RACSAM
Rev. R. Acad. Cien. Serie A. Mat.
VOL. **101 (**1), 2007, pp. 71–80
Estadística e Investigación Operativa / Statistics and Operations Research
Artículo panorámico / Survey

Misclassified multinomial data: a Bayesian approach

C. J. Pérez, F. J. Girón, J. Martín, M. Ruiz, and C. Rojano

Abstract. In this paper, the problem of inference with misclassified multinomial data is addressed. Over the last years there has been a significant upsurge of interest in the development of Bayesian methods to make inferences with misclassified data. The wide range of applications for several sampling schemes and the importance of including initial information make Bayesian analysis an essential tool to be used in this context. A review of the existing literature followed by a methodological discussion is presented in this paper.

Datos Multinomiales Imperfectos: Un Enfoque Bayesiano

Resumen. En este artículo se trata el problema de la inferencia con datos multinomiales imperfectos. Durante los últimos años ha habido un resurgimiento del interés en el desarrollo de métodos bayesianos para hacer inferencias con datos imperfectos. El análisis Bayesiano se ha convertido en una herramienta fundamental en este contexto debido al gran número de aplicaciones existentes para diversos tipos de muestreo y a la importancia de incorporar información a priori. Se presenta una revisión de la literatura existente seguida de una discusión metodológica con ejemplos.

1. Introduction

Classification is a task performed quite naturally by human beings. Many activities have classification at their foundation. It is possible to classify the individuals of a population in multiple ways. Examples include classifying subjects by sex (male/female), smoking status (smoker/non-smoker), health status (ill/sane), and so on. These examples are based on dichotomous data. Multiple classifications are also possible, e.g. individuals may be classified in the following groups: (A) single, (B) married, (C) divorced, and (D) widowed. Classifications may also be carried out by two or more criteria.

When information is collected in the real world, the data do not often reflect the true status of the elements in the sample, i.e. the data-generating process is often noisy. This fact can happen due to several causes. In consumer surveys, consumers may not remember their previous behaviors accurately, may misunderstand survey questions or may intentionally misreport. In medical diagnosis, test failures and miscoded information are causes of distortion. In election surveys, voters are often reluctant to provide their true opinions. These and other situations occur in many other applications. The main consequence is that such distortions or noises can have an important effect on inferences because the effective amount of information obtained from the sample is considerably reduced. If the noise in a data-generating process is not appropriately modeled, the information may be perceived as being more accurate than it actually is,

Presentado por Sixto Ríos García.
Recibido: 16 de enero de 2007. *Aceptado*: 7 de marzo de 2007.
Palabras clave / Keywords: Bayesian inference, contingency tables, MCMC, misclassified data, multinomial sampling
Mathematics Subject Classifications: 62F15

leading in many cases to a non-optimal decision making. The underlying statistical problem is known as inference with misclassified data.

The effects of ignoring misclassification were first noted by Bross [5], who showed that classical estimators base sampling on a dichotomous process under the assumption of known noise parameters. These parameters are needed to correct the bias resulting from estimation based on the observed proportion. One way of assessing the extent of the misclassification with a particular method of observation is to compare it to a definitive method (also called gold standard or infallible) which is error free. A portion of the sample is processed by an infallible classifier, and then it is cross-classified by both fallible and infallible classifiers. This fact allows to estimate false positive and false negative rates. This technique is known as double sampling. A maximum likelihood approach to double sampling was presented by Tenenbein [29] and Tenenbein [30] and generalized by Hochberg [16], Chen [6], and Ekholm and Palmgren [7] among others. However, double sampling is not the panacea because gold standard methods are often not available for many applications, can be prohibitively expensive or can be misdefined.

Many investigators (as the previous ones) have considered misclassified data from the frequentist viewpoint (see Chen [6] and Walter and Irwing [32] for two particular reviews). However, over the last years there has been a significant upsurge of interest in the development of Bayesian methods to make inferences on misclassified data. The reason is that Bayesian methodology provides a complete paradigm for statistical inference under uncertainty that allows to combine information derived from observations with information elicited from experts (a summary of some Bayesian activity is presented by Berger [3]). The incorporation of relevant initial information through the prior distribution is determinant for this kind of problems. Also, the advent of Markov Chain Monte Carlo (MCMC) methods has been decisive in the growing of Bayesian literature in general, and, in the increase of Bayesian literature for misclassified data in particular. The implementation of these simulation-based techniques allows to obtain numerical solutions for problems based on truly complex models (see Brooks [4] for a review and Gilks et al. [14] for a monograph). In this paper, inference with misclassified multinomial data is addressed from a Bayesian viewpoint.

