Inference for Misclassified Multinomial Data with Covariates

Shijia Wang, Liangliang Wang and Tim B. Swartz [*]

Abstract

This paper considers multinomial data subject to misclassification in the presence of covariates which affect both the misclassification probabilities and the true classification probabilities. A subset of the data may be subject to a secondary measurement according to an infallible classifier. Computations are carried out in a Bayesian setting where it is seen that the prior has an important role in driving the inference. In addition, a new and less problematic definition of nonidentifiability is introduced and is referred to as hierarchical nonidentifiability.

Keywords : gold standard data, latent variables, misclassification, Markov chain Monte Carlo, nonidentifiability.

[*] S. Wang is an Assistant Professor in the School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, No 94 Weijin Road, Tianjin 300071, China. L. Wang and T. Swartz are Associate Professor and Professor, respectively, in the Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. L. Wang and Swartz have been partially supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors thank Dani Chu for assistance in acquiring the dataset in Section 4.2. The authors also thank the Associate Editor and two anonymous reviewers for helpful comments that improved the paper.

1 INTRODUCTION

The analysis of categorical data (Agresti 2013) has a longstanding and extensive literature that is applicable to many disciplines including the social, biomedical and marketing sciences. In some cases, categorical data are misclassified. For example, a subject whose "true" classification is the first category may be incorrectly classified in the second category.

Bross (1954) developed the standard framework for the analysis of binomial data subject to misclassification. In addition, he demonstrated that severely biased estimators can occur when the effect of misclassification is ignored. Since the seminal paper by Bross (1954), there has been a considerable amount of research directed towards misclassification in categorical data. Much of the effort has been from a Bayesian point of view, perhaps due to the fact that misclassification problems yield additional parameters where the number of parameters can sometimes exceed the dimensionality of the data. A literature review of the Bayesian analysis of misclassified multinomial data is given by Perez et. al. (2007).

The work presented here extends two previous papers in significant directions. Swartz et. al. (2004) investigated the Bayesian analysis of multinomial data where special attention was given to the inherent problem of nonidentifiability. Here, we extend their model beyond a single multinomial cohort to the case of subject-specific covariates. In addition, the possibility of gold standard data are considered. Gerlach and Stamey (2007) developed Bayesian methods for variable selection in logistic regression models where misclassification is present. Here, we extend their model to the multinomial context where Dirichlet distributions are assigned to the primary parameters.

In Section 2, we present the multinomial model subject to misclassification in the presence of covariates. Whereas our focus is on data that are categorized subject to misclassification, we also provide modelling for *gold standard data* that are classified by an infallible classifier. A prior distribution is proposed where alternative parameterizations are provided according to default (i.e. reference) distributions or highly subjective priors.

In Section 3, we explore pragmatic issues resulting from the highly parametrized model. In particular, the problem of nonidentifiability is investigated, and a new definition of nonidentifiability is proposed that is particularly relevant to Bayesian settings. Computation is also discussed, and this is paramount in high-dimensional problems. In Section 4, we address the analysis of two datasets. The simulated dataset allows us to assess the reliability of inferences when the true paramenters are known. We also are able to investigate the importance of the priors as we alter the amount of gold standard data. We then consider an actual dataset where the results appear to be sensible. Some concluding remarks are provided in Section 5.

2 MODEL DEVELOPMENT

2.1 The Data Model

Consider n_i independent observations belonging to the ith covariate pattern, $i = 1, \ldots, r$, and assume that each observation is classified according to one of q categories. Then the jth observation adhering to the ith covariate pattern takes the form $y_{ij} = (y_{ij1}, \ldots, y_{ijq})'$ where $y_{ijl} = 1$ denotes that the observation is classified according to category l and $y_{ijk} = 0$ for $k \neq l$. Without misclassification, and assuming independence between observations, this is a standard product multinomial model and the likelihood is given by

$$L_0 = \prod_{i=1}^{r} \prod_{j=1}^{n_i} \prod_{k=1}^{q} p_{ik}^{y_{ijk}}$$

where $p_{ik} = \text{Prob}(y_{ijk} = 1)$ for $k = 1, \ldots, q$, $j = 1, \ldots, n_i$, $i = 1, \ldots, r$.

In the presence of misclassification, we denote p_{il} as the probability that a subject belonging to the ith covariate pattern has the true (but unobserved) classification l. In other words, whereas y_{ijk} is the observed classification, the true classification is a latent variable. Further, we let π_{ilk} denote the probability that a subject with the ith covariate pattern is classified as k given that its true classification is l. By the law of total probability, it follows that the probability of classification in category k for a subject with the ith covariate pattern is $\sum_{l=1}^{q} p_{il}\pi_{ilk}$. This leads to the more complex likelihood

$$\begin{aligned} L_1 &= \prod_{i=1}^{r}\prod_{j=1}^{n_i}\prod_{k=1}^{q}(\sum_{l=1}^{q} p_{il}\pi_{ilk})^{y_{ijk}} \\ &= \prod_{i=1}^{r}\prod_{k=1}^{q}(\sum_{l=1}^{q} p_{il}\pi_{ilk})^{\sum_{j=1}^{n_i} y_{ijk}}. \qquad (1) \end{aligned}$$

Therefore the parametrization of the standard product multinomial model has increased considerably with the addition of the $rq(q-1)$ misclassification parameters π_{ilk}.

