The Canadian Journal of Statistics
Vol. 32, No. 3, 2004, Pages 1–18
La revue canadienne de statistique

Bayesian identifiability and misclassification in multinomial data

Tim SWARTZ, Yoel HAITOVSKY, Albert VEXLER and Tae YANG

Key words and phrases: Convergence of Markov chains; Dirichlet priors; Gibbs sampling; latent variables; misclassification; nonidentifiability.

MSC 2000: Primary 62P10; secondary 62P20.

Abstract: The authors consider the Bayesian analysis of multinomial data in the presence of misclassification. Misclassification of the multinomial cell entries leads to problems of identifiability which are categorized into two types. The first type, referred to as the permutation-type nonidentifiabilities, may be handled with constraints that are suggested by the structure of the problem. Problems of identifiability of the second type are addressed with informative prior information via Dirichlet distributions. Computations are carried out using a Gibbs sampling algorithm.

Identifiabilité et erreurs de classification dans l'analyse bayésienne de données multinomiales

Résumé : Les auteurs s'intéressent à l'analyse bayésienne de données multinomiales en présence d'erreurs de classification. De telles erreurs causent des problèmes d'identifiabilité de deux types. Les problèmes d'identifiabilité de type permutationnel sont traités à l'aide de contraintes suggérées par le contexte. Les problèmes d'identifiabilité de l'autre type sont réglés par l'emploi de lois a priori de Dirichlet informatives. Les calculs sont effectués au moyen d'un algorithme d'échantillonneur de Gibbs.

1. INTRODUCTION

A widespread difficulty in drawing inference from categorical data is the existence of misclassification errors, i.e., the disagreement between the observed and the true classification. Misclassification errors occur in many fields of scientific inquiry, including the medical and social sciences, engineering and commercial applications. The problem is especially important in the health sciences where misdiagnoses are sometimes made: sick individuals may be diagnosed as healthy, healthy individuals may be diagnosed as sick, or the severity of the case may be misjudged. Misclassification in medical research is considered by, for example, Hadgu (1996) and Joseph, Gyorkos & Coupal (1995).

Another important area is marketing research where data in consumer surveys are sometimes collected in error (Peterson & Kerin 1981; Gaba & Winkler 1992). In addition to coding errors, consumers may intentionally misreport, may not remember their behaviour accurately, may misunderstand survey questions, etc. Examples include studies of magazine readership (Clancy, Ostlund & Wyner 1979), ad recognition (Singh & Churchill 1986), purchase behaviour (Wind & Lerner 1979) and purchase intentions (Kalwani & Silk 1982).

The effect of misclassification on estimation and hypothesis testing has been investigated by many authors. Bross (1954) set up the standard framework in the case of binomial data by introducing parameters that define the misclassification probabilities. He showed that although binomial data contain no information concerning the misclassification probabilities, profoundly biased estimators can occur when misclassification is ignored. Tenenbein (1972) extended the analysis to multinomial data and proposed a double sampling scheme to obtain information concerning the misclassification probabilities. Double sampling and its variants are now popular in handling misclassification problems; see Thall, Jacoby & Zimmerman (1996), Stewart et al. (1998), and the references therein. More recently, Yang & Kuo (2003) use mixtures of Dirichlet processes in the analysis of binary data subject to misclassification.

The basic difficulty in misclassification problems is that of identifiability. Without the presence of additional information beyond the raw data, it is not possible to take into account the effect of misclassification. Although double sampling approaches can provide the required additional information, a difficulty with double sampling is the necessity of an infallible classifier, which may not exist or may be prohibitively expensive. In this paper, we use a Bayesian approach where prior distributions provide the additional information needed to make sensible inference. Our approach to the analysis of misclassified multinomial data extends the work of Evans, Guttman, Haitovsky & Swartz (1996) who considered binomial data. Moreover, we use Markov chain methods to investigate the high dimensional posterior distributions which arise.

In Section 2, we begin with a discussion of identifiability, with an emphasis on Bayesian identifiability. Although most of the material in this section is not original, the discussion is necessary for the development of the multinomial misclassification model. In particular, we define permutation-type nonidentifiability and discuss the treatment of such nonidentifiabilities in the Bayesian model building process. In Section 3, we develop the multinomial misclassification model following the Bross (1954) framework. Four sets of constraints are proposed, each of which eliminates the permutation-type nonidentifiabilities that exist in the model. It is also demonstrated that in the case of a poor assessor, it may be possible to eliminate permutation-type nonidentifiabilities by switching diagnoses. A latent variable is then introduced which permits convenient posterior calculations via the Gibbs sampling algorithm. To help emphasize concepts, the simplest of the multinomial misclassification models is considered in Section 3. In Section 4, two more examples are analyzed. The first example is based on an artificial data set which sheds light on the performance of the model in higher dimensions. One of the practical lessons obtained from the example is that the permutation-type nonidentifiabilities become relatively less important as the dimension increases. The second example of Section 4 is based on actual medical assessments taken from Dawid & Skene (1979). The analysis of the Dawid & Skene (1979) data requires an extension of the model developed in Section 3. We conclude with a short discussion in Section 5.

2. BAYESIAN IDENTIFIABILITY

With advances in computer hardware and computational algorithms, statisticians are entertaining increasingly complex models. A consequence of this is that some models may not be well understood to the extent that nonidentifiabilities or near nonidentifiabilities may inadvertently creep in and cause havoc. A formal definition of nonidentifiability is given by Basu (1983).

DEFINITION 1. Let U be an observable random variable with distribution function F_θ and let F_θ belong to a family $\mathcal{F} = \{F_\theta : \theta \in \Omega\}$ of distribution functions indexed by a parameter θ. Here θ could be scalar or vector-valued. We say that θ is nonidentifiable by U if there is at least one pair (θ, θ'), $\theta \neq \theta'$, where θ and θ' both belong to Ω such that $F_\theta(u) = F_{\theta'}(u)$ for all u. In the contrary case we shall say θ is identifiable.

The concept of *near nonidentifiability* is less rigorous, but referring to the above definition, we may say that θ is nearly nonidentifiable if $F_\theta(u) \approx F_{\theta'}(u)$ for all u. In short, nonidentifiablility does not allow the data to distinguish between parameter values.

In classical settings, difficulties associated with nonidentifiability are widely recognized. For example, consider the common one-way ANOVA model. The response variables are given by $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, the overall mean is μ, the factor levels α_i are indexed by $i = 1, \ldots, k$, the replicates are indexed by $j = 1, \ldots, n_i$ and the ε_{ij} are independent and identically distributed Normal$(0, \sigma^2)$ random errors. The parameter vector $(\mu, \alpha_1, \ldots, \alpha_k, \sigma)$ is nonidentifiable since $(\mu', \alpha_1', \ldots, \alpha_k', \sigma') = (\mu - \tau, \alpha_1 + \tau, \ldots, \alpha_k + \tau, \sigma)$ gives the same probability distribution of the data for all values τ. As the model stands, a consequence is that one cannot sensibly estimate the parameters. A standard way to fix the problem is to introduce the constraint $\alpha_1 + \cdots + \alpha_k = 0$.

To initiate thought concerning nonidentifiability in Bayesian contexts, let us consider the following simple hierarchical model. Let $y_1, \ldots, y_n$ be a sample from the $\mathrm{Normal}(\mu, 1)$ distribution, and let the prior specification be $\mu \mid \alpha \sim \mathrm{Normal}(\alpha, 1)$ and $\alpha \sim \mathrm{Normal}(\alpha_0, 1)$, where α_0 is specified. Now this is a Bayesian model, and we therefore consider the posterior distribution of the full parameter $\theta = (\mu, \alpha)$, where μ is the primary parameter of interest. It is straightforward to show that the posterior distribution of θ is $\mathrm{Normal}_2(\lambda, \Sigma)$, where $(2n+1)\lambda_1 = 2n\overline{y} + \alpha_0$, $(2n+1)\lambda_2 = n\overline{y} + (n+1)\alpha_0$, $\sigma_{11} = 2/(2n+1)$, $\sigma_{22} = (n+1)/(2n+1)$ and $\sigma_{12} = 1/(2n+1)$. For large n, it follows that $\mathrm{E}(\mu \mid \underline{y}) \approx \overline{y}$, $\mathrm{E}(\alpha \mid \underline{y}) \approx (\overline{y} + \alpha_0)/2$, $\mathrm{var}(\mu \mid \underline{y}) \approx 0$, $\mathrm{var}(\alpha \mid \underline{y}) \approx 1/2$ and $\mathrm{corr}(\mu, \alpha \mid \underline{y}) \approx 0$. Clearly, there is no practical problem with this model. Yet, under the Basu (1983) definition, θ is nonidentifiable since $F_{(\mu,\alpha)} = F_{(\mu,\alpha')}$ for any $\alpha \neq \alpha'$. Moreover, nonidentifiability according to Basu (1983) exists for any hierarchical model!