The rest of the article is organized as follows. Section 2 presents a brief review of Bayesian inference for misclassified binary data. Section 3 discusses the multinomial case, including an illustrative example. Finally, the conclusion is presented in Section 4.

2. Binomial data

Inference with misclassified data has been widely studied for data obtained from a dichotomous process. This kind of process is usually modeled by using Bernoulli distributions. The usual presence of noise complicates the process of making inferences on the proportion of individuals presenting a certain characteristic.

The problem of inference with misclassified binary data has been addressed from a frequentist viewpoint. Several techniques have been developed. All of them provide joint information about the parameter of interest and the noise parameters, but not for each group of parameters separately. By performing a likelihood analysis, Winkler and Gaba [34] found a nonidentifiability problem in a binomial model with an unknown noise parameter. This problem can be overcome by using a Bayesian approach. Furthermore, Bayesian inference can provide separate information for each group of parameters (proportion and noises) through their respective posterior distributions. Although the posterior distribution may be difficult (or impossible) to calculate directly, the aforementioned MCMC methods may be used. Sometimes, it is necessary to introduce auxiliary or latent random variables to make the generating process easier. Note that no closed-form expressions are obtained for these procedures because they are iterative, but the quantities of interest can be easily estimated. Another advantage of the Bayesian approach is that initial information can be used. Any information about the parameter of interest or the noise parameters can be included into the model through the prior distribution. For these reasons, Bayesian modeling allows a flexible handling of misclassified binary data.

The following literature review does not pretend to be exhaustive; however, it reflects the very important advances in the Bayesian treatment of binomial misclassified data under several sampling schemes. In 1976, Lindley and Phillips [21] presented an explanatory article on Bernoulli processes without misclassification from a Bayesian perspective. This paper shows the foundations of the Bayesian approach and its advantage over the frequentist approach in this simple problem. The consequence of ignoring misclassification in this model was noted by Winkler [33]. He indicated that the noise should be adequately modeled to avoid information loss. The effect of noise is generally to shift the posterior distribution toward the prior and to increase its dispersion. A Bayesian model with a single unknown noise parameter was presented by Winkler and Gaba [34]. They noticed a nonidentifiability problem by performing a likelihood analysis. This problem was avoided by using a Bayesian model with a joint prior distribution for the parameter of interest and the noise parameter. An extension to two unknown noise parameters was presented by Johnson and Gastwirth [17]. They described a large sample-based Bayesian model in the context of screening data for rare diseases. Later, Gaba and Winkler [11] presented a general Bayesian model with two unknown noise parameters. Both Johnson and Gastwirth [17] and Gaba and Winkler [11] considered that the noise parameters are independent of the proportion in the prior distribution[1]. The extension to the dependent case was presented by Gaba [10], who described a Bayesian model formalizing a prior dependence between the proportion and the noise parameter. A Bayesian approach to double sampling incorporating covariates was presented by York et al. [35]. They used MCMC methods to approximate the posterior distribution. Evans et al. [8] extended the analysis from Gaba and Winkler [11] to the multiple measurement case and to the comparison of two proportions for independent populations. Through some particular examples, they showed that the Gauss-Jacobi quadrature requires less computation to reach the same level of accuracy as the Gibbs sampling. The Bayesian sample size determination was investigated by Rahme et al. [24], whereas Kuo and Yang [18] described a Bayesian nonparametric approach using mixtures of Dirichlet processes to model the misclassified binary data. Ruiz et al. [26] developed a Bayesian model to make inferences on the parameters and on the true status for a particular individual. They introduce auxiliary variables to obtain an easy-to-implement Gibbs sampling-based algorithm. Paulino et al. [23] and Achcar et al. [1] studied misclassified binary data in the presence of covariates by assuming a logistic regression model. Finally, it is remarkable the work of Paulino et al. [22] that presented a random effect binary logistic regresion model.

