Bayesian Inference for Categorical Data with Misclassification Errors *

Masahiro KURODA
Department of Computer Science and Mathematics,
Kurashiki University of Science and the Arts,
Nishinoura 2640 Turajima-cho Kurasihi-shi,
Okayama 712-8505, Japan,
e-mail: kuroda@soft.kusa.ac.jp

Zhi GENG
Department of Probability and Statistics,
Peking University,
Beijing, 100871, China,
e-mail: zgeng@math.pku.edu.cn

Abstract

In epidemiological studies, observed data are often collected subject to misclassification errors. In this paper, we discuss the Bayesian estimation for contingency table with misclassification errors. Employing the exact Bayesian computations to obtain posterior means as estimates, we are faced with computational difficulties. In order to find the posterior distribution, we apply the data augmentation(DA) algorithm to misclassified categorical data.

Keywords: Misclassification, Bayesian estimation, data augmentation algorithm, contingency table, posterior means.

1 Introduction

In epidemiological studies, observed data are often collected subject to misclassification errors. Such misclassification errors cause bias of estimation and reduce efficiency in the analysis of contingency tables. Many investigators have discussed how to adjust for the effects of misclassification, see the review papers, Chen (1989) and Walter and Irwig (1987). Espeland and Odoroff (1985) discussed the maximum likelihood estimation for a recursive system of log-linear models based on double sampling schemes by the EM algorithm. Espeland and Hui (1987) applied the Fisher scoring algorithm to evaluating variances and covariances of estimates. From the Bayesian viewpoint, Geng and Asano (1989) considered estimation methods for misclassified categorical data making use of prior information and double sampling schemes, and obtained

*published in *ADVANCES IN STATISTICS, COMBINATORICS AND RELATED AREAS*, (edited by Chandra Gulati, Yan-Xia Lin, John Rayner and Satya Mishra), World Scientific Publishing, 143 - 151.

posterior means as estimates. Viana (1994) applied Bayesian computations based on the matrix of misclassification probabilities to small-sample multinomial data. Evans, Guttman, Hatiovsky and Swartz (1996) discussed the implementation of the Gauss-Jacobi quadrature and the Gibbs sampling algorithm for the posterior analysis of binary response data with misclassification.

In this paper, we present a Bayesian approach that utilizes prior knowledge about misclassification and incorporates this prior knowledge with observations subject to misclassification. Although the EM algorithm or the Fisher scoring algorithm are often applied to estimating model parameters, these algorithms can not evaluate posterior distributions on the model parameters. Furthermore, these algorithms do not apply to our estimation problem because of the unidentifiability of the model parameters. However, our Bayesian approach, assuming a prior distribution on the model parameters, can overcome these problems. In order to find the posterior distribution of model parameters and calculate posterior means as estimates of them, we use the data augmentation(DA) algorithm by Tanner and Wong (1987).

In Section 2, we show the Bayesian computation to find a posterior distribution given misclassified observed data. In Section 3, we give the DA algorithm to approximate the posterior distribution, because of difficulties with the calculation of the posterior distribution. Section 4 presents two numerical experiments to examine the performance of the DA algorithm.

2 Misclassified observed data and Bayesian computation

Let X and Y be categorical variables having I and J categories, respectively, and let Y' be a misclassified variable of Y having K categories. We assume that two types of misclassified data are observed: (i) data for X and Y', denoted as $n = \{n_{i+k} \mid i \in \{1,\ldots,I\}, k \in \{1,\ldots,K\}\}$, and (ii) data for Y and Y', denoted as $m = \{m_{+jk} \mid j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\}\}$, where the symbol "+" means the sum over corresponding variables, for example, $n_{i+k} = \sum_j n_{ijk}$. Let p_{ijk} denote a probability for $(X,Y,Y') = (i,j,k)$ and $\theta_{XYY'} = \{p_{ijk} \mid i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\}\}$ denote a set of probabilities.

In this paper, the goal is to find the posterior distribution of model parameters $\theta_{XY} = \{p_{ij+} \mid i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}\}$ which are the marginal probabilities of X and Y, and obtain the posterior means of θ_{XY} as estimates.