We also allow for the case where some observations are doubly classified; first as above where the possibility of misclassification exists. Secondly, these observations are classified according to an infallible classifier where the classification is known to be correct. An infallible classifier is typically more expensive or less readily available than a fallible classifier. With double classification, the jth observation adhering to the ith covariate pattern takes the form (x_{ij}, z_{ij}) where $x_{ij} = (x_{ij1}, \ldots, x_{ijq})'$ corresponds to the fallible classifier and $z_{ij} = (z_{ij1}, \ldots, z_{ijq})'$ corresponds to the infallible classifier. The pair (x_{ij}, z_{ij}) are known as gold standard data where each vector consists of $q-1$ zeros and a single one. We let m_i denote the number of observations in the ith covariate class that are doubly classified. This leads to the gold standard likelihood contribution

$$\begin{aligned} L_2 &= \prod_{i=1}^{r}\prod_{j=1}^{m_i} \text{Prob}(x_{ij}, z_{ij}) \\ &= \prod_{i=1}^{r}\prod_{j=1}^{m_i} \text{Prob}(z_{ij})\ \text{Prob}(x_{ij} \mid z_{ij}) \\ &= \prod_{i=1}^{r}\prod_{j=1}^{m_i}(\prod_{k=1}^{q} p_{ik}^{z_{ijk}})(\prod_{l=1}^{q}\prod_{k=1}^{q} \pi_{ilk}^{z_{ijl}x_{ijk}}) \\ &= \prod_{i=1}^{r}\prod_{k=1}^{q}(p_{ik}^{\sum_{j=1}^{m_i} z_{ijk}})(\prod_{l=1}^{q} \pi_{ilk}^{\sum_{j=1}^{m_i} z_{ijl}x_{ijk}}). \qquad (2) \end{aligned}$$

We note that with some covariate patterns, there may be no data that are solely classified with a fallible classifier (i.e. $n_i = 0$), or there may be no doubly classified data (i.e. $m_i = 0$). Putting together the likelihoods associated with the data that are solely classified with a fallible classifier (1) and the gold standard data (2), the overall likelihood for the misclassified multinomial model in the presence of covariates is given by

$$L = L_1 \cdot L_2. \tag{3}$$

Although a less realistic scenario, it is also possible to have data that are only classified by the infallible classifier. Such data provide information about the true classification probabilities p_{ik}. We define the data vector w_{ij} with s_i cases corresponding to the ith covariate pattern. Although we do not study this case, for reference, we provide its likelihood contribution

$$\begin{aligned} L_3 &= \prod_{i=1}^{r} \prod_{j=1}^{s_i} \text{Prob}(w_{ij}) \\ &= \prod_{i=1}^{r} \prod_{j=1}^{s_i} (\prod_{k=1}^{q} p_{ik}^{w_{ijk}}) \\ &= \prod_{i=1}^{r} \prod_{k=1}^{q} (p_{ik}^{\sum_{j=1}^{s_i} w_{ijk}}). \end{aligned}$$

2.2 The Prior Distribution

In specifying the prior distribution, we use the generic notation $[A \mid B]$ to denote the conditional probability density function of A given B. In addition to the true classification probabilities p and the misclassification probabilities π developed above, we introduce additional parameters a, b, σ_a and σ_b. Via conditional probability, the prior distribution is given by

$$[p, \pi, a, b, \sigma_a, \sigma_b] \;=\; [p, \pi \mid a, b, \sigma_a, \sigma_b] \; [a, b \mid \sigma_a, \sigma_b] \; [\sigma_a, \sigma_b] \tag{4}$$

where the parameter hierarchy is apparent; p and π are primary parameters, a and b are secondary parameters and σ_a and σ_b are tertiary parameters. The hierarchical labelling is relevant to the discussion in Section 3.1.

In Gerlach and Stamey (2007), it was convenient to model the logits of the true classification probabilities as linear combinations of covariate vectors u with an additive normal error. Given constraints $\sum_{k=1}^{q} p_{ik} = 1$ in the multinomial setting, we instead assign a prior distribution to the p_{ik} parameters via

$$(p_{i1}, \ldots, p_{iq})' \sim \text{Dirichlet}(\exp(a_1' u_i), \ldots, \exp(a_q' u_i)) \tag{5}$$

where the true classifications are covariate dependent. Note that the distribution in (5) corresponds to the ith covariate pattern and independence is assumed for $i = 1, \ldots, r$. Typically, it would be sensible to include a constant (intercept) term in the covariate vector u_i.

As in Gerlach and Stamey (2007), it is also reasonable to consider misclassification rates that are covariate dependent. We therefore assign a prior distribution to the misclassification rates via

$$(\pi_{il1}, \ldots, \pi_{ilq})' \sim \text{Dirichlet}(\exp(b_{l1}' u_i), \ldots, \exp(b_{lq}' u_i)) \tag{6}$$

where $i = 1, \ldots, r$, $l = 1, \ldots, q$ and (6) is conditionally independent of (5). The prior assignments in (5) and (6) are the first steps in the hierarchical model building process.

The second step in the hierarchical model building process involves the assignment of hyperpriors for the a's and b's. With covariate vectors of dimension v, these priors are assigned as follows:

$$\begin{aligned}(a_{k1}, \ldots, a_{kv})' &\sim \text{Normal}((a_{k1}^{(0)}, \ldots, a_{kv}^{(0)})', D_{\sigma_a})\\ (b_{lk1}, \ldots, b_{lkv})' &\sim \text{Normal}((b_{lk1}^{(0)}, \ldots, b_{lkv}^{(0)})', D_{\sigma_b})\end{aligned}$$

where $l, k = 1, \ldots, q$, and D_{σ_a} and D_{σ_b} are diagonal matrices with diagonal vectors σ_a and

σ_b, respectively. The components of σ_a and σ_b have priors that are independent Inverse Gamma($r_a^{(0)}, s_a^{(0)}$) and Inverse Gamma($r_b^{(0)}, s_b^{(0)}$) distributions, respectively. Parameters with a superscript (0) are user-specified.

Hence the posterior distribution is proportional to the product of (3) and (4) with possible parameter constraints used to ameliorate nonidentifiability issues (Swartz et. al. 2004).

3 PRAGMATIC ISSUES

3.1 Nonidentifiability

As modern statistical practice entertains increasingly complex models, the problem of nonidentifiability has become an increasingly important topic. Not only is it sometimes difficult to handle nonidentifiability issues, it is sometimes difficult to even detect nonidentifiability. Strangely, the topic of nonidentifiability does not appear to be receiving its due coverage in current statistics curricula. In a non-scientific survey of the third author's bookshelf, he discovered 12 texts on mathematical statistics, and only two of the texts had "identifiability" or "nonidentifiability" or "unidentifiability" listed in the table of contents.