Although the range of applications for binomial sampling with misclassified data is very wide, most authors have focused on medical applications. The special importance of decisions in the medical context has made this kind of analysis necessary. Examples in medical settings can be found in some of the previous references ([1,8,17,18,22,23]). A particular example of diagnostic test presented in Ruiz et al. [26] is used to illustrate the Bayesian analysis of misclassified binary data.

Example 1 (Diagnostic test) *The confirmation of many chronic diseases is a complex and expensive task. It is often necessary to propose screening tests that can be rapidly and economically applied to many people. However, these tests are not error-free. Table* 1 *shows the general representation of a diagnostic test.*

		Disease (Reality)		
		$+$	$-$	Total
Test (Classification)	$+$	True $+$ (a)	False $+$ (b)	$a+b$
	$-$	False $-$ (c)	True $-$ (d)	$c+d$
	Total	$a+c$	$b+d$	n

Table 1: General representation of a diagnostic test.

For every diagnostic test there are two critical values that determine its accuracy: sensitivity and specificity. Sensitivity is the probability that a test turns out to be positive, given that the person has the disease, while specificity is the probability that a test turns out to be negative, given that the person does not have

[1] Some remarks on independence applicable to these kinds of models can be found in Girón et al. [15].

the disease. A valid test should have both a high sensitivity and specificity. Positive and negative predictive values are also very important in this context. For more information about diagnostic tests, see Altman [2].

As an example, assume that in a study about AIDS a random sample of $192{,}415$ *individuals is selected from a particular population and that* 20 *individuals are classified as disease-affected by screening their blood samples. Assume that a process with misclassified data is modeled for this application as in Ruiz et al.* [26]. *Let* θ *be the prevalence of the disease, and let* λ_1 *and* λ_0 *be the noise parameters (sensitivity and false positive coefficient, respectively). So, the likelihood is:*

$$L(\theta, \lambda_1, \lambda_0) \propto (\theta\lambda_1 + (1-\theta)\lambda_0)^{20}(\theta(1-\lambda_1) + (1-\theta)(1-\lambda_0))^{192395}.$$

The prior densities proposed by Johnson and Gastwirth [17] *are* $\theta \sim \mathrm{Be}(15, 94092)$, $\lambda_1 \sim \mathrm{Be}(142, 1)$ *and* $\lambda_0 \sim \mathrm{Be}(3, 1363)$. *In this case, the prior distributions are conjugate for each parameter. This fact allows for an easy implementation of the posterior generation process. The model is also applicable with non-conjugate distributions.*

After the chain has been considered to have converged, a sample of size 100000 *is generated from the posterior distributions of* θ, λ_1, *and* λ_0. *Histograms for the posterior distributions of* θ, λ_0, *and* λ_1 *are represented in Figure* 1. *A simple look at the figures shows the main characteristics of the prior and the posterior distributions for the proportion and the noise parameters. Note that the prior distribution for* λ_0 *is substantially transformed by the data into the posterior distribution. Bayesian analysis provides inference not only on the proportion of AIDS but also on the sensitivity* λ_1 *and specificity* $1 - \lambda_0$.

In many applications, data are classified in more than two categories. Then, a generalization of the problem considering multinomial data is useful in practice. Next section addresses the problem of inference with misclassified multinomial data.

3. Multinomial data

Many fewer developments have been realized for the multinomial sampling case. Here, the main interest is focused on making inferences on the proportions for the m mutually exclusive categories the individuals are classified in. Inferences on the noise parameters are also interesting. The advantages of the Bayesian modeling pointed out formerly still hold for this more general case.

