{A} = i) = p_{ij}/p_{i+} .$$

The MLEs from incomplete data are usually computed using an EM method. Maximizing the incomplete data likelihood function can be formulated via an EM algorithm. So we can represent the two steps of EM algorithm as follows.

E-step: Estimate the complete-data sufficient statistics $s(\mathrm{x})$ by

$$s^{(t)} = E[s(\mathrm{x}) \mid \mathrm{z}, \mathrm{p}^{(t)}]$$

M-step: Determine $\mathrm{p}^{(t+1)}$ to be a value maximizing $L(\mathrm{p})$ subject to (2.3), where $s(\mathrm{x})$ is replaced by $s^{(t)}$.

For observed incomplete data, $\{z_{ij}^{AB}\}$, $\{z_i^A\}$, $\{z_j^B\}$, the E-step can be expressed as

$$x_{ij}^{(t)} = z_{ij}^{AB} + z_i^A(m_{ij}^{(t)}/m_{i+}^{(t)}) + z_j^B(m_{ij}^{(t)}/m_{+j}^{(t)}) \tag{3.1}$$

where $m_{ij}^{(t)}$ is the expected cell count such that $p_{ij}^{(t)} = m_{ij}^{(t)}/m_{++}$. Equation (3.1) says that the expected cell mean at step t equals the count in cell (i,j) from the fully categorized table $\{z_{ij}^{AB}\}$ plus a proportional allocation from the cells of supplemental tables.

As mentioned by Fuchs(1982) use of the EM algorithm has two particular advantages.

First, it has an intuitive interpretation of adjustment of the values of sufficient statistics followed by the computation of the MLE for a sample of complete data.

Second, we may use the available algorithms for obtaining MLEs in samples with complete data to perform the M-step. For example, the algorithm by Dykstra and Robertson(1982) may be useful.

We briefly outline the isotonic regression algorithm by Dykstra and Robertson(1982). The algorithm consists of three steps: Step 1 performs isotonization over rows, and Step 2 does it over columns from the initial values at Step 1. Step 3 repeats the previous two steps until the solution in isotonic regression converges.

At each step of the EM algorithm, the E-step requires a simple calculation while the M-step may be iterative. In this paper we implement the whole steps of EM algorithm using the same Fortran code of Bril et al.(1984) by inserting the iterative computation of equation (3.1). This is the main point of this paper in the sense that we can compute the MLEs based on incomplete data under a stochastic order restriction using the existing Fortran program.

3.2. Convergence and initial values

In this section we discuss the convergence property of the EM algorithm suggested in Section 3.1.

First, according to Dykstra and Robertson(1982), at each M-step the isotonic regression algorithm converges to a true value. We note that the global maximization nature of M-step does not imply the convergence of whole EM algorithm.

Many authors, including Dempster et al.(1977) and Wu(1983), have discussed the convergence properties of EM algorithm. But it is known that no general optimization algorithms are guaranteed to converge to local maxima. According to Wu(1983), if the unobserved complete-data likelihood can be described by a (curved) exponential family with compact parameter space, then all the limit points of any EM sequence are stationary

points of the likelihood function. Without loss of generality, for the one-sample problems the complete-data likelihood in (2.2) satisfies the above conditions and we can find a stationary point. If a stationary value can be identified then the problem is to establish it as the unique or global maximum. Since the convergence to stationary value or local maximum or global maximum depends on the choice of starting points, it is recommended in general to try several EM iterations with different starting points.

In (3.1) when some of the z_{ij}^{AB}'s are zeros, the allocation has some noteworthy characteristics. The use of zeros as initial values is equivalent to setting those cells as a priori empty. Thus the practical implication is that the first step of the algorithm should be adding a very small positive value to the cells. Zero cell counts cause some problems in the M-step because the MLE in (2.6) under order restrictions cannot be obtained directly.

Sampson and Whitaker(1989) approximated MLEs by replacing the zero cell probabilities as ε when they are zeros, where ε is a very small positive value. In the Fortran code of Bril et al.(1988) ε is taken to be 10^{-5}. Intuitively we can take the observed cell frequencies of fully cross-classified tables as the initial estimates in equation (3.1). The EM algorithm is known to be insensitive to the starting values as long as they are obtained from the same structure of the probability model which we are estimating for.