A statistician's first encounter with nonidentifiability often occurs in a regression context. Consider the simple one-way ANOVA

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where y_{ij} is the jth response under treatment i, μ is the overall mean, τ_i is the ith treatment effect, $i = 1, 2$, and the ϵ_{ij} are random error terms. In matrix notation, $y = (y_{11}, \ldots, y_{in_1}, y_{21}, \ldots, y_{2n_2})'$ and the design matrix is given by $X = (\underline{1}, x_1, x_2)$ where $\underline{1}$ is a vector of ones, x_1 is a vector of n_1 ones followed by n_2 zeros, and x_2 is a vector of n_1 zeros followed by n_2 ones. The least squares estimator of $(\mu, \tau_1, \tau_2)'$ given by $(X'X)^{-1}X'y$ cannot be calculated since $X'X$ is not invertible (i.e. X is not full rank). Clearly, this is a problem, but a simple solution is to introduce the constraint $\tau_1 + \tau_2 = 0$.

More formally, Basu (1983) defined nonidentifiability as follows:

Definition. 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.

Basu's definition essentially states that a nonidentifiable model is one where no matter what data are observed, there are parameter values that cannot be distinguished. If we interpret nonidentifiability as a pejorative term concerning models, then perhaps the Basu (1983) definition is overly rigid, and the definition may not be appropriate for Bayesian statistical practice. For example, consider the following posterior density expressed in a common hierarchical structure

$$[\theta_1, \theta_2 \mid y] \propto [y \mid \theta_1]\ [\theta_1 \mid \theta_2]\ [\theta_2]. \tag{7}$$

Now (7) may be considered the bread and butter of Bayesian modelling. There is certainly nothing wrong with (7), and such models are widespread in statistical practice. However, according to the Basu (1983) definition, there exists a nonidentifiability in model (7) as both (θ_1, θ_2) and (θ_1, θ_2') yield the same sampling distribution given by $[y \mid \theta_1]$. In fact, all hierarchical models, which are the mainstream of Bayesian modelling are nonidentifiable.

On the opposite side of the spectrum, there is an extreme point of view with respect to nonidentifiability and Bayesian statistics. Lindley (1971) stated, "In passing, it might be noted that unidentifiability causes no real difficulty in the Bayesian approach." Reading into Lindley's statement, he may have intended to convey the message that after writing down the posterior which is proportional to the product of the likelihood and prior, there is no obstacle to integrating.

Despite Lindley's claim, nonidentifiability is a problem for Bayesian statistical practice, and this has been explored by many authors including Gustafson (2005, 2010) and Martin and Gonzalez (2010). A common problem concerning nonidentifiability is that likelihood is assigned to regions of the parameter space which are not apriori plausible. For example, consider the simple problem discussed in Evans et al. (1996). Here, there is a random patient having probability p of being sick, probability θ of being misdiagnosed as healthy and probability ϕ of being misdiagnosed as sick. Therefore, the probability that the patient is diagnosed sick is $f(p, \theta, \phi) = p(1-\theta) + (1-p)\phi$. This leads to a nonidentifiable model as $f(p, \theta, \phi) = f(1-p, 1-\phi, 1-\theta)$. But moreover, we see that the model does not correspond to common sense as $f(p, \theta, \phi)$ is not necessarily an increasing function of p (without introducing appropriate constraints).

Essentially, problematic issues involving nonidentifiability in Bayesian statistics may be viewed as an artefact of non-sensical model building with respect to the sampling model.

In Swartz et. al. (2004), various constraints were introduced on the parameter space to avoid non-sensical model building in the misclassified multinomial model. A simple constraint which was suggested there, and will be utilized in this paper is

$$\pi_{ilk} < \pi_{ill} \tag{8}$$

for $i = 1, \ldots, r$ and all $l \neq k$. The constraint (8) essentially states that it is more probable for a subject to be classified correctly than incorrectly. Although (8) does not entirely take care of nonidentifiability in our model, it goes a long way in improving inferences as it eliminates unacceptable regions of the posterior space.

Constraints other than (8) have also been discussed in the literature with respect to nonidentifiability. For example, Wang and Gustafson (2014) consider the misclassification of an ordinal exposure variable in the context of hypothesis testing. Such frameworks are common in studies where disease outcome is regressed against exposure. Wang and Gustafson provide discussion of how power is lost and gained when misclassification probabilities are *tapered* (i.e. decrease in probability as one moves away from the true classification). Ogburn and Vanderweele (2013) investigate the setting where there is a binary treatment and an ordinal confounding variable which is subject to misclassification. Ogburn and Vanderweele (2013) impose monotonicity assumptions on the effect of the misclassified confounder on the treatment and the outcome.

How do we then reconcile the overly rigid Basu (1983) definition with legitimate concerns for nonidentifiability in Bayesian practice? We suggest an alternative definition which takes the Bayesian hierarchical structure into account. For ease of notation, let $y = \theta_0$, and express the posterior density as

$$\begin{aligned} [\theta_1, \ldots, \theta_k \mid \theta_0] &\propto [\theta_0 \mid \theta_1]\ [\theta_1 \mid \theta_2] \cdots [\theta_{k-1} \mid \theta_k] \\ &= \prod_{i=1}^{k} [\theta_{i-1} \mid \theta_i] \end{aligned}$$

where $\theta_1, \ldots, \theta_k$ are potentially vector-valued and there is uniqueness in the hierarchical structure. We then say that a Bayesian model contains a *hierarchical nonidentifiability* if for some $i = 1, \ldots, k$, there exists $\theta_i \neq \theta_i'$ for which $[\theta_{i-1} \mid \theta_i] = [\theta_{i-1} \mid \theta_i']$ for all values θ_{i-1}. The idea is that nonidentifiability occurs when at least one of the links in the hierarchical model do not identify between parameters. This is a weaker definition than the Basu (1983) definition, and the proposed definition prevents the immediate disqualification of hierarchical models as being nonidentifiable.

In the model considered in this paper, we refer to the expressions in Section 2.2, and obtain the posterior density

$$[p, \pi, a, b, \sigma_a, \sigma_b \mid y, x, z] \propto [y, x, z \mid p, \pi]\ [p, \pi \mid a, b, \sigma_a, \sigma_b]\ [a, b \mid \sigma_a, \sigma_b]\ [\sigma_a, \sigma_b] \qquad (9)$$

where $[y, x, z \mid p, \pi] = L_1 L_2$ from (1) and (2). Viewing the terms in (9), the only possibility of hierarchical nonidentifiability occurs in the $[y, x, z \mid p, \pi]$ specification. In fact, if complete gold standard data exists (i.e. $m_i \geq 1$, $i = 1, \ldots, r$), then the model is hierarchically identifiable. In the absence of gold standard data (i.e. no L_2 term), there are nonidentifiability issues as discussed in Swartz et. al. (2004). The impact of gold standard data, the prior specification and constraints with respect to nonidentifiability and inference are explored in the simulated data example of Section 4.1.