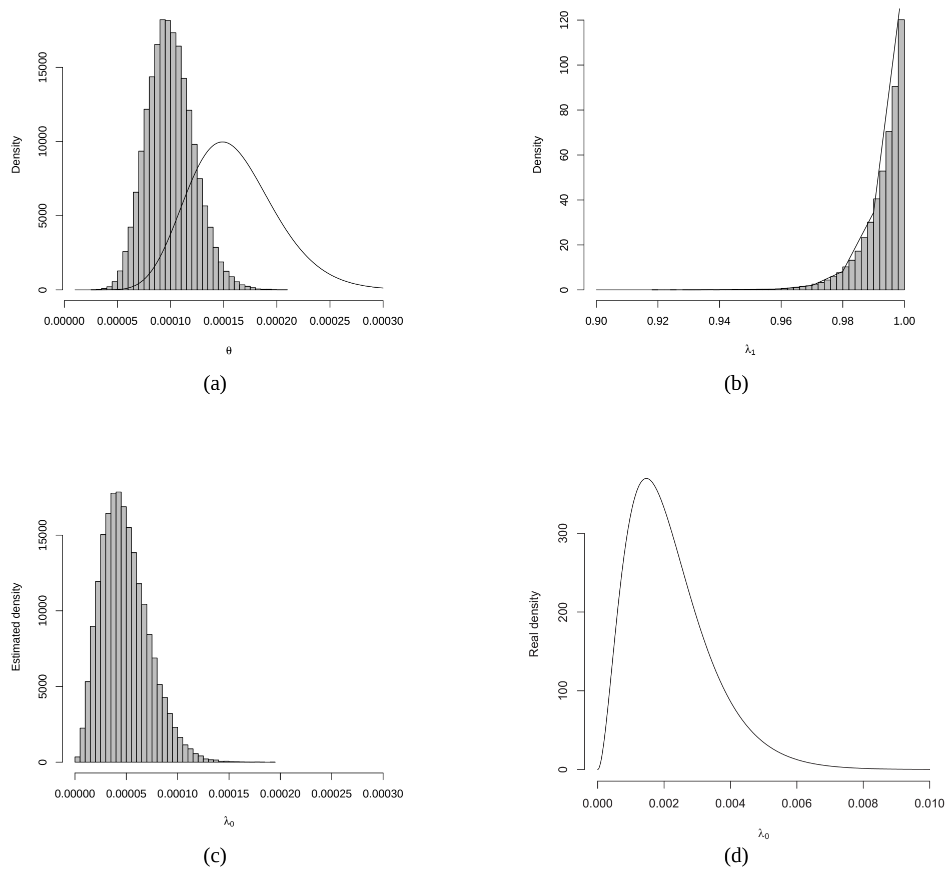
Firstly, the case where the individuals are classified by only one criterion is considered. Then, only one classification variable is involved. Ruiz et al. [25] developed a Bayesian model for multinomial sampling with misclassified data that is a generalization of the model described in Ruiz et al. [26]. A brief description of the model is exposed in order to show its advantages through an illustrative example.

A population P is divided into m disjoint categories denoted by A_1, A_2, ..., A_m. O_n represents an observed sample of size n, and the individuals from O_n can be misclassified. The main interest is focused on making inferences on the proportions θ_1, θ_2, ..., θ_m of individuals belonging to the classes A_1, A_2, ..., A_m. The noise parameters are characterized by the transition matrix:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \cdots & \lambda_{mm} \end{pmatrix},$$

where λ_{ij} denotes the probability that an individual from A_i is classified in A_j. Then, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)'$ denotes the vector of proportions and $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{im})'$ denotes the transition vector for the class A_i ($i = 1, 2, \ldots, m$).

A particular notation is introduced into the model to differentiate between the observed and the true number of individuals in each class. $\boldsymbol{X}$ ($\boldsymbol{Y}$) is defined as a random vector representing the number of



(a) (b)

(c) (d)

Figure 1: Prior distributions (solid line) and estimated posterior distributions (histogram) for θ, λ_1, and λ_0.

individuals that belong to (that are classified in) each category. The objective is to make inferences on $\boldsymbol{X}$, $\boldsymbol{\theta}$, and $\boldsymbol{\lambda}_i$, $i = 1, 2,\ldots, m$ by using the information given by the observed data and the prior knowledge.

Some hypothesis about conditional independence and the distribution of some random vectors involved in the model are assumed. Then, the likelihood is:

$$f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\Lambda}) = \frac{n!}{y_1! \cdots y_m!} \prod_{j=1}^{m} \left(\sum_{i=1}^{m} \theta_i \lambda_{ij} \right)^{y_j} .$$

Due to the difficulty to deal with the likelihood (and hence with the posterior distribution), auxiliary or latent random vectors are introduced. This fact allows to develop an easy-to-implement Gibbs sampling-based algorithm in order to generate samples from distributions of interest. Then, inferences on the proportion and noise parameters can be carried out. Monte Carlo standard error estimates can also be obtained.