Assume that n and m have independently multinomial distributions with parameters $\theta_{XY'} = \{p_{i+k} \mid i \in \{1,\ldots,I\}, k \in \{1,\ldots,K\}\}$ and $\theta_{YY'} = \{p_{+jk} \mid j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\}\}$, respectively, that is,

$$f(n \mid \theta_{XY'}) = \frac{n_{+++}!}{\prod_{i,k} n_{i+k}!} \prod_{i,k} p_{i+k}^{n_{i+k}},\ f(m \mid \theta_{YY'}) = \frac{m_{+++}!}{\prod_{j,k} m_{+jk}!} \prod_{i,k} p_{+jk}^{m_{+jk}}, \tag{1}$$

and that the prior distribution of $\theta_{XYY'}$ is a Dirichlet distribution which has the density function

$$\pi(\theta_{XYY'} \mid \alpha_{XYY'}) \;=\; \frac{\Gamma[\alpha_{+++}]}{\prod_{i,j,k}\Gamma[\alpha_{ijk}]}\prod_{i,j,k} p_{ijk}^{\alpha_{ijk}-1}, \tag{2}$$

where $\alpha_{XYY'} = \{\alpha_{ijk} \mid i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\}\}$ is a set of hyper-parameters of the prior distribution of $\theta_{XYY'}$. From the equations (1) and (2), we obtain the mixture posterior distribution given n and m. The posterior density is given by

$$\begin{aligned}
&\pi(\theta_{XYY'} \mid n, m) \\
&\quad\propto\; f(n \mid \theta_{XY'})\, f(m \mid \theta_{YY'})\, \pi(\theta_{XYY'} \mid \alpha_{XYY'}) = \prod_{i,k} p_{i+k}^{n_{i+k}} \prod_{j,k} p_{+jk}^{m_{+jk}} \prod_{i,j,k} p_{ijk}^{\alpha_{ijk}-1} \\
&\quad=\; \prod_k \left\{ \prod_i \sum_{\Omega(n)} \frac{n_{i+k}!}{\prod_j \tilde{n}_{ijk}!} \prod_j \sum_{\Omega(m)} \frac{m_{+jk}!}{\prod_i \tilde{m}_{ijk}!} p_{ijk}^{\alpha_{ijk}+\tilde{n}_{ijk}+\tilde{m}_{ijk}-1} \right\},
\end{aligned} \tag{3}$$

where $\sum_{\Omega(n)}$ denotes the sum over all possible $\{\tilde{n}_{ijk}\}$ under the conditions $\tilde{n}_{ijk} \geq 0$ for all i, j and k, and $\sum_j \tilde{n}_{ijk} = n_{i+k}$, and $\sum_{\Omega(m)}$ denotes the sum over all possible $\{\tilde{m}_{ijk}\}$ under the conditions $\tilde{m}_{ijk} \geq 0$ for all i, j and k, and $\sum_i \tilde{m}_{ijk} = m_{+jk}$. However, the posterior density (3) has a very complicated function, because of combinational explosion. It is extremely difficult to calculate exactly the posterior distribution and these calculation may take a long time when observed data are moderately large.

For the case that the incomplete-data posterior density is complicated as the equation (3) and the complete-data posterior density is relative easy to handle and draw from, the DA algorithm is very suitable.

In next section, we present the DA algorithm to approximate the posterior distribution to find the posterior distribution and estimate posterior means of model parameters θ_{XY},

3 DA algorithm for misclassified data

The DA algorithm consists of iterating between the *imputation*-step and the *posterior*-step. For this misclassified multinomial model, the DA algorithm is given by the following iterative scheme:

Imputation-step: Repeat the following two steps for $l = 1,\ldots,L$ to obtain the imputed data of n and m such as $\tilde{n} = \{\tilde{n}_{ijk} \mid i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\}, \sum_j \tilde{n}_{ijk} = n_{i+k}\}$ and $\tilde{m} = \{\tilde{m}_{ijk} \mid i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\}, \sum_i \tilde{m}_{ijk} = m_{+jk}\}$.