In the example above, and more generally, it is the introduction of prior information in Bayesian settings that can rectify issues of nonidentifiability. In fact, there is a school of thought that suggests that nonidentifiability is never a problem for a Bayesian. One simply introduces a prior, obtains the posterior and integrates as required. Quoting Lindley (1971), "In passing it might be noted that unidentifiability causes no real difficulty in the Bayesian approach."

We take the view that nonidentifiability poses concerns for the Bayesian. The concerns can be both practical and philosophical. A practical concern for the Bayesian involving nonidentifiability or near nonidentifiability is that there tends to be strong correlations between parameters in the posterior distribution. The presence of strong correlations can result in the poor mixing of Markov chains used to explore the posterior, thus rendering estimators that may fail to converge within reasonable computing times. Another practical concern is that even large sample sizes may be unable to overcome prior information when nonidentifiabilities occur (Johnson, Gastwirth & Pearson 2001). We suggest that a philosophical concern occurs when the statistician imposes a flat prior, or more generally, a prior density that has constant values along the contours of the nonidentifiabilities. We believe that the statistician is guilty of nonsensical model building as neither the data nor the prior has the capacity to distinguish between parameters involved in a nonidentifiability. We suggest a principle of Bayesian model building where the prior is required to be informative with respect to nonidentifiability. This is in keeping with one of the basic themes espoused by Gustafson (2004).

Having indicated that nonidentifiability can also be problematic in Bayesian analyses, we argue that there are different types of nonidentifiability. We say that a *permutation-type nonidentifiability* exists when a permutation of the parameters $(\theta_1, \ldots, \theta_n)$ results in nonidentifiability. Permutation-type nonidentifiability typically has the effect that the likelihood has a symmetry that is counter to common sense. For example, it is unacceptable when an a priori plausible region of the parameter space induces $n! - 1$ regions of nonsensical parameters each with the same likelihood. This, in turn, may yield unrealistic inferences based on the posterior distribution. Note that mappings of plausible regions to nonsensical regions may occur under other algebraic structures such as reflections in hyperplanes. For example, if $F_{\theta_1,\theta_2} = F_{\theta_1,-\theta_2}$ for all data values, then nonidentifiability exists and manifests itself as a reflection in the line $\theta_2 = 0$. For permutation-type nonidentifiability, and more generally, for nonidentifiabilities where implausible parameters are induced from plausible parameters, we advocate a modification of the model. In Section 3, this is accomplished via constraints in the parameter space whereby any subregion of the constrained space has mapped regions outside of the constrained space.

When permutation-type nonidentifiabilities have been removed, it is possible that there exist remaining nonidentifiabilities that take the form of contours when the probability distribution of the data is viewed as a function of the parameters. In these circumstances, informative prior information is combined with the likelihood according to the Bayesian paradigm. Gustafson (2004) has exhibited realistic scenarios where a moderate amount of prior information has led to reasonable inferences in the presence of these types of nonidentifiabilities.

3. MULTINOMIAL MISCLASSIFICATION MODEL

We begin with the basic multinomial misclassification model and indicate various generalizations and modifications in Section 4. Consider N sampling units that are each classified into one of m categories. In a medical context, this may correspond to a group of patients who are diagnosed on an m-level scale. We denote the classification of the kth sampling unit using the random variable X_k, $k = 1, \ldots, N$ and define

$$q_i = \mathrm{P}(X_k = i)$$

for $i = 1, \ldots, m$, $k = 1, \ldots, N$. Since we allow for the possibility of misclassification, we denote the true (but unobserved) value of the kth sampling unit using the random variable T_k, where

$$p_i = \mathrm{P}(T_k = i)$$

for $i = 1, \ldots, m$, $k = 1, \ldots, N$. Note that the basic model (which can be extended) assumes that the patients are homogeneous in the sense that p_i does not depend on k. This leads to the definition of misclassification probabilities

$$\pi_{ji} = \mathrm{P}(X_k = i \,|\, T_k = j)$$

for $i, j = 1, \ldots, m$, $k = 1, \ldots, N$. The likelihood is then proportional to

$$L = L(p_1, \ldots, p_m, \pi_1, \ldots, \pi_m) = \prod_{k=1}^{m} q_k^{n_k} = \prod_{k=1}^{m} \Big(\sum_{\ell=1}^{m} p_\ell \pi_{\ell k} \Big)^{n_k}, \tag{1}$$

where $\pi_j = \big(\pi_{j1}, \ldots, \pi_{jm}\big)$ and n_i is the number of sampling units that are classified according to category i, $i = 1, \ldots, m$ such that $N = n_1 + \cdots + n_m$. Since

$$\sum_{k=1}^{m} p_k = 1 \quad \text{and} \quad \sum_{k=1}^{m} \pi_{jk} = 1$$

for $j = 1, \ldots, m$, the likelihood is a function of $(m-1) + m(m-1) = m^2 - 1$ parameters. In many applications, such as the evaluation of questionnaires, the true categorical probabilities $p_1, \ldots, p_m$ are the parameters of primary interest. However, one can also imagine applications such as the evaluation of new diagnostic tests where there is interest in the misclassification probabilities π_{ji}.

To address the identifiability issue, suppose that $i < j$ and let

$$L_{ij} = L\big(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_m, \pi_1, \ldots, \pi_j, \ldots, \pi_i, \ldots, \pi_m\big), \tag{2}$$

i.e., p_i and p_j have swapped positions, and π_i and π_j have swapped positions. Then it is clear that a nonidentifiability exists between (p_i, p_j, π_i, π_j) and (p_j, p_i, π_j, π_i) since $L_{ij} = L$ for all data values $n_1, \ldots, n_m$. Moreover, this is a permutation-type nonidentifiability, and more generally, it occurs for any of the $m!$ permutations of both sets of indices in (1).

To overcome the permutation-type nonidentifiabilities, we introduce four alternative sets of constraints on the parameters. The first three constraints follow a hierarchy of strongest, weaker and weakest, while the fourth is a different type of constraint. The strongest of the first three constraints may often be applicable in the case of ordinal categorical data and is given by

$$\pi_{j1} < \cdots < \pi_{j,j-1} < \pi_{jj} > \pi_{j,j+1} > \cdots > \pi_{jm} \tag{3}$$

for $j = 1, \ldots, m$. Constraint (3) says in effect that it becomes less likely to make a misclassification as the incorrect category moves away from the true category. For example, consider a

5-point scale of health where 0 denotes extremely poor health, 2 denotes average health and 4 denotes excellent health. If a subject truly has average health, then it is sensible that diagnoses of 2, 3 and 4 are in decreasing order of probability, and that diagnoses of 2, 1 and 0 are in decreasing order of probability.

To see that (3) breaks up the aforementioned nonidentifiabilities, L in (1) becomes

$$\begin{aligned} L &= \prod_{k=1}^{m}\Big(\sum_{\ell=1}^{m} p_\ell \pi_{\ell k}\Big)^{n_k} \cdot I\big(\pi_{j1} < \cdots < \pi_{j,j-1} < \pi_{jj} > \pi_{j,j+1} > \cdots > \pi_{jm}; \forall j\big) \\ &= \prod_{k=1}^{m}\Big(\sum_{\ell=1}^{m} p_\ell \pi_{\ell k}\Big)^{n_k} \cdot I\big(\pi_{ji} < \pi_{jj}\big) \cdot I\big(\pi_{ii} > \pi_{ij}\big) \cdot I^*, \end{aligned}$$

where $i < j$ and I^* generically denotes the remaining constraints. It follows that L_{ij} in (2) becomes

$$L_{ij} = \prod_{k=1}^{m}\Big(\sum_{\ell=1}^{m} p_\ell \pi_{\ell k}\Big)^{n_k} \cdot I\big(\pi_{ii} < \pi_{ij}\big) \cdot I\big(\pi_{ji} > \pi_{jj}\big) \cdot I^* \neq L$$

for any pair $i < j$. Therefore the constraint given by (3) has removed the invariance in the likelihood between (p_i, p_j, π_i, π_j) and (p_j, p_i, π_j, π_i) for all $i < j$ and has eliminated the permutation-type nonidentifiabilities.