The initial information about the parameters is included into the model through the prior distribution. The prior densities for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}_i$ depend on the hyperparameters $\boldsymbol{\rho}$ and $\boldsymbol{\pi}_i$, $i = 1, 2, \ldots, m$, respectively. For each set of hyperparameters $\boldsymbol{\pi} = \{\boldsymbol{\rho}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_m\}$, the joint prior distribution is assumed to be:

$$f(\boldsymbol{\theta}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_m|\boldsymbol{\pi}) = f(\boldsymbol{\theta}|\boldsymbol{\rho}) \prod_{i=1}^{m} f(\boldsymbol{\lambda}_i|\boldsymbol{\pi}_i).$$

When conjugate prior distributions are used, the generating process is even easier.

This model can be applied to a wide range of applications. An illustrative example related to elections is presented in Ruiz et al. [25] and discussed here.

Example 2 *Opinion surveys typically provide misclassified information. Voters reluctance to provide their true opinions and possible opinion changes are the two main reasons why observed data do not reflect the true vote intention.*

The information obtained by a survey related to the 2003 *local election, which took place in Málaga (Spain), is analyzed. Note that the political situation was tense when the survey was made because of some events affecting the national government (PP party) like Iraq war or the Prestige case (sinking of an oil ship in northwest Spain). However, the economical situation of the country was relatively satisfactory. These facts influenced the vote changes for the local election.*

The survey was conducted on a sample of 420 *individuals, and the results were:* 186 *for PP (*A_1*),* 152 *for PSOE (*A_2*),* 49 *for IU (*A_3*),* 20 *for PA (*A_4*), and* 13 *for the remaining minority parties (*A_5*). Then, the observed vector is* $\boldsymbol{y} = (186, 152, 49, 20, 13)'$*. The vector* $\boldsymbol{X}$ *representing the true number of votes for each party is assumed to follow a multinomial law with parameter* $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)'$ *and its assigned prior density is a Dirichlet with hyperparameter vector* $\boldsymbol{\rho} = (4, 3, 1, 0.5, 0.3)$*. In our opinion this density reflects reasonably well the prior information on the political evolution of this city and the country from the last local elections. Also, this prior knowledge is adequately reflected on the transition vectors, whose prior densities are Dirichlet laws characterized by the sets of hyperparameters given in the rows of the following matrix:*

$$\boldsymbol{A} = \begin{pmatrix} 6 & 0.5 & 0.02 & 0.3 & 0.05 \\ 0.1 & 5 & 0.4 & 0.1 & 0.05 \\ 0.01 & 0.1 & 6 & 0.05 & 0.1 \\ 0.1 & 0.5 & 0.05 & 5 & 0.1 \\ 0.05 & 0.3 & 0.3 & 0.3 & 5 \end{pmatrix}$$

A sample of size 5000 *was generated after the chain was considered to have converged. Figure* 2 *represents the scatter-plot and the histograms for the generated proportions of the two main parties, whereas the estimations for parameter* $\boldsymbol{\theta}$ *by using this model and a noise-free one are represented in Table* 2*. The true proportions (obtained after the election results were known) for each party are also represented.*