1. Generate cell probabilities $\{p^*_{ijk}\}$ from the current estimate of the posterior distribution,

2. Generate the imputed data $\tilde{n}^{(l)}$ and $\tilde{m}^{(l)}$ from the predictive distributions which have the conditional multinomial distributions, given n and m, with densities

$$\begin{aligned} f(\tilde{n} \mid \{p^*_{j|i,k}\}, n) &= \prod_{i,k} \frac{n_{i+k}!}{\prod_j \tilde{n}_{ijk}!} \prod_{i,j,k} p^{*\tilde{n}_{ijk}}_{j|i,k}, \\ f(\tilde{m} \mid \{p^*_{i|j,k}\}, m) &= \prod_{j,k} \frac{m_{i+k}!}{\prod_i \tilde{m}_{ijk}!} \prod_{i,j,k} p^{*\tilde{m}_{ijk}}_{i|j,k}, \end{aligned}$$

where $p^*_{j|i,k} = p^*_{ijk}/p^*_{i+k}$ and $p^*_{i|j,k} = p^*_{ijk}/p^*_{+jk}$.

Posterior-step: Update the current approximation of the posterior distribution of $\theta_{XYY'}$, given these imputed data $\tilde{n}^{(l)}$ and $\tilde{m}^{(l)}$, for $l = 1, \ldots, L$, by the Monte Carlo method,

$$\pi(\theta_{XYY'} \mid n, m) = \frac{1}{L} \sum_{l=1}^{L} \pi(\theta_{XYY'} \mid \tilde{n}^{(l)}, \tilde{m}^{(l)}).$$

Until the approximated distribution converges to a stationary distribution, the imputation-step and the posterior-step are iterated. Getting a stationary distribution, the values of L may be increased to improve the accuracy with respect to the Monte Carlo method.

Then we can find the posterior distribution of θ_{XY} and can obtain easily the posterior means and variances of them. Furthermore, it is possible to calculate the highest posterior density(HPD) region that is the Bayesian analog of the confidence intervals.

4 Numerical experiments

We provide two numerical experiments to examine the performance of the DA algorithm described the previous section.

4.1 Comparison of estimates and exact values

In the following numerical experiment, we compare the estimates obtained by the DA algorithm with the posterior means using the exact Bayesian calculation given by Geng and Asano (1987) regarding the data from Diamond and Lilienfeld (1962) that reported a case-control study concerning the circumcision status of male partners of woman with and without cervical cancer. The study sample was categorized by cervical cancer status, X (Case and Control), and self-reported circumcision status, Y' (Yes or No), in the left side of Table 1. In order to gain the information on the degree of misclassification of circumcision status, the supplemental sample concerning the relationship between actual circumcision status, Y (Yes or No), and Y' was gathered from the separate population shown by in the center of Table 1. Espeland and Hui (1987) described that, for the misclassified multinomial model, the conditional independence model between X and Y given Y' from the class of hierarchical log-linear models

was appropriate, since no observed data for X, Y and Y' were obtained. Furthermore, for the conditional independence model, Geng and Asano (1989) gave the hypothetical prior information shown in the right side of Table 1. We use their prior information as hyper-parameters $\alpha_{XYY'} = \{\alpha_{ijk} \mid i \in \{\text{Case, Control}\}, j \in \{\text{Yes, No}\}, k \in \{\text{Yes, No}\}\}$ and then obtain a posterior distribution and estimates of $\theta_{XY} = \{p_{ij+} \mid i \in \{\text{Case, Control}\}, j \in \{\text{Yes, No}\}\}$ for X and Y by posterior means.

In this numerical experiment, we evaluate the accuracy of the estimates using the DA algorithm in comparison with the exact posterior means given by Geng and Asano (1989). Table 2 shows the exact values, and the posterior means, the standard deviations(SDs) and the posterior 95 % credible intervals(CIs) of θ_{XY} obtained from simulated 10,000 samples after a *burn-in* of 1,000 samples. It can be seen that the estimates have approximately three-digit accuracy for the exact values. From the numerical results, we can see that the DA algorithm works quite well to estimate posterior means.