Our second constraint is a weaker version of the previous constraint and is given by

$$\pi_{ji} < \pi_{jj} \tag{4}$$

for all $i \neq j$. The constraint (4) is appealing since it says that it is more probable to be classified correctly than incorrectly. This constraint is more likely to be applicable when the categorical data is nominal rather than ordinal. An argument similar to that above can be used to demonstrate that (4) eliminates the permutation-type nonidentifiabilities.

An even weaker constraint is given by

$$\pi_{ij} + \pi_{ji} < \pi_{ii} + \pi_{jj} \tag{5}$$

for all $i < j$. The motivation for (5) is less compelling than (3) or (4) but may be regarded as an average control on misclassification. Since $\pi_{ii} - \pi_{ij}$ and $\pi_{jj} - \pi_{ji}$ may both be thought of as differences in probabilities between making correct and incorrect classifications, the constraint (5) therefore stipulates that their average should be positive. Again, a similar argument to that above can be used to demonstrate that (5) eliminates the permutation-type nonidentifiabilities.

A fourth constraint which can also be shown to remove the permutation-type nonidentifiabilities takes the form

$$\pi_{ji} < \pi_{ii} \tag{6}$$

for all $j \neq i$. If we think of the matrix $\pi = (\pi_{ji})$, we observe that constraint (6) takes a different form than the previous constraints: constraint (6) imposes conditions on the columns of π, whereas constraints (3), (4) and (5) impose conditions on the rows of π. Now although (6) is not immediately intuitive, in the Appendix we prove that two conditions each imply constraint (6), which may be appealing in particular applications. In fact, the first condition is both necessary and sufficient. Specifically, fix the matrix (π_{ji}) such that $\pi_{1i} + \cdots + \pi_{mi} > 0$ for all $i = 1, \ldots, m$. Then the first condition $\mathrm{P}(T = i \,|\, X = i) > \mathrm{P}(T = i \,|\, X \neq i)$ holds for all i and all $p_1, \ldots, p_m$ with $0 < \mathrm{P}(X = i) < 1$, if and only if (6) is true. In words, the first condition states that it is more probable that the true classification is i when the diagnosis is i than when the diagnosis is something other than i. The second condition states that for all $j \neq i$, $D_i = p_1\pi_{1i} + \cdots + p_m\pi_{mi}$ is an increasing function of p_i when only p_i and $p_j > 0$ are allowed to vary. In words, D_i is

the probability of the diagnosis i, and the second condition states that as the occurrence of the true category i becomes more probable and the occurrence of the true category j becomes less probable, the more probable it is to have a diagnosis of i, $j \neq i$.

We note that both (5) and (6) generalize the constraint $\pi_{12} + \pi_{21} < 1$ proposed by Evans, Guttman, Haitovsky & Swartz (1996) in the case of binomial data (i.e., $m = 2$). However, (6) is the natural generalization since Evans, Guttman, Haitovsky & Swartz (1996) similarly motivated their constraint by requiring that $p_1(1 - \pi_{12}) + (1 - p_1)\pi_{21}$ be an increasing function of p_1.

There exist other constraints that can eliminate the permutation-type nonidentifiabilities. For example, $\pi_{ii} > 1/2$ for $i = 1, \ldots, m$ is such a constraint. However, we do not find these constraints as compelling as the four proposed constraints, and therefore, we will consider them no further.

To complete the specification of the Bayesian multinomial misclassification model, we specify prior distributions

$$\begin{aligned} (p_1, \ldots, p_m) &\sim \text{Dirichlet}(a_1, \ldots, a_m) \\ \text{and} \quad (\pi_{j1}, \ldots, \pi_{jm}) &\sim \text{Dirichlet}(b_{j1}, \ldots, b_{jm}) \quad \text{for } j = 1, \ldots, m \end{aligned}$$

where the a's and b's are positive hyperparameters set by the experimenter. Dirichlet distributions are ideal as they satisfy the need for the p's and π's to be defined on the simplex. We also note that together with the chosen constraint (3), (4), (5) or (6), the Dirichlet distributions for $(\pi_{j1}, \ldots, \pi_{jm})$ are essentially truncated-Dirichlet distributions.

Now having introduced a constraint to eliminate the permutation-type nonidentifiabilities, the Dirichlet hyperparameters need to be assigned. In many Bayesian applications, Dirichlet hyperparameters are set equal to 1.0 and it is suggested that they provide vague prior distributions. As argued in Section 2, we strongly recommend against this practice in the multinomial misclassification problem, particularly in the case of the misclassification hyperparameters b_{ji}. We say this since there exist remaining nonidentifiabilities that ought to be partially distinguished by the prior distributions. More specifically, we observe that there exists a nonidentifiability between the parameters $(p_1^{(1)}, \ldots, p_m^{(1)}, \pi_1^{(1)}, \ldots, \pi_m^{(1)})$ and $(p_1^{(2)}, \ldots, p_m^{(2)}, \pi_1^{(2)}, \ldots, \pi_m^{(2)})$ whenever

$$\sum_{\ell=1}^{m} p_\ell^{(1)} \pi_{\ell i}^{(1)} = \sum_{\ell=1}^{m} p_\ell^{(2)} \pi_{\ell i}^{(2)}$$

for $i = 1, \ldots, m$. This type of nonidentifiability defines "contours" in the likelihood that can be "bent" by nonflat prior distributions. In the multinomial misclassification problem, it is typically unreasonable to assign flat priors to $(\pi_{j1}, \ldots, \pi_{jm})$, $j = 1, \ldots, m$, for this would imply that disparate parameters would be equally plausible (a priori and a posteriori) under any of the four constraints. We argue for a robust Bayesian analysis where various priors are entertained and the resultant inferences are compared. This is carried out in the examples considered in this paper.

3.1. Example 1: simple model.

To illustrate our treatment of nonidentifiability in the multinomial misclassification model, we consider the simplest case corresponding to $m = 2$ categories. In the simple setting, we are better able to visualize concepts as we have a low dimensional model (i.e., three parameters) with parameter vector $(p_1, \pi_{11}, \pi_{21})$. However, the concepts discussed in this example also extend to the problem for general m.

In Figure 1, we provide a plot of the parameters that yield the particular likelihood

$$\begin{aligned} L &= \{p_1\pi_{11} + (1 - p_1)\pi_{21}\}^{n_1} \{p_1(1 - \pi_{11}) + (1 - p_1)(1 - \pi_{21})\}^{n_2} \\ &= (0.60)^{n_1}(0.40)^{n_2}. \end{aligned} \tag{7}$$

Note that the likelihood is the same for all of the plotted parameters regardless of the data n_1, n_2. In other words, the displayed parameters are nonidentifiable. To better visualize the plot in Figure 1, keep in mind that for fixed $p_1 \in [0,1]$, the profile (π_{11}, π_{21}) is always a straight line. Therefore, the plot consists of a twisted sheet. At the top of the twisted sheet (i.e., $p_1 = 1.0$), the line $\pi_{11} = 0.6$ occurs in the (π_{11}, π_{21}) plane. The twisted sheet extends down to the middle of the plot (i.e., $p_1 = 0.5$) where the line $\pi_{11} + \pi_{21} = 1.2$ occurs in the (π_{11}, π_{21}) plane. At the bottom of the twisted sheet (i.e., $p_1 = 0.0$), the line $\pi_{21} = 0.6$ occurs in the (π_{11}, π_{21}) plane. Note that we could have chosen other likelihoods—for example, $L = (0.15)^{n_1}(0.85)^{n_2}$—that result in different plots but share the same structure of a twisted sheet.