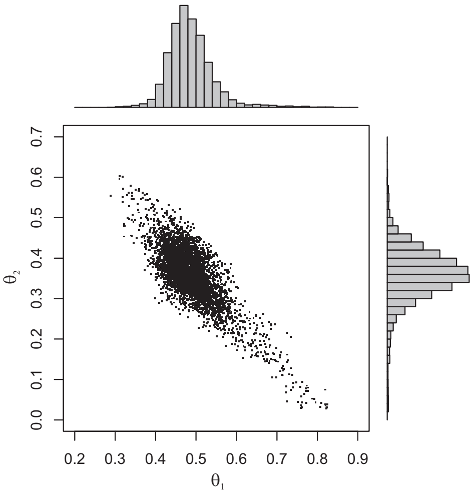
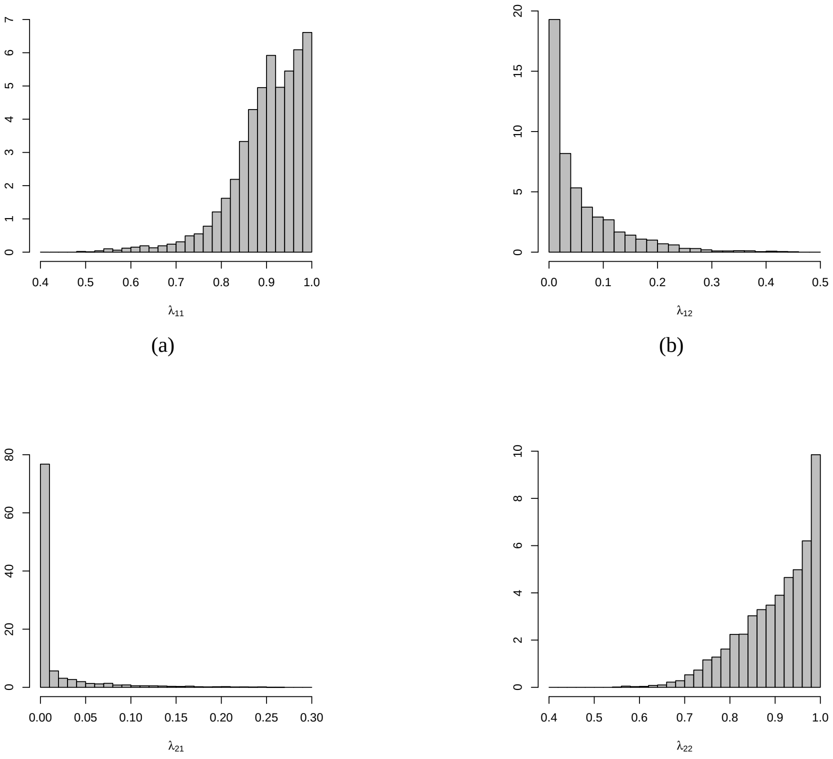


Figure 2: Scatter-plot and histograms for the generated vote proportions of A_1 and A_2.

	θ_1	θ_2	θ_3	θ_4	θ_5
True proportion	0.499	0.347	0.085	0.042	0.027
Noise model	0.491	0.356	0.094	0.033	0.026
Noise-free model	0.443	0.361	0.117	0.048	0.031

Table 2: Comparison of estimates by using the proposed noise model and a noise-free one.

Finally, Figure 3 *represents the generated λ_{ij} for the two main parties. Note that Figure* 3 (b) *shows a high proportion of hidden vote in favor of A_1, while Figure* 3 (c) *shows the opposite behavior for A_2. This is in agreement with what happened when the voting results were obtained.*



(a)

(b)

(c)

(d)

Figure 3: Generated data for (a) λ_{11} (b) λ_{12} (c) λ_{21} and (d) λ_{22}.

This model could be linked to a generalized linear model to allow for the introduction of covariates. The presence of covariates in a Bayesian approach to model misclassified data would be very interesting. For example, in the election context, a covariate could be the real vote of the surveyed individual in the last elections. Furthermore, it could be considered that the covariate has been measured with error. Age or social status could be two more covariates susceptible to be included in the model.

Two more references on Bayesian models for the treatment of misclassified multinomial data are Viana [31] and Swartz et al. [27]. Viana [31] applied Bayesian computations based on the matrix of misclassification probabilities to small-sample multinomial data. Swartz et al. [27] extended the work of Evans et al. [8]. They also discussed on the nonidentifiability problem and made emphasis on practical concerns for Bayesian analysis involving nonidentifiability.

Sometimes, the population is divided into some subpopulations and the interest is focused on estimating the proportion and the noise parameters separately for each subpopulation. The model to be used in this context could be an extension of the hierarchical model described by Gelman et al. [12] for only one proportion and, also, an extension of the model given in Ruiz et al. [25].

This model could be used to analyze misclassified data obtained from a random sample of size n_l for each subpopulation P_l, $l = 1, 2, \ldots, L$ of a population P. Now, $\boldsymbol{\theta}_l = (\theta_{l1}, \theta_{l2}, \ldots, \theta_{lm})'$ and Λ_l denote the proportion vector and the transition matrix for the subpopulation l, respectively.

Two situations can be considered. First, the case where the transition matrix is shared by all the subpopulations, that is, $\Lambda_l = \Lambda$ for $l = 1, 2, \ldots, L$. For many problems, it is common that the data are affected by misclassification in a similar way for different subpopulations. For example, in an election context, the same transition matrix could be used by two or more provinces into one region. The model applicable in this context is a direct extension of the one given in Ruiz et al. [25]. This new model maintains the same good properties described for the one population case. The second case eliminates the hypothesis of a global transition matrix, which may be restrictive in some situations. Furthermore, the transition matrices could be dependent on the corresponding proportion vector. The resulting hierarchical model should be easy to generate from, mainly by MCMC methods. This is the following step in the Bayesian treatment of multinomial data in several subpopulations. This model can be used when the classification errors in the subpopulations are heterogeneous.