4.2 Performance of the DA algorithm

In the next numerical experiment, we examine the performance of the DA algorithm in comparison with the EM algorithm and the Fisher scoring algorithm. We apply the DA algorithm to the *double sampling* data from Hochberg (1977). The data were the highway safety research data relating the seat-belt usages to driver injuries. The main sample was of 80,084 accidents that were recorded by police subject to misclassification errors. The subsample was of 1,796 accidents that were recorded by both *imprecise* police reports and *precise* hospital interviews. Then, by the double sampling design, the subsample was randomly selected from the main sample. Thus, the subsample and the main sample have independent and identical distributions.

The main sample and the subsample in Table 3 were categorized by four variables X, X', Y and Y', where X and Y denote precise personal survey for seat-belt usages and driver injuries, and X' and Y' denote imprecise police reports for them.

In this experiment, we estimate model parameters under the saturated multinomial model, because our purpose is to investigate whether the DA algorithm is applicable to estimate model parameters, but not to analyze the misclassified observed data.

For these data, we assume that the main sample data and the subsample data have independent and identical multinomial distributions with

$$\theta_{XX'YY'} = \{p_{ijkl} \mid i \in \{\text{Yes, No}\}, j \in \{\text{Yes, No}\}, k \in \{\text{Yes, No}\}, l \in \{\text{Yes, No}\}\},$$

where $p_{ijkl} = \Pr(X = i, X' = j, Y = k, Y' = l)$ and the prior distribution for $\theta_{XX'YY'}$ has the Dirichlet distribution with hyper-parameters $\alpha_{XX'YY'} = \{\alpha_{ijkl} \mid i \in \{\text{Yes, No}\}, j \in \{\text{Yes, No}\}, k \in \{\text{Yes, No}\}, l \in \{\text{Yes, No}\}\}$. Then the model parameters are marginal probabili-

ties of X and Y,

$$\theta_{XY} = \{p_{i+k+} \mid i \in \{\text{Yes, No}\}, k \in \{\text{Yes, No}\}\},$$

where $p_{i+k+} = \sum_{j,l} p_{ijk}$. We utilize the subsample in Table 3 as hyper-parameters $\alpha_{XX'YY'}$ and obtain estimates of θ_{XY} by the DA algorithm. Table 4 shows the estimates and the SDs of θ_{XY} obtained by the DA algorithm, the exact Bayesian calculation, the Fisher scoring algorithm and the EM algorithm. The estimates using the DA algorithm can be found from simulated samples 100,000 after a burn-in samples 10,000 in two chains. The exact values of estimates of θ_{XY} using the Bayesian calculation are given by Geng and Asano (1989) who assumed the Jeffreys noninformative prior. The estimation using the Fisher scoring algorithm were carried out with ℓ_{EM} developed by Vermunt (1997).

From these numerical results, it can be seen that the DA algorithm has the equivalent performance of the EM and the Fisher scoring algorithm in comparison with these estimates and SDs. Applying the EM and the Fisher scoring algorithm, their algorithms have such disadvantages as it is impossible to find the posterior distribution of model parameters, may not be applied them to estimation of parameters owing to unidentifiability of the model and may be difficult to calculate the Fisher information matrix needed in the Fisher scoring algorithm.

5 Concluding remarks

In this paper, we discussed the DA algorithm to estimate model parameters for misclassified categorical data. We gave the posterior distribution by exact Bayesian computation. To avoid complicated calculation, we used the DA algorithm and find the posterior distribution. It is easily seen that the DA algorithm is the iterative simulation version of the EM algorithm that the imputation-step corresponds to the E-step and the posterior-step corresponds to the M-step.

In order to explore the possibility of parameter estimation by the DA algorithm, we provided two numerical experiments. In the first experiment, we evaluated accuracy of estimates in comparison with exact values. In the second experiment, we examined the performance of the DA algorithm. The results of both the two numerical experiments showed the advantage of applying the DA algorithm in terms of accuracy of estimates and in terms of algorithm simplicity to find the posterior distribution.

For the inference of multidimensional contingency tables, the Bayesian inference by the DA algorithm can be easily extended and also widely utilized. Then we may need to take account for conditional independence between variables in models. For parameters assuming the conditional independence model, A prior Dirichlet distribution that has hyper Markov laws by Dawid and Lauritzen (1993) is very suitable. Our future problem is how to incorporate prior information with the hyper-parameters of a hyper Dirichlet prior distribution without consistency.