In Figure 1, for every point $(p_1, \pi_{11}, \pi_{21})$ in the parameter space, there is a corresponding point $(1 - p_1, \pi_{21}, \pi_{11})$ with equal likelihood. These pairs of points reflect the permutation-type nonidentifiability and are the source of the symmetry seen in the plot. The symmetry has deleterious effects as can be seen by assigning a flat prior. For example, the posterior density $f(p_1) = f(p_1 \mid n_1, n_2)$ is given by

$$\begin{aligned} f(p_1) &\propto \int_0^1 \int_0^1 \sum_{i=0}^{n_1} \binom{n_1}{i} (p_1\pi_{11})^i \{(1-p_1)\pi_{21}\}^{n_1-i} \\ &\quad \cdot \sum_{j=0}^{n_2} \binom{n_2}{j} \{p_1(1-\pi_{11})\}^j \{(1-p_1)(1-\pi_{21})\}^{n_2-j} \, d\pi_{11}\, d\pi_{21} \\ &= \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} \binom{n_1}{i} \binom{n_2}{j} \frac{i!j!}{(i+j+1)!} \frac{(n_1-i)!(n_2-j)!}{(n_1+n_2-i-j+1)!} p_1^{i+j} (1-p_1)^{n_1+n_2-i-j}. \end{aligned}$$

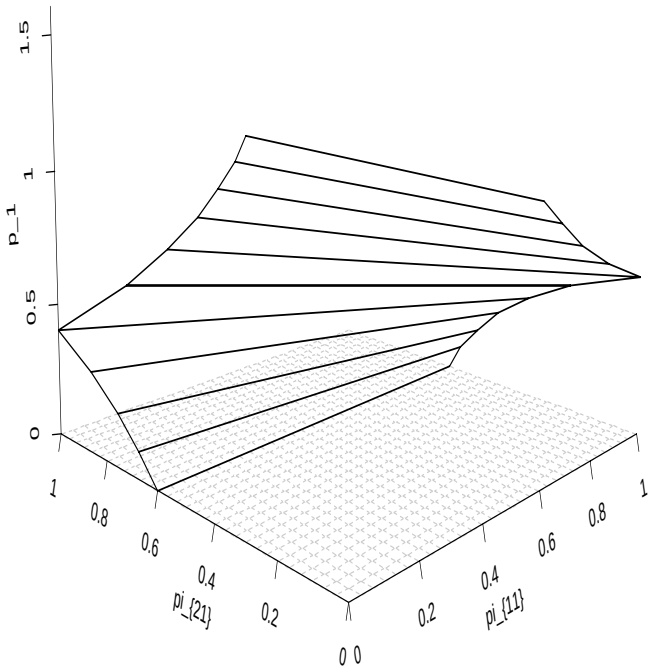

FIGURE 1: Plot of the nonidentifiable parameters corresponding to the likelihood $L = (0.60)^{n_1}(0.40)^{n_2}$ in Example 1.

By changing variables, it can be shown that $f(1/2 + \delta) = f(1/2 - \delta)$ for all $0 < \delta < 1/2$. In turn, this implies $\mathrm{E}(p_1 \mid n_1, n_2) = 1/2$ for all data values n_1, n_2. As the prior expectation $\mathrm{E}(p_1) = 1/2$, it might be said that little has been learned about the true categorical probability p_1.

To deal with the permutation-type nonidentifiabilities, we introduce constraints motivated by the nature of the problem. Constraints (3) and (4) both reduce to $\pi_{11} > 1/2$ and $\pi_{21} < 1/2$. Constraints (5) and (6) both reduce to $\pi_{11} > \pi_{21}$. In Figure 2, we provide a plot of the parameters that yield the likelihood (7) and are subject to the constraints $\pi_{11} > 1/2$ and $\pi_{21} < 1/2$. It is readily seen that the symmetry disappears. In addition, the region of constant likelihood is a smaller, more plausible set than in Figure 1. Together, with weak prior information on $(p_1, \pi_{11}, \pi_{21})$, inference can then proceed.

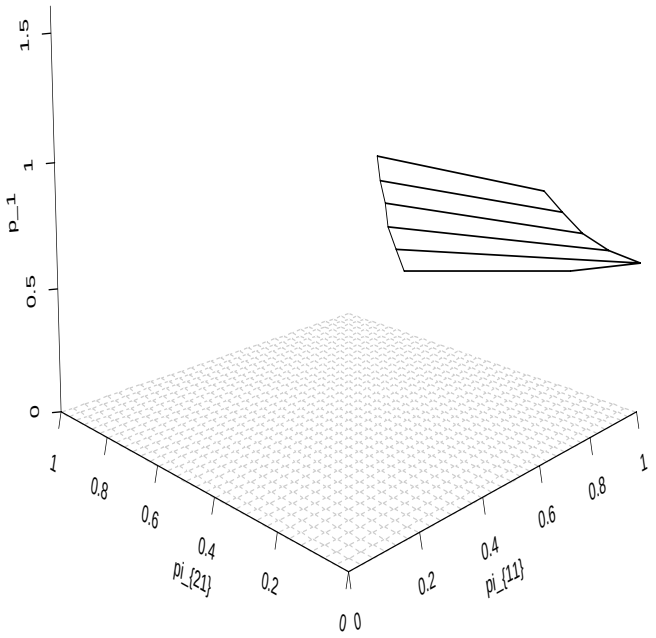

FIGURE 2: Plot of the nonidentifiable parameters corresponding to the likelihood $L = (0.60)^{n_1}(0.40)^{n_2}$ and subject to constraint (3) (i.e. $\pi_{11} > 0.5$ and $\pi_{21} < 0.5$) in Example 1.

To summarize, permutation-type nonidentifiabilities exist in the multinomial misclassification model and can lead to distorted posterior inferences if they are not properly handled. Fortunately, the structure of the problem typically suggests constraints (either (3), (4), (5) or (6)) to eliminate the permutation-type nonidentifiabilities. Remaining nonidentifiabilities in the multinomial misclassification may be handled through informative Dirichlet distributions. The effect of the Dirichlet distributions on posterior inferences is demonstrated in the examples which follow.

3.2. Dealing with a poor assessor.

Now there may be rare situations where an assessor (or a diagnostic test) is so poor that incorrect assessments are made more often than correct assessments. Suppose for example, that an assessor has severe difficulty distinguishing between categories i_1 and i_2 to the extent that $\pi_{i_1 i_2} > \pi_{i_1 i_1}$ and $\pi_{i_2 i_1} > \pi_{i_2 i_2}$. If this were really true, it would be illogical, for example, to invoke constraint (4) to overcome the permutation-type nonidentifiabilities. The question then is what to do in a problem like this?

One possibility is to throw away all of the diagnoses made by this assessor. However, there may be a solution which is not wasteful of data. In the example given above, suppose that you believe that all of the constraints in (4) are satisfied except for $\pi_{i_1 i_2} > \pi_{i_1 i_1}$ and $\pi_{i_2 i_1} > \pi_{i_2 i_2}$. For example, the condition may be detected when the diagnostic test is compared against

a prohibitively expensive gold standard. Then one should simply change all of the assessor diagnoses of i_1 to i_2, and vice versa. This has the effect of exchanging the i_1th column with the i_2th column in the matrix $\pi = (\pi_{ji})$. Consequently, constraint (4) is now satisfied and can be utilized to remove the permutation-type nonidentifiabilities. More generally, switching diagnoses according to some permutation $i_1 \to i_2 \to \cdots \to i_k$ has the effect of permuting the corresponding columns of the matrix π. If it is reasonable to assign one of the four proposed constraints to the permuted matrix, then the permutation-type nonidentifiabilities are effectively handled. Of course, all of this presumes that one has ample prior information on misclassification to make judgements on whether given constraints on the π_{ji} are reasonable.

To see that the strategy has some flexibility, consider the following variation of the above example. Suppose that you believe that all of the constraints in (6) are satisfied except for $\pi_{i_1 i_2} > \pi_{i_2 i_2}$ and $\pi_{i_2 i_1} > \pi_{i_1 i_1}$. Then one should simply change all of the assessor diagnoses of i_1 to i_2, and vice versa. This has the effect of exchanging the i_1th column with the i_2th column in the matrix $\pi = (\pi_{ji})$. Consequently, constraint (6) is now satisfied and can be utilized to remove the permutation-type nonidentifiabilities.