3.2 Computation

Misclassified multinomial data in the presence of covariates leads to a complex model where the number of observations $\sum_{i=1}^{r}(n_i + m_i)$ is sometimes exceeded by the number of parameters. The model parameters are summarized in Table 1.

Parameters	Description	Number of Parameters
p_{ik}	primary parameters of interest which describe true classification probabilities	$r(q-1)$
π_{ilk}	primary parameters of interest which describe misclassification probabilities	$rq(q-1)$
a_{kt}	secondary parameters of interest relating covariates to true classification probabilities	qv
b_{lkt}	secondary parameters of interest relating covariates to misclassification probabilities	q^2v
σ_a, σ_b	tertiary parameters of interest	$2v$

Table 1: Parameter summary for the misclassified multinomial model with covariates.

With so many parameters and constraints, inferential procedures relying on exact integration are clearly infeasible. Approximation techniques such as Laplace's method, importance sampling and quadrature (Evans and Swartz 2000) are also not well-suited to this application. It seems that a Markov chain procedure may be the only avenue for obtaining inferential summaries. In a Markov chain approach, a Markov chain is constructed whose equilibrium distribution corresponds to the posterior. By averaging over samples, posterior estimates can be approximated.

It was originally our intention to use a Bayesian programming language such as WinBUGS (Lunn et al. 2013) or JAGS (Plummer 2015) to implement the models considered in this paper. Such software is attractive as it avoids the need of special purpose Markov chain Monte Carlo (MCMC) code. However, we were unsuccessful with an RJAGS implementation as the Markov chains would hang after a number of iterations.

Our next computational attempt involved the consideration of a Gibbs sampling algorithm. This may be facilitated via data augmentation where the latent variable T_{ij} is introduced in the context of data that are only classified by a fallible classifier. For $i = 1, \ldots, r$, $j = 1, \ldots, n_i$, we define

$T_{ij} \equiv k$ if the jth observation sharing the ith covariate pattern has true classification k

Accordingly, the posterior density simplifies and takes the form

$$[p, \pi, a, b, \sigma_a, \sigma_b, T \mid y, x, z] \propto \left(\prod_{i=1}^{r} \prod_{j=1}^{n_i} p_{i,T_{ij}} \pi_{i,T_{ij},\sum_{k=1}^{q} k y_{ijk}} \right) \cdot L_2 \cdot [p, \pi, a, b, \sigma_a, \sigma_b]$$

where the likelihood L_2 and the prior $[p, \pi, a, b, \sigma_a, \sigma_b]$ are given in Section 2.2.

Although it is possible to derive the required full conditional distributions from the above expression, it is clearly a laborious process that is prone to coding errors. It is especially tedious when small changes to the model or changes in parameter constraints invoke considerable changes to the full conditional distributions.

For this problem, we then turned to a hybrid of Gibbs sampling and slice sampling (Neal 2003) to generate variates from the joint posterior distribution $[p, \pi, a, b, \sigma_a, \sigma_b, T \mid y, x, z]$. Details concerning the implementation of the algorithm are provided in the Appendix.

4 EXAMPLES

4.1 Simulated Data

We considered $r = 50$ covariate patterns and $q = 5$ categorical responses which is typical of surveys carried out on a Likert scale. In the baseline dataset, we began with 20% gold standard data where there are 1000 observations. This was accomplished by setting $n_i = 0$ for covariate patterns $i = 5, 10, 15, \ldots, 50$ and $n_i = 20$ otherwise using the fallible classifier according to (1). For the gold standard data, we set $m_i = 20$ for covariate patterns $i = 5, 10, 15, \ldots, 50$ and $m_i = 0$ otherwise according to (2).

The covariates u_i in (5) and (6) have dimension $v = 2$ where $u_i = (1, i)'$ contains a constant term and a term which increases linearly according to the covariate pattern i. The parameter coefficients corresponding to the covariate in (5) were set according to $a_1 = (0.1, 0.05)'$, $a_2 = (0.2, 0.05)'$, $a_3 = (0.3, 0.05)'$, $a_4 = (0.4, 0.05)'$ and $a_5 = (0.5, 0.05)'$. This implies that the probabilities of classification p_{ik} have prior probability distributions that are increasing across the five cells but the effect diminishes as i increases.

For the misclassification probabilities π_{ilk} in (6), we set the parameter coefficients in (6) according to

$$\begin{array}{lllll}
b_{11} = (2.0, 0.0)' & b_{12} = (1.0, 0.0)' & b_{13} = (0.0, 0.0)' & b_{14} = (0.0, 0.0)' & b_{15} = (0.0, 0.0)' \\
b_{21} = (1.0, 0.0)' & b_{22} = (2.0, 0.0)' & b_{23} = (1.0, 0.0)' & b_{24} = (0.0, 0.0)' & b_{25} = (0.0, 0.0)' \\
b_{31} = (0.0, 0.0)' & b_{32} = (1.0, 0.0)' & b_{33} = (2.0, 0.0)' & b_{34} = (1.0, 0.0)' & b_{35} = (0.0, 0.0)' \\
b_{41} = (0.0, 0.0)' & b_{42} = (0.0, 0.0)' & b_{43} = (1.0, 0.0)' & b_{44} = (2.0, 0.0)' & b_{45} = (1.0, 0.0)' \\
b_{51} = (0.0, 0.0)' & b_{52} = (0.0, 0.0)' & b_{53} = (0.0, 0.0)' & b_{54} = (1.0, 0.0)' & b_{55} = (2.0, 0.0)' \ .
\end{array}$$

This implies that misclassification occurs in neighbouring categories with the greatest prior probability and does not depend on the covariate pattern i.

Using the above settings, we generated the underlying classification probabilities p_{ik} and the underlying misclassification probabilities π_{ilk} according to the Dirichlet distributions in (5) and (6). In turn, we used multinomial distributions to generate the data y_{ij} according to the fallible classifier and to generate the doubly classified gold standard data (x_{ij}, z_{ij}).