Acknowledgements

The authors would like to thank an editor and a referee for helpful suggestions and correcting our paper.

References

[1] Chen, T.T. (1989). A review of methods for misclassified categorical data in epidemiology. Statistics in Medicine 8, 1095–1106.

[2] Dawid, A.P. and Lauritzen, S.L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. Ann. Statist. 21, 1272–1317.

[3] Diamond, E.L. and Lilienfeld, A.M. (1962). Effects of errors in classification and diagnosis in various types of epidemiological studies. American Journal of Public Health 52, 1137–1144.

[4] Espeland, M.A. and Odoroff, C.L. (1985). Log-linear models for doubly sampled categorical data fitted by the EM algorithm, Journal of the American Statistical Association 80, 663–670.

[5] Espeland, M.A. and Hui, S.L. (1987). A general approach to analyzing epidemiologic data that contain misclassification errors, Biometrics 43, 1001 – 1012.

[6] Evans, M., Guttman, I., Hatiovsky, Y. and Swartz, T. (1996). Bayesian analysis of binary data subject to misclassification, in Bayesian analysis in statistics and economics, (edited by Berry, D.A., Chaloner, K.M. and Geweke, J.K.), John Wiley & Sons, 67 – 77.

[7] Geng, Z. and Asano, Ch. (1989). Bayesian estimation methods for categorical data with misclassification, Communications in Statistics 8, 2935–2954.

[8] Hochberg, Y. (1977). On the use of double sampling schemes in analyzing categorical data with misclassification errors, Journal of the American Statistical Association 72, 914–921.

[9] Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation, Journal of the American Statistical Association 82, 528–540.

[10] Vermunt, J.K. (1997). ℓ_{EM}: A general program for the analysis of categorical data, Tilburg University.

[11] Viana, M.A.G. (1994). Bayesian small-sample estimation of misclassified multinomial data, Biometrics 50, 237 – 243.

[12] Walter, S.D. and Irwig, L.M. (1987). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review, Journal of Clinical Epidemiology 41, 923 – 937.

Table 1: Data from Diamond and Lilienfeld (1962) and hypothetical prior information $\alpha_{XYY'}$

Y'	Y	Study Sample X		Supple. Sample X	Hypothetical prior X	
		Case	Control	Unknown	Case	Control
Yes	Yes	5	14	37	80	10
	No			19	20	40
No	Yes	95	86	47	40	20
	No			89	10	80

Table 2: Posterior means and SDs and 95% CIs using the DA algorithm and the exact posterior means given by Geng and Asano(1989)

X	Y	Exact Bayes	DA	
		Posterior means	Posterior means ± SD	CI (lower-upper)
Case	Yes	0.3794	0.3786 ± 0.0127	0.3512 - 0.4017
	No	0.1116	0.1134 ± 0.0159	0.0838 - 0.1460
Control	Yes	0.0921	0.0927 ± 0.0107	0.0737 - 0.1142
	No	0.4169	0.4152 ± 0.0113	0.3916 - 0.4364

Table 3: Data of highway safety research(Hochberg, 1977)

Y'	Y	Main Sample X' = Yes	Main Sample X' = No	Subsample X' = Yes X = Yes	Subsample X' = Yes X = No	Subsample X' = No X = Yes	Subsample X' = No X = No
Yes	Yes	1196	13562	17	3	10	258
	No			3	4	4	25
No	Yes	7151	58175	16	3	25	194
	No			100	13	107	1014

Table 4: Estimates and their SDs of θ_{XY}

X	Y	Exact Bayes	DA	Fisher scoring	EM
		Posterior means ± SD	Posterior means ± SD	Estimates ± SD	
Yes	Yes	0.0397 ± 0.0043	0.0389 ± 0.0041	0.0394 ± 0.0045	0.0394
	No	0.1293 ± 0.0065	0.1311 ± 0.0073	0.1190 ± 0.0076	0.1294
No	Yes	0.2558 ± 0.0079	0.2577 ± 0.0078	0.2563 ± 0.0103	0.2559
	No	0.5752 ± 0.0093	0.5722 ± 0.0092	0.5870 ± 0.0116	0.5752