3.3. Computations.

Although full posterior distributions contain all of the information regarding unknown parameters, the posterior in this application is not in a form that is readily interpretable. Our posterior density is a function of dimension $m^2 - 1$, and is proportional to the product of (1), the prior density and the indicator function corresponding to the chosen constraint. Instead of focusing on the full posterior, we consider posterior summaries of some of the parameters. For example, we may be interested in the posterior means and posterior standard deviations of the p_i's and the π_{ji}'s. As these summaries involve intractable $(m^2 - 1)$-dimensional integrations, we take a sampling based approach using the theory of Markov chains. Data augmentation in conjunction with the Gibbs sampler (Gilks, Richardson & Spiegelhalter 1996) is ideal for this problem as it is convenient to sample from the requisite full conditional distributions. Recall that the augmented latent variable $T_k = j$ implies that the kth sampling unit truly belongs to category j, and let $[A \,|\, \cdot]$ denote the conditional density or probability mass function of A given the data and all other parameters. An implementation of the Gibbs sampling algorithm with the latent variable T requires that we sample iteratively according to the following distributions

$$\begin{aligned}
[T_k = j \,|\, \cdot] &= p_j \pi_{j x_k} \Big/ \sum_{\ell=1}^{m} p_\ell \pi_{\ell x_k} \quad \text{for } k = 1, \ldots, N, j = 1, \ldots, m \\
[p_1, \ldots, p_m \,|\, \cdot] &\sim \text{Dirichlet}\Big(a_1 + \sum_{k=1}^{N} I(T_k = 1), \ldots, a_m + \sum_{k=1}^{N} I(T_k = m) \Big) \\
[\pi_{j1}, \ldots, \pi_{jm} \,|\, \cdot] &\sim \text{truncated-Dirichlet}(c_{j1}, \ldots, c_{jm}), \\
&\qquad c_{ji} = b_{ji} + \sum_{k=1}^{N} I(T_k = j) I(X_k = i) \quad \text{for } j = 1, \ldots, m
\end{aligned} \tag{8}$$

where the truncated-Dirichlet distributions are truncated according to the chosen constraint (3), (4), (5) or (6).

Fortunately, sampling from each of the distributions in (8) is straightforward as software is readily available to generate from Dirichlet distributions and from finite discrete distributions. In the case of the truncated-Dirichlet distributions, we simply use the rejection sampling algorithm where we sample from the corresponding parent Dirichlet and only accept samples that satisfy the given constraint.

Occasionally, one may find that the rejection sampling approach is too inefficient to be used in practice. In other words, the algorithm is too slow as a high percentage of the candidate variates are rejected. This may occur when the number of categories m is large and/or

when the constraints are severe. In this case, we consider a more efficient approach to generating from truncated-Dirichlet distributions which may be regarded as a Gibbs sampling algorithm imbedded within a Gibbs sampling algorithm. In general, consider $[\pi_{j1}, \ldots, \pi_{jm} \,|\, \cdot] \sim$ truncated-Dirichlet$(c_{j1}, \ldots, c_{jm})$, where the truncation is specified according to any of the four proposed constraints. It follows that the full conditional distribution of π_{ji} is truncated on some interval (q_{ji1}, q_{ji2}), where $0 \leq q_{ji1} < q_{ji2} \leq 1$ for $i = 1, \ldots, m-1$. To facilitate simulation, some elementary distribution theory gives

$$[y_{ji} \,|\, \pi_{j1}, \ldots, \pi_{j,i-1}, \pi_{j,i+1}, \ldots, \pi_{j,m-1}] \sim \text{truncated-Beta}(c_{ji}, c_{jm})$$

where $y_{ji} = \pi_{ji}/\pi_{j(i)}$, $\pi_{j(i)} = 1 - \pi_{j1} - \cdots - \pi_{j,i-1} - \pi_{j,i+1} - \cdots - \pi_{j,m-1}$ and y_{ji} is truncated on the interval $(q_{ji1}/\pi_{j(i)}, q_{ji2}/\pi_{j(i)})$ Therefore, the recipe is clear; we proceed through the indices $i = 1, \ldots, m-1$, where we generate y_{ji} from its corresponding truncated-Beta distribution and set $\pi_{ji} = y_{ji}\pi_{j(i)}$.

3.4. Example 1 continued: computations in the simple model.

We consider again the simplest case corresponding to $m = 2$ categories. Suppose that we have data $n_1 = 8$, $n_2 = 2$ and that the Dirichlet priors are specified by $a_1 = a_2 = 1.0$, $b_{11} = b_{22} = 2.1$, and $b_{12} = b_{21} = 0.9$. These are not very informative priors since $\text{var}(p_i) = (0.29)^2$ and $\text{var}(\pi_{ii}) = (0.23)^2$, $i = 1, 2$. We are interested in the effect of the various constraints, and consider the cases of no constraint, constraint (5) $\equiv$ constraint (6) and constraint (3) $\equiv$ constraint (4). Posterior means and posterior standard deviations of the parameters are given in Table 1.

TABLE 1: Posterior means and posterior standard deviations (in parentheses) of the parameters in Example 1 continued.

Constraint	p_1	π_{11}	π_{21}
No constraint	0.64 (0.26)	0.82 (0.15)	0.42 (0.25)
Weakest : $(5) \equiv (6)$	0.67 (0.24)	0.85 (0.11)	0.37 (0.23)
Strongest: $(3) \equiv (4)$	0.74 (0.19)	0.85 (0.11)	0.24 (0.15)

We observe meaningful differences in the estimates due to the different constraints, and as expected, we observe tighter posterior distributions corresponding to the stronger constraints. Additional runs (not shown) indicate that differences between the three cases diminish as the priors become more informative. We are also interested in the practical question concerning the convergence of the Markov chains used to produce Table 1. Although the estimates in Table 1 are reliable, the lag-5 autocorrelations in p_1 are 0.41, 0.39, and 0.21 for the three cases, respectively. Similarly, the lag-10 autocorrelations in p_1 are 0.18, 0.17, and 0.03. This highlights that even in the simplest case (dimension 3), the regions of nonidentifiability cause poor mixing of the chains. It also suggests that more serious difficulties may arise in higher dimensions (i.e., larger m) where there are vast regions of nearly flat posterior density. Although we have not shown any results, the convergence issues become less problematic as the Dirichlet priors become more informative for this effectively reduces the size of the regions where the chains wander. One simple (although wasteful) method of dealing with correlated output from a Markov chain would be to collect only every qth variate, with the spacing q chosen sufficiently large to dampen the correlation. Batching (Ripley 1987) is another simple method for handling correlated output.

4. MORE EXAMPLES

4.1. Example 2: artificial data.

We illustrate the Bayesian multinomial misclassification model using an artificial data set for which the true categorical probabilities and the misclassification probabilities are known.

We consider a problem with $N = 1000$ sampling units and $m = 6$ categories where the true categorical probabilities $p_i^{(0)}$ and the misclassification probabilities $\pi_{ji}^{(0)}$ are specified in Table 2. Briefly, we have a situation where there is a great deal of misclassification; in every category, there is only a probability of 0.5 that the classification is correct. Also, the data correspond to a situation involving ordinal responses where the misclassification probabilities decrease as we move away from the true categories. We refer to the data as artificial since the categorical responses $X_1, \ldots, X_N$ have been constructed so that their proportions agree with the probabilities in Table 2.

TABLE 2: The true categorical probabilities $p_i^{(0)}$ and the misclassification probabilities $\pi_{ji}^{(0)}$ in Example 2.

		$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$
$p_i^{(0)}$		0.20	0.16	0.16	0.16	0.16	0.16
$\pi_{ji}^{(0)}$	$j=1$	0.50	0.20	0.15	0.08	0.05	0.02
	$j=2$	0.15	0.50	0.15	0.10	0.08	0.02
	$j=3$	0.10	0.15	0.50	0.15	0.08	0.02
	$j=4$	0.02	0.08	0.15	0.50	0.15	0.10
	$j=5$	0.02	0.08	0.10	0.15	0.50	0.15
	$j=6$	0.02	0.05	0.08	0.15	0.20	0.50

We focus on the posterior means of the true categorical probabilities p_i under various constraints and priors, and these are reported in Table 3.