We then repeat the entire simulation process described above $N = 20$ times. This provides us with N datasets from which we can assess the performance of the model and the MCMC procedure.

The model is completely specified except for the hyperparameters. For the hyperparameters, we begin with settings $a_{kj}^{(0)} = b_{lkj}^{(0)} = 0$ for all $k = 1, \ldots, q$, $l = 1, \ldots, q$ and $j = 1, \ldots, v$. We also introduce diffuse hyperpriors $\sigma_a \sim$ Inverse Gamma$(1.0, 0.01)$ and $\sigma_b \sim$ Inverse Gamma$(1.0, 0.01)$.

There is no shortage of ways that one might investigate the simulation results. We choose to concentrate on the true classification probabilities p_{ik} since these are the parameters of primary interest. Let p_{ik} denote the underlying values that were generated and let $\tilde{p}_{ik}$ denote the estimated posterior means from the Markov chain. For the jth dataset, $j = 1, \ldots, N$, we define the statistic

$$D^{(j)} = \frac{1}{rq} \sum_{i=1}^{r} \sum_{k=1}^{q} (p_{ik} - \tilde{p}_{ik})^2 \tag{10}$$

as a measure of the ability of the model to estimate the classification probabilities. We interpret $D^{(j)}$ in (10) as the average distance between an estimated p_{ik} and its underlying value. Smaller values of $D^{(j)}$ indicate successful estimation. An overall measure of estimation (i.e. over all datasets) is given by $D = (1/N) \sum_{j=1}^{N} D^{(j)}$ and $SD = \sqrt{\sum_{j=1}^{N} (D^{(j)} - D)^2/(N-1)}$. Standard diagnostics were assessed to confirm convergence of the Markov chains.

In the baseline model, although there was evidence of autocorrelation in some of the parameters b_{lk2}, we observed stability in the parameter estimates. Here, we observed $D = 0.0069$ ($SD = 0.0012$) based on $100,000$ iterations of the Markov chain. We then increased the size of the dataset from 1000 observations to 5000 observations and retained the 20% gold standard data ratio. In this case, we observed $D = 0.0051$ ($SD = 0.0008$). This suggests (as expected) that more data improves estimation.

To investigate the effect of the constraint (8) which is intended to reduce nonidentifiability issues, we repeated the simulation procedure with the original 1000 observations under the baseline model without the constraint. In this case, we observed a larger $D = 0.0073$ ($SD = 0.0014$) which suggests that the constraint is mildly helpful in this example.

It is also the case that a higher proportion of gold standard data improves estimation. This was observed by setting $n_i = 0$ for covariate patterns $i = 2, 4, 6, \ldots, 50$ and $n_i = 20$ for covariate patterns $i = 1, 3, 5, \ldots, 49$ using the fallible classifier according to (1). For the gold standard data, we set $m_i = 20$ for covariate patterns $i = 2, 4, 6, \ldots, 50$ and

$m_i = 0$ otherwise according to (2). In this case, we had 50% gold standard data and 1000 observations as before. In this case, there was a reduction from $D = 0.0069$ in the baseline model to $D = 0.0056$ ($SD = 0.0007$).

We then investigated the sensitivity of our procedures with respect to the prior specification. We set the hyperparameters according to $a_k^{(0)} = (0.3, 0.05)'$, $b_{kk}^{(0)} = (2.0, 0.0)'$ and $b_{lk}^{(0)} = (0.0, 0.0)'$ ($l \neq k$) for all $k = 1, \ldots, q$ and $l = 1, \ldots, q$. Previously, the hyperpriors on σ_a and σ_b were diffuse. With a more informative prior Inverse Gamma$(10000.0, 1.0)$, less probability is assigned to implausible regions in the contours resulting from nonidentifiability. This resulted in quicker mixing and improved estimation; we obtained $D = 0.0064$ ($SD = 0.0011$) which is improved when compared to the baseline results.

Finally, we report on some details associated with the baseline case in the simulation study. We required 661.6 seconds (i.e. roughly 11 minutes) of computation on an Intel Core i5, 1.3 GHz processor based on 10,000 iterations of the Markov chain. The average acceptance rate of the rejection sampler was 91.7% which caused no difficulties for our sampling scheme. The average MCMC effective sample size for p_{ik} was $7,525$ based on $50,000$ MCMC iterations after burn-in.

4.2 Actual Data and Analysis

The following example is relatively simple. However, the application area is novel, and the results are both readily interpretable and interesting.

Here we consider data taken from Replay Challenges during the last five regular seasons (2014-2018) of the National Football League (NFL). During this period, there have been no rule changes with respect to Replay Challenges. In an attempt to improve decision making by game officials, each of the two coaches in an NFL game is allowed to appeal the decisions of up to two plays using video replay technology. In the case where a coach has made two successful challenges, he is allowed to make a third challenge.

The data were obtained using the R package *nflscrapr* (Horowitz, Yurko and Ventura 2018), and we have restricted the dataset to the 410 coaches' challenges that involved pass completion rulings. In this dataset, we assume the result of the video replay is correct and this provides us with gold standard data. Therefore, following (2), (x_{ij1}, x_{ij2}) and (z_{ij1}, z_{ij2}) correspond to the jth challenge under the ith covariate pattern where $(x_{ij1}, x_{ij2}) = (1, 0)$ if the play was ruled a completion on the field, $(x_{ij1}, x_{ij2}) = (0, 1)$ if the play was ruled an incompletion on the field, $(z_{ij1}, z_{ij2}) = (1, 0)$ if the play was ruled a completion by video replay and $(z_{ij1}, z_{ij2}) = (0, 1)$ if the play was ruled an incompletion by video replay. In this application, we are mostly interested in the misclassification probabilities π_{i11} that the video replay confirmed a completion and π_{i12} that the video replay reversed an incompletion to a completion. The misclassification probabilities are important as they address the accuracy of the original calls.

For simplicity, we consider $r = 2$ covariate patterns:

- $i = 1 \equiv$ challenge by the home team
- $i = 2 \equiv$ challenge by the road team

It is instructive to note that the home team will only challenge calls corresponding to an incompletion by the home team or a completion by the road team. Similarly, the road team will only challenge calls corresponding to an incompletion by the road team or a completion by the home team.