Finally, when more than one criterion is considered to classify the individuals, contingency tables can be used. In this case, the relation among the two or more classification variables can be investigated rather than only focusing on the estimation of the proportions. Fleiss et al. [9] studied how misclassification of one variable affects measures of association. They also considered the case where both variables are observed with error.

Geng and Asano [13] considered Bayesian estimation methods for categorical data occurring in the form of a contingency table. They used a double sampling scheme representing observations in a contingency table categorized by error-free and error-prone categorical variables. The posterior means of probabilities in cells are used as estimators. When the posterior means are difficult to obtain directly, they use the expectation-maximization (EM) algorithm. Kuroda and Geng [19] noted that the EM algorithm can not evaluate posterior distribution on the model parameters for contingency tables in a non-double scheme sampling. They proposed a Bayesian approach that uses the data augmentation algorithm of Tanner and Wong [28] to estimate cell proportions by using posterior means. This approach overcomes the unindentifiability of the model parameters. Later, Kuroda and Nakagawa [20] proposed a more efficient data augmentation algorithm for graphical models from the computation time viewpoint.

4. Conclusion

Multinomial data subject to misclassification can be well handled by Bayesian methods. When available, the inclusion of initial relevant information into the model is very useful to model adequately the process for both proportion and noise parameters. Computational difficulties can be avoided by using simulation-based methods as MCMC. The increasing number of papers addressing applications on this topic shows its importance.

Acknowledgement. This research has been partially supported by *Consejería de Educación y Ciencia, Junta de Andalucía,* Spain (Group FQM 140) and by *Ministerio de Educación y Ciencia,* Spain (Project TSI2004-06801-C04-03)

References

[1] Achcar, J. A., Martínez, E. Z. and Louzada-Neto, F., (2004). *Proceedings of the Computational Statistics Conference*, chapter Binary data in the presence of misclassifications, 581–588. Physica-Verlag.

[2] Altman, D. G., (1991). *Practical Statistics for Medical Research*. Chapman and Hall.

[3] Berger, J. O., (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *J. Amer. Statist. Assoc.*, **95(452)**, 1269–1276.

[4] Brooks, S. P., (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69–100.

[5] Bross, I., (1954). Misclassification in 2×2 tables. *Biometrics*, **10**, 478–486.

[6] Chen, T. T., (1989). A review of methods for misclassified categorical data in epidemiology. *Stat. Med.*, **8**, 1095–1106.

[7] Ekholm, A. and Palmgren, J., (1987). Correction for misclassification using doubly sampled data. *Journal of Official Statistics*, **3**, 419–429.

[8] Evans, M., Guttman, I., Haitovsky, Y., and Swartz, T., (1996). *Bayesian Analysis in Statistics and Econometrics, Essays in Honor of Ärnold Zellner (eds. D. A. Berry, K. M. Chaloner and J. K. Geweke)*, chapter Bayesian Analysis of Binary Data Subject to Misclassification, 67–77. Wiley.

[9] Fleiss, J. L., Levin, B. and Paik, M. C., (2003). *Statistical Methods for Rates and Proportions*, chapter Misclassification: Effects, Control, and Adjustment. Mathematics and Statistics. Wiley.

[10] Gaba, A., (1993). Inferences with an unknown noise level in a Bernoulli process. *Management Science*, **39**, 1227–1237.

[11] Gaba, A. and Winkler, R. L., (1992). Implications of errors in survey data: A Bayesian approach. *Management Science*, **38**, 913–925.

[12] Gelman, A., Carlin J. B., Rubin, D. B. and Stern, H. S., (1998). *Bayesian Data Analysis*. Chapman & Hall.

[13] Geng, Z. and Asano, C., (1989). Bayesian estimation methods for categorical data with misclassification. *Comm. Statist. Theory Methods*, **18(8)**, 2935–2954.