TABLE 3: Posterior means and posterior standard deviations (in parentheses) of the true categorical probabilities p_i under various constraints and priors as discussed in Example 2.

	p_1	p_2	p_3	p_4	p_5	p_6
Case A	0.15 (0.10)	0.18 (0.11)	0.20 (0.12)	0.19 (0.12)	0.17 (0.11)	0.12 (0.09)
Case B	0.19 (0.09)	0.17 (0.10)	0.16 (0.10)	0.16 (0.10)	0.16 (0.09)	0.16 (0.08)
Case C	0.20 (0.07)	0.16 (0.08)	0.16 (0.08)	0.16 (0.08)	0.16 (0.08)	0.16 (0.06)
Case D	0.18 (0.04)	0.17 (0.04)	0.18 (0.04)	0.18 (0.05)	0.17 (0.04)	0.14 (0.04)
Case E	0.18 (0.04)	0.17 (0.04)	0.18 (0.05)	0.18 (0.05)	0.17 (0.04)	0.14 (0.04)
Case F	0.21 (0.04)	0.16 (0.04)	0.16 (0.04)	0.16 (0.04)	0.16 (0.04)	0.16 (0.04)
Case G	0.18 (0.04)	0.17 (0.04)	0.18 (0.04)	0.18 (0.05)	0.17 (0.04)	0.14 (0.04)
Case H	0.19 (0.02)	0.17 (0.02)	0.17 (0.02)	0.17 (0.02)	0.16 (0.02)	0.15 (0.01)
$p_i^{(0)}$	0.20	0.16	0.16	0.16	0.16	0.16

We begin with Case A which uses constraint (4) and is flat in both the p_i and the π_{ji} (i.e., $a_i = b_{ji} = 1.0$ for $i, j = 1, \ldots, 6$). This is a prior which we would never recommend in practice due to the identifiability concerns. We observe that the inferences are poor. For example, an interval of

five posterior standard deviations about the posterior mean of p_1 does not capture the true $p_1^{(0)}$. We also remark that estimates from the Gibbs sampler require many more iterations ($\sim 10^6$) using this prior than when using more informative priors. The slow convergence of the Gibbs sampler is due to the high correlations in the parameters resulting from the nonidentifiabilities. For example, the lag-50 autocorrelation in p_1 is 0.45.

Cases B and C both use constraint (4). We set a flat prior $a_i = 1$, $i = 1, \ldots, 6$ for the true categorical probabilities and set the misclassification prior with hyperparameters $b_{ji} = K\pi_{ji}^{(0)}$ for $i, j = 1, \ldots, 6$. Cases B and C use good misclassification priors in the sense that the prior means equal the underlying misclassification errors. The larger the value of K, the better our knowledge of the misclassification errors. For Case B, we set $K = 10.0$, and for Case C, we set $K = 40.0$. We observe that Case B does a good job in estimating the p_i with only a little difficulty involving p_1 and p_2. It is encouraging that good estimation occurs in Case B, where the prior is not that informative. As expected, the inference is improved under the more informative prior in Case C.

To obtain a sense of the impact of the four constraints (3), (4), (5) and (6), we first note that the data were constructed so as to satisfy the strongest constraint (3). We consider a situation in which we have good prior sense of the true categorical probabilities p_i and the overall misclassification probabilities $\sum_{k \neq j} \pi_{jk}$, but we are a priori unsure as to how the misclassification probabilities are distributed amongst the $m - 1 = 5$ cells. That is, we choose hyperparameters $a_1 = 10.0$, $a_2 = \cdots = a_6 = 8.0$, $b_{ii} = 12.5$ and $b_{ji} = 2.5$ for $i, j = 1, \ldots, 6$, $i \neq j$. Note that this choice gives prior expectations $\mathrm{E}(p_i) = p_i^{(0)}$ and $\mathrm{E}(\pi_{ii}) = \pi_{ii}^{(0)}$ for $i = 1, \ldots, 6$. We anticipate improved inferences as we go from constraint (5) (Case D) to constraint (4) (Case E) to constraint (3) (Case F). In fact, this is observed although there are negligible differences (i.e., third digit) in the effects between Case D and Case E. As expected, there are smaller posterior standard deviations associated with Case F than with Cases D and E since Case F incorporates the strongest of the constraints. We also observe that Case G which uses constraint (6) gives the same results as Cases D and E. When compared to Example 1, the implication here is that the constraints become less important relative to the prior as the dimension of the problem increases. Finally, when we multiply the a_i's and the b_{ji}'s by a factor of 10.0, we have more informative priors and the results are the same under all four constraints. These results are reported under Case H.

4.2. Example 3: Dawid and Skene data.

There are many generalizations of the basic multinomial misclassification model presented in this paper. For example, in the medical context, there could be multiple assessments of patients, diagnoses by multiple physicians, and patients arising from different cohorts. It is also possible that the true and the observed number of categories may not be the same value m.

We now consider one such generalization to give the reader a sense of how these problems can be approached. The data are taken from Dawid & Skene (1979); they are based upon a standard form that was completed on 45 patients, which contains information reflecting the state of health of each patient. The forms were then subsequently reviewed independently by five anaesthetists who classified each patient on a scale from 1 to 4 to assess whether the patient was sufficiently fit to undergo a general anaesthetic. The first anaesthetist provided three assessments, each assessment separated by some weeks. The data are recorded in Table 4.

We make the assumption that each patient belongs to the same cohort giving rise to probabilities p_j that the true classification of a random patient is category $j = 1, \ldots, 4$. Defining $n_{\ell i}^{(k)}$ as the number of times physician $\ell = 1, \ldots, 5$ classifies patient $k = 1, \ldots, 45$ to category $i = 1, \ldots, 4$ and assuming unique misclassification probabilities π_{lji} for each anaesthetist, we

obtain the likelihood

$$L = \prod_{k=1}^{45} \prod_{\ell=1}^{5} \prod_{i=1}^{4} \Big(\sum_{j=1}^{4} p_j \pi_{\ell j i} \Big)^{n_{\ell i}^{(k)}} . \tag{9}$$

TABLE 4: Assessments of fitness for anaesthesia taken from Dawid & Skene (1979) as discussed in Example 3.

	Anaesthetist						Anaesthetist						Anaesthetist				
Patient	1	2	3	4	5	Patient	1	2	3	4	5	Patient	1	2	3	4	5
1	111	1	1	1	1	16	111	2	1	1	1	31	111	1	1	1	1
2	333	4	3	3	4	17	111	1	1	1	1	32	333	3	2	3	3
3	112	2	1	2	2	18	111	1	1	1	1	33	111	1	1	1	1
4	222	3	1	2	1	19	222	2	2	2	1	34	222	2	2	2	2
5	222	3	2	2	2	20	222	1	3	2	2	35	222	3	2	3	2
6	222	3	3	2	2	21	222	2	2	2	2	36	433	4	3	4	3
7	122	2	1	1	1	22	222	2	2	2	1	37	221	2	2	3	2
8	333	3	4	3	3	23	222	3	2	2	2	38	232	3	2	3	3
9	222	2	2	2	3	24	221	2	2	2	2	39	333	3	4	3	2
10	232	2	2	2	3	25	111	1	1	1	1	40	111	1	1	1	1
11	444	4	4	4	4	26	111	1	1	1	1	41	111	1	1	1	1
12	222	3	3	4	3	27	232	2	2	2	2	42	121	2	1	1	1
13	111	1	1	1	1	28	111	1	1	1	1	43	232	2	2	2	2
14	222	3	2	1	2	29	111	1	1	1	1	44	121	1	1	1	1
15	121	1	1	1	1	30	112	1	1	2	1	45	222	2	2	2	2