We set default covariates $u_1 = u_2 = 1.0$, and default hyperparameters $a^{(0)}_{11} = a^{(0)}_{21} = 0.0$, $b^{(0)}_{111} = b^{(0)}_{121} = b^{(0)}_{211} = b^{(0)}_{221} = 0.0$, $r^{(0)}_a = r^{(0)}_b = 1.0$, $s^{(0)}_a = s^{(0)}_b = 0.01$. For this application, we did not impose constraints as in (8) since the gold standard data alleviates nonidentifiability.

For this simple problem, the Markov chain converged rapidly. We obtained the following posterior estimates of the true classification and the misclassification probabilities: $p_{11} = 0.53$, $p_{21} = 0.48$, $\pi_{112} = 0.37$, $\pi_{121} = 0.55$, $\pi_{212} = 0.36$ and $\pi_{221} = 0.51$. Panel 1 and panel 2 of Figure 1 display posterior histograms of $\pi_{112} - \pi_{212}$ and $\pi_{121} - \pi_{221}$. Panel 3 and panel 4 of Figure 1 display scatterplots for posterior samples of π_{112}, π_{212} and π_{121}, π_{221}. Based on the point estimates, our most stunning observation is that incorrect on-field decisions in favour of the road team are slightly more probable than incorrect on-field decisions in favour of the home team (i.e. $\pi_{112} > \pi_{212}$ and $\pi_{121} > \pi_{221}$). However, referring to panel 1 of Figure 1, the posterior variability suggests that there there is no evidence of the phenomenon known as the "home team advantage" in the context of replay challenges in the NFL; see Swartz and Arce (2014) and the references therein. This is in keeping with recent discussions (https://twitter.com/StatsbyLopez/status/1161775001596567559) indicating that the home team advantage has been decreasing over time.

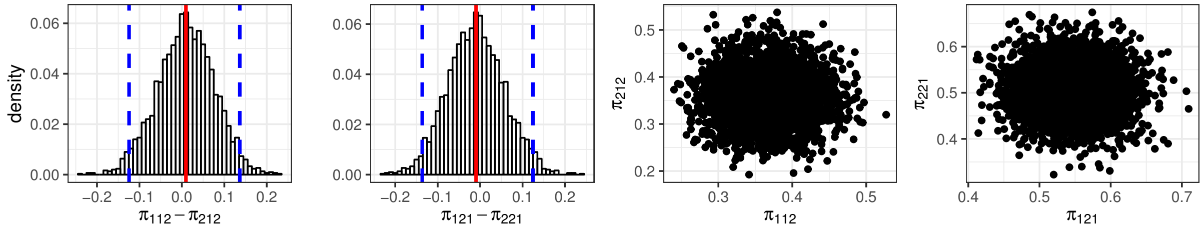

Figure 1: Histograms of $\pi_{112} - \pi_{212}$ and $\pi_{121} - \pi_{221}$, and scatterplots for posterior samples of π_{112} vs π_{212} and π_{121} vs π_{221}. The red vertical lines indicate posterior means and the blue vertical lines indicate 2.5% and 97.5% quantiles respectively.

Furthermore, under covariate pattern i, we can express

$$P_{iU} = \text{Prob(on-field decision upheld)} = p_{i1}\pi_{i11} + p_{i2}\pi_{i22} \tag{11}$$

and

$$P_{iR} = \text{Prob(on-field decision reversed)} = 1 - P_{iU}\ .$$

Using expression (11), we obtain estimated posterior means $P_{1U} = 0.55$ and $P_{2U} = 0.56$. Therefore, when challenged, the initial calls on the field are correct more than half the time. This may be a surprising since coaches typically invoke challenges only when they believe that officials have made mistakes. Figure 2 displays the posterior histograms of P_{1U}/P_{1R} and P_{2U}/P_{2R}. The plots confirm that in the context of replay challenges, the on-field calls are correct most of the time.

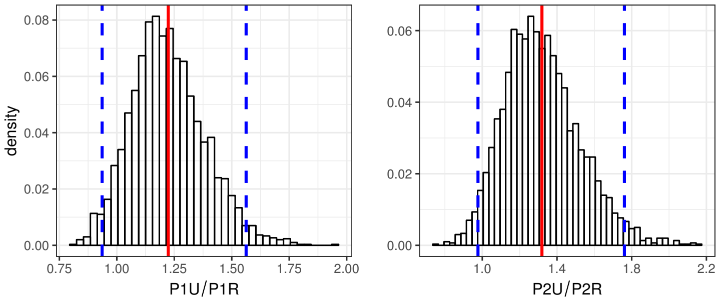

Figure 2: Histograms of P_{1U}/P_{1R} and P_{2U}/P_{2R}. The red vertical lines indicate posterior means and the blue vertical lines indicate 2.5% and 97.5% quantiles respectively.

4.3 Reduced Gold Standard Data

We consider the NFL Replay Challenge data from Section 4.2 where we utilize the full likelihood (3). To do this, we eliminate some of the gold standard data (z_{ij1}, z_{ij2}) so that we have both a fallible classifier (i.e. the on-field calls) and gold standard data (i.e. the video review results). Of the 410 challenges, we randomly eliminate 10%, 20% and 30% of the gold standard data corresponding to 41, 82 and 123 replay challenge decisions. We use the same prior distributions as in Section 4.2.

In these analyses, we again consider estimation of the misclassification probabilities π_{112}, π_{121}, π_{212} and π_{221}. The results are given in Table 2. As expected, we observe that when the gold standard data are plentiful (e.g. 10% reduction), inferences are nearly the same as in the full data analysis from Section 4.2. As more gold standard data are eliminated, the estimates begin to drift. However, even with a 30% reduction in gold standard data, there appears to be sufficient information in the dataset to give estimates that are in the vicinity of the original estimates. We also observe that the posterior standard deviations increase with less gold standard data.