As a convergence criterion for the EM algorithm we may choose a criterion based on the likelihood function or the distance of the estimates between the previous and the current steps. In this paper we take the criterion

$$|| \hat{p}^{(t)} - \hat{p}^{(t+1)} ||^2 \leq d \qquad (3.2)$$

where $\hat{p}^{(t)}$ denotes a vector of the estimates at the t^{th} iteration and the constant d is a predetermined positive value. We note that the value d can be taken appropriately according to the cell sizes of the contingency table and the desired accuracy of the estimates.

4. An Example

We apply the EM method to 2 dimensional contingency tables with partially cross-classified observations. Table 4.1 denotes counts of total 2,294 young men who failed in the Armed Forces Qualification(AFQ) test in USA(Fienberg, 1977). To apply the method suggested in previous section we rearrange the data according to the race, black versus white as in Table 4.1.

Table 4.1 Observed Cross-Classification of 2294 Young Men Who Failed to AFQ Test (Talbot and Mason, 1975)

(a) White

	1	2	3	total
1	270	144	59	473
2	21	29	14	64
3	29	37	51	117
4	249	128	43	416

(b) Black

	1	2	3	total
1	129	173	122	424
2	23	55	32	110
3	13	39	21	73
4	227	285	105	617

The row and column variables represent father's and respondent's educational levels, respectively. The educational levels are recoded as

1. Grammar School, 2. Some High School,
3. High School Graduate, 4. Non Response.

About 39 and 50 percents of whites and blacks are missing in father's educational levels, respectively. So it is reasonable to include partially cross-classified counts in the estimation process.

To apply the estimation procedure proposed in the previous section we may assume that the probability distributions for cell probabilities satisfy stochastic order relation, for example, the blacks are stochastically larger than the whites in the sense of equation (2.1). The assumption seems intuitive. The whites who have low educational levels in both father and himself may be lower than blacks in the sense of income, social status, mental age and so on.

The hypothesis of stochastic ordering may be tested but testing has not been considered in this paper. Interested readers may refer to Robertson and Wright(1981) and Lucas and Wright(1991).

This example is a two-sample problem and so we first estimate the combined cell probabilities as in Table 4.2. Next we obtain the MLEs for the whites under the assumption that the probability distribution of the whites is stochastically smaller than the known distribution of Table 4.2. The MLEs for blacks are computed in a similar way. Table 4.3 represents MLEs of the cell probabilities under the assumption that the blacks are stochastically larger than the whites.

For the convergence criterion given in (3.2) with $d = 10^{-7}$ the EM iteration converged at 23^{rd} and 25^{th} to obtain Table 4.3. (a),(b), respectively. In these cases the initial values are taken as the sample cell proportions of fully cross-classified tables.

Table 4.2 Estimated Combined Cell Probabilities For AFQ Test Data of Table 4.1.

	1	2	3
1	.3432	.2578	.1180
2	.0378	.0683	.0300
3	.0361	.0618	.0469

Table 4.3 MLEs for the cell probabilities of AFQ Test Data Under the Assumption that Black $\geq^{st}$ White

(a) White

	1	2	3
1	.4285	.2261	.0919
2	.0378	.0485	.0223
3	.0443	.0537	.0469

(b) Black

	1	2	3
1	.2046	.2760	.1912
2	.0384	.0937	.0513
3	.0240	.0740	.0469

5. Conclusion and comments

We suggested a method to compute MLEs for the stochastically ordered populations when some observations are missing or incomplete. In this paper the proposed EM method was applied to 2 dimensional contingency tables with ordered categories, where some observations are partially cross-classfied. Since the algorithm for isotonic regression can be extended to more than two variables, an extension of the proposed method is also possible.

We note that our attention has been confined to estimation but it would be valuable to consider testing for stochastic ordering between incomplete tables. The MLEs under stochastic order constraint can be used for likelihood ratio statistics for testing stochastic ordering between two multivariate multinomial populations. For complete data Lucas and Wright(1991) suggested testing procedures for testing stochastic ordering.