We further define the latent variable $T_k = j$ if patient k truly belongs to category j and assign independent prior distributions $(p_1, p_2, p_3, p_4) \sim \text{Dirichlet}(a_1, a_2, a_3, a_4)$ and $(\pi_{\ell j 1}, \pi_{\ell j 2}, \pi_{\ell j 3}, \pi_{\ell j 4}) \sim \text{Dirichlet}(b_{\ell j 1}, b_{\ell j 2}, b_{\ell j 3}, b_{\ell j 4})$ for $\ell = 1, \ldots, 5$ and $j = 1, \ldots, 4$. We introduce the constraint $\pi_{\ell j i} < \pi_{\ell j j}$ for $\ell = 1, \ldots, 5$ and $i \neq j$. The constraint is a modification of (4) and is motivated by the permutation-type nonidentifiability obtained through permutations of the index j in (9). It then follows that an implementation of the Gibbs sampling algorithm is based upon iterative sampling from the following distributions

$$\begin{aligned}
[T_k = j \,|\, \cdot\,] &= p_j \prod_{\ell=1}^{5} \prod_{i=1}^{4} \pi_{\ell j i}^{n_{\ell i}^{(k)}} \Big/ \sum_{j=1}^{4} p_j \prod_{\ell=1}^{5} \prod_{i=1}^{4} \pi_{\ell j i}^{n_{\ell i}^{(k)}}, \\
[p_1, p_2, p_3, p_4 \,|\, \cdot\,] &\sim \text{Dirichlet}\Big(a_1 + \sum_{k=1}^{45} I(T_k = 1), \ldots, a_4 + \sum_{k=1}^{45} I(T_k = 4) \Big), \\
[\pi_{\ell j 1}, \pi_{\ell j 2}, \pi_{\ell j 3}, \pi_{\ell j 4} \,|\, \cdot\,] &\sim \text{truncated-Dirichlet}(c_{\ell j 1}, c_{\ell j 2}, c_{\ell j 3}, c_{\ell j 4}), \\
& \qquad c_{\ell j i} = b_{\ell j i} + \sum_{k=1}^{45} I(T_k = j) n_{\ell i}^{(k)}
\end{aligned}$$

where the truncated-Dirichlet distributions are truncated according to $\pi_{\ell j i} < \pi_{\ell j j}$, $\ell = 1, \ldots, 5$ and $i \neq j$. This is a problem where prior information is vital since the ratio of observations to parameters 315/63 is small. It is also instructive to note from the above algorithm that replicate information (i.e., multiple assessments by the same rater on the same individual) is helpful. In

the case of replicate information, the counts $n_{\ell i}^{(k)}$ are no longer only 0's and 1's. This leads to larger parameters in the truncated-Dirichlet distributions, narrower simulated output of the $\pi_{\ell j i}$, and in turn, narrower inferences.

Our primary interest in this problem is the determination of the true category for each patient (i.e., we are interested in the posterior probability of $T_k = j$ for $j = 1, \ldots, 4$ and $k = 1, \ldots, 45$).

TABLE 5: Estimates of probabilities of the true classifications T_k for selected patients based on the Dawid & Skene (1979) data as discussed in Example 3.

		Category			
Approach	Patient	1	2	3	4
Maximum likelihood	1	1.00	0.00	0.00	0.00
with modified constraint (4)	3	0.00	1.00	0.00	0.00
	7	0.99	0.01	0.00	0.00
	36	0.00	0.00	0.00	1.00
Case 1: Prior A	1	1.00	0.00	0.00	0.00
with modified constraint (4)	3	0.22	0.78	0.00	0.00
	7	0.97	0.03	0.00	0.00
	36	0.00	0.00	0.96	0.04
Case 2: Prior A′	1	1.00	0.00	0.00	0.00
with modified constraint (4)	3	0.11	0.89	0.00	0.00
	7	0.94	0.06	0.00	0.00
	36	0.00	0.00	0.98	0.02
Case 3: Prior A′	1	1.00	0.00	0.00	0.00
with modified constraint (3)	3	0.20	0.80	0.00	0.00
	7	0.96	0.04	0.00	0.00
	36	0.00	0.00	0.94	0.06
Case 4: Prior B	1	1.00	0.00	0.00	0.00
with modified constraint (3)	3	0.52	0.48	0.00	0.00
	7	0.97	0.03	0.00	0.00
	36	0.00	0.00	0.60	0.40
Case 5: Prior B′	1	1.00	0.00	0.00	0.00
with modified constraint (3)	3	0.57	0.43	0.00	0.00
	7	0.92	0.08	0.00	0.00
	36	0.00	0.00	0.78	0.22
Case 6: Prior B′	1	1.00	0.00	0.00	0.00
with modified constraint (4)	3	0.57	0.43	0.00	0.00
	7	0.92	0.08	0.00	0.00
	36	0.00	0.00	0.78	0.22

In Table 5, we present estimates of these probabilities for selected patients using various approaches. We include the maximum likelihood results of Dawid & Skene (1979) based on the same constraint where computations are carried out using the EM algorithm. We observe that the maximum likelihood estimates are unrealistically accurate (i.e., mostly 0 and 1). Dawid &

Skene (1979) also comment on the failings of the EM algorithm in this application as convergence is guaranteed for only local maxima.

Case 1 in Table 5 is based on an empirical Bayes procedure using a weakly informative prior. We set $a_1 = a_2 = 4.3$, $a_3 = 1.0$ and $a_4 = 0.4$ to roughly mimic the proportion of diagnoses in the dataset. The hyperparameters $b_{\ell ji}$ are set by observing in Table 4 that there are a minimum of $X = 52$ misclassifications over the $45(7)$ assessments. We then assign $b_{\ell ji} = 10X/[(45)(7)(3)]$ for $i \neq j$ and $b_{\ell jj} = 10[(45)(7) - X]/[(45)(7)]$ so that the prior means for the $\pi_{\ell ji}$ equal the average minimum misclassification rate and the prior standard deviation for $\pi_{\ell jj}$ is 0.11. We refer to the prior as Prior A. We observe that the estimate for patient 36 (and all of the subsequent Bayesian estimates for patient 36) are in huge disagreement with the maximum likelihood estimate of Dawid & Skene (1979). Again, Dawid & Skene (1979) suggest convergence problems with the maximum likelihood approach, and we observe that the majority of the diagnoses (4 out of 7) support the direction of the Bayesian estimates for patient 36.

Case 2 is the same as Case 1 except that the $b_{\ell ji}$ are multiplied by a factor of 10.0 leading to a more informative prior which is referred to as Prior A$'$. The results using the more informative prior are not too dissimilar from those obtained using the weakly informative prior. This demonstrates that even weak prior information can sometimes overcome the difficulties associated with nonidentifiability.

We now consider the effect of using the constraint

$$\begin{aligned}
\pi_{\ell 11} &> \pi_{\ell 12} > \pi_{\ell 13} > \pi_{\ell 14} \\
\pi_{\ell 21} &< \pi_{\ell 22} > \pi_{\ell 23} > \pi_{\ell 24} \\
\pi_{\ell 31} &< \pi_{\ell 32} < \pi_{\ell 33} > \pi_{\ell 34} \\
\pi_{\ell 41} &< \pi_{\ell 42} < \pi_{\ell 43} < \pi_{\ell 44}
\end{aligned}$$

for $\ell = 1, \ldots, 5$. The above constraint is a modification of the strongest constraint (3) and may be reasonable as the data are ordinal. For Case 3, we use the modified constraint (3) with Prior A$'$. We observe the results for patient 3 are a little different when compared to Case 2 and this indicates that the constraints are meaningful when used in conjunction with the particular prior (Prior A$'$).

For Case 4, we use the modified constraint (3) with a different prior which we refer to as Prior B. Prior B continues to use $a_1 = a_2 = 4.0$, $a_3 = 1.0$ and $a_4 = 0.4$ although the misclassification hyperparameters are now set according to $b_{\ell ji} = (3.0)(0.3)^{|j-i|}$. The misclassification hyperparameters in Prior B are in keeping with the modified constraint (3), where probabilities of misclassification dampen out as we move away from the true category. The prior standard deviation of $\pi_{\ell jj}$ is 0.08, indicating that the prior is not too strongly informative. Case 4 differs considerably (patients 3 and 36) from Cases 1, 2 and 3, indicating that the choice of prior has a major impact on inferences.

Case 5 is the same as Case 4 except that the $b_{\ell ji}$ are multiplied by a factor of 10.0, leading to a more informative prior which we refer to as Prior B$'$. The similarity between the results of Case 5 and Case 4 again suggests that even weak prior information can sometimes overcome the difficulties associated with nonidentifiability.

In Case 6, we retain Prior B$'$ and go back to the modified constraint (4). We observe that the results for Case 6 are the same as for Case 5. This is not surprising due to the synergy between Prior B' and the modified constraint (3). Since the modified constraint (3) adds very little to what is already accomplished by Prior B$'$, replacing the modified constraint (3) with a weaker constraint results in little change to the posterior.