Parameters	Full Data	10% Reduction	20% Reduction	30% Reduction
π_{112}	0.366 (0.043)	0.362 (0.044)	0.359 (0.045)	0.366 (0.046)
π_{122}	0.548 (0.046)	0.542 (0.048)	0.533 (0.051)	0.545 (0.051)
π_{212}	0.359 (0.052)	0.357 (0.053)	0.329 (0.054)	0.317 (0.056)
π_{221}	0.506 (0.052)	0.506 (0.054)	0.479 (0.054)	0.472 (0.056)

Table 2: Parameter estimates based on fractional reductions of the gold standard data. Posterior standard deviations are given in parentheses.

5 CONCLUDING REMARKS

This paper has expanded earlier models on multinomial misclassification by considering the inclusion of subject specific cohorts in the presence of both fallible and infallible classifiers. In addition, an alternative definition of nonidentifiability has been proposed which appears well-suited to hierarchical models.

From our investigations on sample datasets, we have observed that (1) nonidentifiability resulting from misclassification complicates inference in this class of problems, (2) that sufficient gold standard data helps alleviate the nonidentifiability issue and that (3) a sufficiently informative prior helps alleviate the nonidentifiability issue.

In this project, MCMC algorithms have been implemented in R to sample from the corresponding distributions. The code is available from the authors upon request.

6 APPENDIX

6.1 MCMC sampling for $[p, \pi, a, b, \sigma_a, \sigma_b, T \mid y, x, z]$

For this problem, we implemented a hybrid of Gibbs sampling and slice sampling (Neal 2003) to generate variates from the joint posterior distribution $[p, \pi, a, b, \sigma_a, \sigma_b, T \mid y, x, z]$.

We sample $[p, \pi, T| \cdot]$ according to following distributions

$$\begin{aligned}
[T_{ij} = m| \cdot] &= p_{im} \cdot \pi_{im \sum_{k=1}^{q} ky_{ijk}} / \sum_{l=1}^{q} p_{il} \cdot \pi_{il \sum_{k=1}^{q} ky_{ijk}}, \text{ where } j = 1, \ldots, n_i \\
[p_{i1}, \ldots, p_{iq}| \cdot] &\sim \text{Dirichlet}(\sum_{j=1}^{n_i} I(T_{ij} = 1) + \sum_{j=1}^{m_i} z_{ij1} + \exp(a_1' u_i), \ldots, \\
&\quad \sum_{j=1}^{n_i} I(T_{ij} = q) + \sum_{j=1}^{m_i} z_{ijq} + \exp(a_q' u_i)) \\
[\pi_{il1}, \ldots, \pi_{ilq}| \cdot] &\sim \text{Dirichlet}(c_{il1}, \ldots, c_{ilq}), \\
&\quad c_{ilm} = \exp(b_{lm}' u_i) + \sum_{j=1}^{n_i} I(T_{ij} = l) I(\sum_{k=1}^{q} ky_{ijk} = m) + \sum_{j=1}^{m_i} z_{ijl} x_{ijm}
\end{aligned}$$

where $i = 1, \ldots, r$, $l = 1, \ldots, q$, $m = 1, \ldots, q$.

For some covariate patterns that data are solely classified with a fallible classifier, we sample $[p, \pi, T| \cdot]$ occurring in the L1 case according to following distributions

$$\begin{aligned}
[T_{ij} = m| \cdot] &= p_{im} \cdot \pi_{im \sum_{k=1}^{q} ky_{ijk}} / \sum_{l=1}^{q} p_{il} \cdot \pi_{il \sum_{k=1}^{q} ky_{ijk}} \\
[p_{i1}, \ldots, p_{iq}| \cdot] &\sim \text{Dirichlet}(\sum_{j=1}^{n_i} I(T_{ij} = 1) + \exp(a_1' u_i), \ldots, \sum_{j=1}^{n_i} I(T_{ij} = q) + \exp(a_q' u_i)) \\
[\pi_{il1}, \ldots, \pi_{ilq}| \cdot] &\sim \text{truncated-Dirichlet}(c_{il1}, \ldots, c_{ilq}), \\
&\quad c_{ilm} = \exp(b_{lm}' u_i) + \sum_{j=1}^{n_i} I(T_{ij} = l) I(\sum_{k=1}^{q} ky_{ijk} = m)
\end{aligned}$$

where $i = 1, \ldots, r$, $l = 1, \ldots, q$, $m = 1, \ldots, q$. The truncated-Dirichlet distributions are truncated according to constraint (8). We use the rejection sampling algorithm and only accept samples that satisfy the constraint. However, rejection sampling may be too inefficient for a large value of q, as a high proportion of proposed candidates are rejected. We could consider a more efficient construction to generate samples for truncated-Dirichlet distributions. As discussed in Swartz et. al. (2004), for $k = 1, \ldots, q-1$, we sample

$$[\tilde{\pi}_{ilk} |\ \pi_{il1}, \ldots, \pi_{il,k-1}, \pi_{il,k+1}, \ldots, \pi_{il,q-1}\ ] \ \sim \ \text{truncated-Beta}(c_{ilk}, c_{ilq}),$$

where $\tilde{\pi}_{ilk} = \pi_{ilk}/\pi_{il(k)}$, $\pi_{il(k)} = 1 - \pi_{il1} - \ldots - \pi_{il,k-1} - \pi_{il,k+1} - \ldots - \pi_{il,q-1}$, $\tilde{\pi}_{ilk}$ is truncated on interval $(0, \pi_{ill}/\pi_{il(k)})$ for $k \neq l$, and is truncated on interval $(\max_{k \neq l} \pi_{ilk}, 1)$ for $k = l$.

For some covariate patterns that no data are solely classified with a fallible classifier, the full conditional densities $[p|\cdot\,]$ and $[\pi|\cdot\,]$ occurring in the L2 case are readily available and are given by

$$\begin{aligned}
[p_{i1}, \ldots, p_{iq}|\cdot\,] &\sim \text{Dirichlet}(\sum_{j=1}^{m_i} z_{ij1} + \exp(a_1' u_i), \ldots, \sum_{j=1}^{m_i} z_{ijq} + \exp(a_q' u_i)) \\
[\pi_{il1}, \ldots, \pi_{ilq}|\cdot\,] &\sim \text{Dirichlet}(\sum_{j=1}^{m_i} z_{ijl} x_{ij1} + \exp(b_{l1}' u_i), \ldots, \sum_{j=1}^{m_i} z_{ijl} x_{ijq} + \exp(b_{lq}' u_i))
\end{aligned}$$

where $i = 1, \ldots, r$, $l = 1, \ldots, q$.