[14] Gilks, W. R., Richarson, S. and Spiegelhalter, D. J., (1998). *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

[15] Girón, F. J., Kadane, J. B. and Moreno, E., (1997). Independence issues in imprecise data models: A Bayesian approach. *C. R. Math. Acad. Sci. Paris*, **324**, 1149–1153.

[16] Hochberg, Y., (1977). On the use of double sampling schemes in analyszing categorical data with misclassification errors. *J. Amer. Statist. Assoc.*, **72**, 914–921.

[17] Johnson, W. O. and Gastwirth, J. L., (1991). Bayesian inference for medical screening tests: Approximations useful for the analysis of acquired immune deficiency syndrome. *J. R. Stat. Soc. Ser. B. Stat Methodol*, **53**, 427–439.

[18] Kuo, L. and Yang, T. Y., (2002). Bayesian nonparametric approach to medical binary data with misclassification errors. In *Fourth biennial international conference on Statistics, Probability and Related Areas*, Northern Illinois University, Dekalb.

[19] Kuroda, M. and Geng, Z., (2002). *Advances in statistics, combinatorics and related areas (eds. C. Gulati, Y. Lin, J. Rayner, and S. Mishra)*, chapter Bayesian inference for categorical data with misclassification errors, 143–151. World Scientific Publishing.

[20] Kuroda, M. and Nakagawa, S., (2002). *Data augmentation algorithm in the analysis of contingency tables with misclassification*, chapter Proceedings of the Conference in Computational Statistics 2002. CompStat.

[21] Lindley, D. V. and Phillips, L. D., (1976). Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.*, **30(3)**, 112–119.

[22] Paulino, C. D., Silva, G. and Achcar, J. A., (2005). Bayesian analysis of correlated misclassified binary data. *Comput. Statist. Data Analysis*, **49**, 1120–1131.

[23] Paulino, C. D., Soares, P. and Neuhaus, J., (2003). Binomial regression with misclassification. *Biometrics*, **59**, 670–675.

[24] Rahme, E., Joseph, L. and Gyorkos, T. W., (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *J. R. Stat. Soc. Ser. C.*, **49**, 119–128.

[25] Ruiz, M., Girón, F. J., Pérez, C. J., Rojano, C. and Martín, J., (2003). A Bayesian model for multinomial sampling with misclassified data. Technical report, University of Málaga.

[26] Ruiz, M., Girón, F. J., Rojano, C., Pérez, C. J. and Martín, J., (2004). *Proceedings of the Computational Statistics Conference*, chapter A Bayesian model for Binomial Imperfect Sampling, 1709–1716. Physica-Verlag.

[27] Swartz, T., Haitovsky, Y., Vexler A. and Yang, T., (2004). Bayesian identifiability and misclassification in multinomial data. *Canad. J. Statist.*, **32(3)**, 285–302.

[28] Tanner, M. A. and Wong, W. H., (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.*, **82**, 528–540.

[29] Tenenbein, A., (1970). A double sampling scheme for estimating from binomial data with misclassification. *J. Amer. Statist. Assoc.*, **65**, 1350–1361.

[30] Tenenbein, A., (1972). A double sampling scheme for estimating from misclassified multinomial data with application to sampling inspection. *Technometrics*, **14**, 187–202.

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

[32] Walter, S. D. and Irwing, L. M., (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology*, **41(9)**, 923–937.

[33] Winkler, R. L., (1985). *Bayesian Statistics 2 (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith)*, chapter Information Loss in Noisy and Dependent Processes, 559–570. Elsevier Science Publishers.

[34] Winkler, R. L. and Gaba, A., (1990). *Bayesian and Likelihood Methods in Statistics and Econometrics: Essay in Honor of George A. Barnad (eds. S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner)*, chapter Inference with Imperfect Sampling from a Bernoulli Process, 303–317. Elsevier Science Publishers.

[35] York, J., Madigan, D., Heuch, I. and Lie, R. T., (1995). Birth defects registered by double sampling: a Bayesian approach incorporating covariates and model uncertainty. *Appl. Statist.*, **44**, 227–242.

C. J. Pérez J. Martín
Departamento de Matemáticas
Universidad de Extremadura
Avda. de la Universidad s/n,
10071 Cáceres, Spain.

F. J. Girón M. Ruiz C. Rojano
Departamento de Estadística e Investigación Operativa,
Universidad de Málaga
Campus de Teatinos, s/n,
29071 Málaga, Spain.