5. DISCUSSION

In the absence of double sampling, it has been historically difficult to obtain reliable inferences in the multinomial misclassification problem. This paper outlines a Bayesian approach where the first task of the experimenter is to eliminate permutation-type nonidentifiabilities. It is argued that the structure of the problem typically suggests appropriate constraints that can be used in this task.

In the second stage, the experimenter is faced with remaining nonidentifiabilities. The goal of the experimenter is to provide some information on the parameters in the form of Dirichlet distributions. This is typically not too difficult as one can translate guesses of the means and standard deviations of the parameters to the specification of the Dirichlet hyperparameters. One may also consider empirical Bayes methods for specifying the hyperparameters as discussed in Section 4.2.

Now there are caveats to the tasks described above. As the dimension of the problem increases (i.e., larger m), the regions of nearly flat likelihood become vast, and consequently Markov chains have difficulty exploring the posterior space if the Dirichlet prior distributions are not sufficiently informative. However, as the priors become more informative, this lessens the impact of the constraints. Nevertheless, as a principle, we still advocate the introduction of constraints as it is typically a straightforward task and the constraints eliminate illogical regions of the parameter space.

The question arises as to how informative the Dirichlet priors need to be. From a practical point of view, if the Gibbs sampling algorithm converges, the inferences are correct. However, inferences also need to be sensible, and to be sensible, the Dirichlet priors should assign little probability to regions that are unacceptable to you. Essentially, the priors should do what they are intended to do—express your prior opinions. Again, this is important because the data cannot distinguish between acceptable regions and unacceptable regions that are linked by nonidentifiabilities.

Finally, experimenters ought to exercise caution with the prior distributions by investigating the robustness of inferences to changes in the prior. It is impossible to get something for nothing; and in this problem involving nonidentifiability, it is the addition of prior information which drives the inference.

6. APPENDIX

We show that two readily interpretable conditions each imply constraint (6). The first condition is also a necessary condition.

PROPOSITION 1. *Fix a matrix (π_{ji}) of misclassification probabilities such that $\pi_{\ell 1}+\cdots+\pi_{\ell m} > 0$ for all $i = 1, \ldots, m$. The condition*

$$\mathrm{P}(T = i \,|\, X = i) > \mathrm{P}(T = i \,|\, X \neq i)$$

holds for all choices of i and all choices of $p_1, \ldots, p_m$ with $0 < \mathrm{P}(X = i) < 1$ if and only if $\pi_{ji} < \pi_{ii}$ for $j \neq i$.

Proof. We have

$$\mathrm{P}(T = i \,|\, X = i) = \frac{\mathrm{Prob}(X = i \,|\, T = i)\mathrm{P}(T = i)}{\mathrm{P}(X = i)} = \frac{p_i \pi_{ii}}{\sum_{\ell=1}^{m} p_\ell \pi_{\ell i}}$$

and similarly,

$$\mathrm{P}(T = i \,|\, X \neq i) = \frac{p_i(1 - \pi_{ii})}{1 - \sum_{\ell=1}^{m} p_\ell \pi_{\ell i}}.$$

Under the stated condition, it follows that

$$\frac{\pi_{ii}}{\sum_{\ell=1}^{m} p_\ell \pi_{\ell i}} > \frac{1-\pi_{ii}}{1-\sum_{\ell=1}^{m} p_\ell \pi_{\ell i}} \tag{10}$$

which implies $\pi_{ii} > \sum_{\ell=1}^{m} p_\ell \pi_{\ell i}$. By setting $p_j = 1$, $j \neq i$, we then have $\pi_{ii} > \pi_{ji}$, $j \neq i$.

For the necessity part of the proof, suppose that $\pi_{ii} > \pi_{ji}$, $j \neq i$. This implies $\pi_{ii} > \sum_{l=1}^{m} p_l \pi_{li}$ which in turn gives the stated condition (10).

PROPOSITION 2. *Define $p_{(ij)}$ as the $(m-2)$-dimensional vector $(p_1, \ldots, p_m)'$ with p_i and p_j removed. If for all $j \neq i$, $D_i = p_1\pi_{1i} + \cdots + p_m\pi_{mi}$ is an increasing function of p_i when $p_{(ij)}$ is fixed and $p_j > 0$, then $\pi_{ji} < \pi_{ii}$ for $j \neq i$.*

Proof. Note that $p_1 + \cdots + p_m = 1$ and differentiate

$$\frac{\partial D_i}{\partial p_i} = \pi_{ii} - \pi_{ji}.$$

Since D_i is an increasing function of p_i, the derivative is positive and hence $\pi_{ii} > \pi_{ji}$ for all $j \neq i$.

ACKNOWLEDGEMENTS

Swartz and Yang were partially supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Korea Science and Engineering Foundation, respectively. The authors are appreciative of the input from three anonymous reviewers and the Editor, Richard A. Lockhart.

REFERENCES

A. P. Basu (1983). Identifiability. In *Encyclopedia of Statistical Sciences*, (S. Kotz & N. L. Johnson, eds.), Wiley Interscience, 4, 2–6.

I. D. J. Bross (1954). Misclassification in 2×2 tables. *Biometrics*, 10, 478–486.

K. J. Clancy, L. E. Ostlund & G. A. Wyner(1979). False reporting of magazine readership. *Journal of Advertising Research*, 19, 23–30.

A. P. Dawid & A. M. Skene (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28, 20–28.

M. Evans, I. Guttman, Y. Haitovsky & T. B. Swartz (1996). Bayesian analysis of binary data subject to misclassification. In *Bayesian Analysis in Statistics and Econometrics*, (D. A. Berry, K. M. Chaloner & J. K. Geweke, eds.), John Wiley, 67–77.

A. Gaba & R. L. Winkler (1992). Implications of errors in survey data. *Management Science*, 38, 913–925.

W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds.) (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall: London.

P. Gustafson (2004). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science*, to appear.

A. Hadgu (1996). The discrepancy in discrepant analysis. *Lancet*, 348, 592–593.

W. O. Johnson, J. L. Gastwirth & L. M. Pearson (2001). Screening without a "gold standard": The Hui–Walter paradigm revisited. *American Journal of Epidemiology*, 153, 921–924.

L. Joseph, T. W. Gyorkos & L. Coupal (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141, 263–272.

M. U. Kalwani & A. J. Silk (1982). On the reliability and predictive validity of purchase intention measures. *Marketing Science*, 1, 243–286.

D. V. Lindley (1971). *Bayesian Statistics; A Review*. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.

R. A. Peterson & R. A. Kerin (1981). The quality of self-report data: review and synthesis. In *Review of*

Marketing, (B. M. Ennis & K. J. Roering, Eds.), American Marketing Association, Chicago: 5–20.

B. D. Ripley (1987). *Stochastic Simulation*. Wiley, New York.

S. N. Singh & G. A. Churchill, Jr. (1986). Using the theory of signal detection to improve ad recognition testing. *Journal of Marketing Research*, 23, 327–336.

S. L. Stewart, K. C. Swallen, S. L. Glaser, P. L. Horn-Ross & D.W. West (1998). Adjustment of cancer incidence rates for ethnic misclassification. *Biometrics*, 54, 774–781.

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

P. F. Thall, D. Jacoby & S. O. Zimmerman (1996). Estimating genomic category probabilities from fluorescent in situ hybridization counts with misclassification. *Applied Statistics*, 4, 431–436.

Y. Wind & D. Lerner (1979). On the measurement of purchase data: surveys versus purchase diaries. *Journal of Marketing Research*, 16, 39–47.

T. Y. Yang & L. Kuo (2003). A Bayesian nonparametric approach to medical binary data with misclassification errors. Manuscript.

Received 11 June 2003
Accepted 19 March 2004

Tim B. SWARTZ: tim@stat.sfu.ca
Department of Statistics and Actuarial Science
Simon Fraser University, Burnaby
British Columbia, Canada V5A 1S6

Yoel HAITOVSKY: msyoel@mscc.huji.ac.il
Departments of Economics and Statistics, Hebrew University
Jerusalem, Israel 91905

Albert VEXLER: valbert@vms.huji.ac.il
Central Bureau of Statistics
Jerusalem, Israel

Tae Y. YANG: tyang@mju.ac.kr
Department of Mathematics, Myongji University
Kyunggi, Korea