We suggest as further research that the EM method under a stochastic ordering may be explored under a nonignorable missing mechanism. The method may be useful for model selection for more than 2 dimensional contigency tables.

References

[1] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions,* Wiley, New York.

[2] Barlow, R.E., and Brunk, H.D. (1972). The Isotonic Regression Problem and Its Dual, *Journal of the American Statistical Association,* Vol. 67, 140-147.

[3] Bril, G., Dykstra, R., Pillers, C., and Robertson, T. (1984). Isotonic Regression in Two independent Variables, *Applied Statistics,* Vol. 33, 352-357.

[4] Brunk, H.D., Franck, W.E., Hanson, D.L., and Hogg, R.V. (1966). Maximum Likelihood Estimation of the Distribution of Two Stochastically Ordered Random Variables, *Journal of the American Statistical Association,* Vol. 61, 1067-1080.

[5] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with Discussion). *Journal of Royal Statistical Society,* Ser B, Vol. 39,1-38.

[6] Dykstra, R.L. (1982). Maximum Likelihood Estimation of the Survival Functions of Stochastically Ordered Random Variables, *Journal of the American Statistical Association,* Vol. 77, 621-628.

[7] Dykstra, R.L. and Robertson, T. (1982). An Algorithm for Isotonic Regression for Two or More Independent Variables, *The Annals of Statistics,* Vol. 10, 708-716.

[8] Fienberg, S.E. (1977). *The Analysis of Cross-Classified Categorical Data,* The MIT Press.

[9] Fuchs, C. (1982). Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data, *Journal of the American Statistical Association,* Vol. 77, 270-278.

[10] Haberman, S.J. (1974). Loglinear Models for Frequency Tables Derived by Indirect Observation: Maximum Likelihood Equations, *The Annals of Statistics,* Vol. 2, 911-924.

[11] Lee, C.C. (1987). Maximum Likelihood Estimates for Stochastically Ordered Multinomial Populations with Fixed and Random Zeros, *Foundations of Statistical Inference,* 187-197.

[12] Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data,* John Wiley & Sons, New York.

[13] Lucas, L.A. and Wright, F.T. (1991). Testing for and Against a Stochastic Ordering Between Multivariate Multinomial Populations, *Journal of Multivariate Analysis,* Vol. 38, 167-186.

[14] Meng, X.L. and Rubin, D.B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework, *Biometrika,* Vol. 80, 267-278.

[15] Robertson, T. and Wright, F.T. (1974). On the Maximum Likelihood Estimation of Stochastically Ordered Random Variates, *The Annals of Statistics,* Vol. 2,

528-534.

[16] Robertson, T. and Wright, F.T. (1981). Likelihood Ratio Tests for and Against a Stochastic Ordering Between Multinomial Populations, *The Annals of Statistics,* Vol. 9, 1248-1257.

[17] Sampson, A.R. and Whitaker, L.R. (1989). Estimation of Multivariate Distributions Under Stochastic Ordering, *Journal of the American Statistical Association,* Vol. 84, 541-548.

[18] Wu, C.F.J. (1983). On the Convergence Properties of the EM Algorithm, *Annals of Statistics,* Vol. 11, 95-103.

확률적 순서를 갖는 다변량분포에서 불완전자료에 의한 추정[3)]

정광모[4)]

요약

확률적 순서관계를 갖는 다변량분포에서 얻어진 자료가 결측값을 갖는 불완전한 자료일 때, EM 알고리즘을 이용한 최우추정법을 논의하였다. 본 논문에서는 관찰값들이 부분적으로 분류된 분할표자료에 국한하여 연구되었으며 기존의 동위회귀추정 프로그램을 써서 EM을 수행할 수 있는 이점이 있다. 예를 통하여 제안된 추정법을 설명한다.

3) 이 논문은 1992년도 교육부지원 한국학술진흥재단의 지방대육성과제 학술연구 조성비에 의하여 연구되었음.

4) (609-735) 부산시 금정구 부산대학교 통계학과.