The full conditional densities $[\sigma_a|\cdot\,]$ and $[\sigma_b|\cdot\,]$ are given by

$$\begin{aligned}
[\sigma_a|\cdot\,] &\sim \text{Inverse Gamma}(r_a^{(0)} + q, s_a^{(0)} + \sum_{k=1}^{q} \sum_{t=1}^{v} a_{kt}^2/2) \\
[\sigma_b|\cdot\,] &\sim \text{Inverse Gamma}(r_b^{(0)} + q^2, s_b^{(0)} + \sum_{l=1}^{q} \sum_{k=1}^{q} \sum_{t=1}^{v} b_{lkt}^2/2).
\end{aligned}$$

The full conditional densities $[a_k|\cdot\,]$ and $[b_{lk}|\cdot\,]$ do not have familiar forms, and hence we implement slice sampling on

$$\begin{aligned}
[a_k|\cdot\,] &\propto \exp\Big\{ - \frac{(a_k - a_k^{(0)})'(a_k - a_k^{(0)})}{2\sigma_a^2} \Big\} \prod_{i=1}^{r} \Big[\frac{\Gamma(\sum_{j=1}^{q} \exp(a_j' u_i))}{\Gamma(\exp(a_k' u_i))} p_{ik}^{\exp(a_k' u_i)} \Big] \\
[b_{lk}|\cdot\,] &\propto \exp\Big\{ - \frac{(b_{lk} - b_{lk}^{(0)})'(b_{lk} - b_{lk}^{(0)})}{2\sigma_b^2} \Big\} \prod_{i=1}^{r} \Big[\frac{\Gamma(\sum_{j=1}^{q} \exp(b_{lj}' u_i))}{\Gamma(\exp(b_{lk}' u_i))} \pi_{ilk}^{\exp(b_{lk}' u_i)} \Big]
\end{aligned}$$

where $l = 1, \dots, q$ and $k = 1, \dots, q$.

6.2 Standard error of MCMC estimates

One technical challenge posed by MCMC algorithms is the estimation of standard errors for posterior quantities of interest (Kass et. al. 1998). The samples in a chain are typically autocorrelated. The standard error of posterior quantities of interest will be underestimated without considering the autocorrelation within a chain. We use the effective sample size (ESS) to measure the number of independent samples within a Markov chain.

The ESS of a chain is defined as a function of autocorrelations within the chain at different lags. We let $\theta(1), \dots, \theta(N)$ denote the sequence of N MCMC samples. We let ρ_t denote the autocorrelation at lag t for θ. The effective sample size of N samples generated by a process with autocorrelations ρ_t is defined by

$$N_{\text{eff}} = \frac{N}{\sum_{t=-\infty}^{\infty} \rho_t} = \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho_t}.$$

If we let $\tilde{\theta}$ denote the mean of $\theta(1), \dots, \theta(N)$, and let $\text{sd}(\theta)$ denote the standard deviation of $\theta(1), \dots, \theta(N)$, then the standard error of $\tilde{\theta}$ is estimated as $\text{sd}(\theta)/\sqrt{N_{\text{eff}}}$.

7 REFERENCES

Agresti, A. (2013). *Categorical Data Analysis, Third Edition*, Wiley-Interscience, New York.

Basu, A.P. (1983). “Identifiability”, in *Encyclopedia of Statistical Sciences*, Editors S. Kotz and N.L. Johnson, Wiley Interscience, New York, 4, 2-6.

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

Evans, M., Guttman, I., Haitovsky, Y. and Swartz, T.B. (1996). “Bayesian analysis of binary data subject to misclassification”, In *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner* (Berry, D., Geweke, J. and Chaloner, K.M. editors), 67-77.

Evans, M. and Swartz, T.B. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*, Oxford University Press, Oxford.

Gerlach, R. and Stamey, J. (2007). "Bayesian model selection for logistic regression with misclassification problems", *Statistical Modelling*, 7, 255-273.

Gustafson, P. (2005). "On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables", *Statistical Science*, 20, 111-140.

Gustafson, P. (2010). "Bayesian inference for partially identified models", *International Journal of Biostatistics 6* (issue 2, article 17, 17 pages).

Horowitz, M., Yurko, R. and Ventura, S. (2018). "nflscrapr: Compiling the NFL play-by-play API for easy use in R. *R package version 1.8.1*, https://github.com/maksimhorowitz/nflscrapR.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). "Markov chain Monte Carlo in practice: a roundtable discussion", *The American Statistician*, 52, 93-100.

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

Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2013). *The BUGS Book: A Practical Introduction to Bayesian Analysis*, Chapman & Hall/CRC Texts in Statistical Science Series, Boca Raton.

Martin, E.S. and Gonzalez, J. (2010). "Bayesian identifiability: Contributions to an inconclusive debate", *Chilean Journal of Statistics*, 1, 69-91.

Neal, R. M. (2003). "Slice sampling", *The Annals of Statistics*, 31, 705-767.

Ogburn, E.L. and Vanderweele, T.J. (2013). "Bias attenution results for nondifferentially mismeasured ordinal and coarsened confounders", *Biometrika*, 100, 241-248.

Perez, C.J., Giron, F.J., Martin, J., Ruiz, M. and Rojano, C. (2007). "Misclassified multinomial data: a Bayesian approach", *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales, Serie A: Matematicas*, 101, 71-80.

Plummer, M. (2015). *JAGS Version 4.0 User Manual*, Accessed online September 19, 2019 at http://www.uvm.edu/∼bbeckage/Teaching/DataAnalysis/Manuals/manual.jags.pdf

Swartz, T.B. and Arce, A. (2014). "New insights involving the home team advantage", *International Journal of Sports Science and Coaching*, 9, 681-692.

Swartz, T.B., Haitovsky, Y., Vexler, A. and Yang, T. (2004). "Bayesian identifiability and misclassification in multinomial data", *The Canadian Journal of Statistics*, 32, 285-302.

Wang, D. and Gustafson, P. (2014). "On the impact of misclassification in an ordinal exposure variable", *Epidemiological Methods*, 3, 97-106.