Biometrics 60, 41–49
March 2004

Bayesian Semiparametric Modeling for Matched Case–Control Studies with Multiple Disease States

Samiran Sinha,* Bhramar Mukherjee, and Malay Ghosh*****

Department of Statistics, University of Florida, Gainesville, Florida 32611, U.S.A.
**email:* ssinha@stat.ufl.edu
***email:* mukherjee@stat.ufl.edu
****email:* ghoshm@stat.ufl.edu

Summary. We present a Bayesian approach to analyze matched "case–control" data with multiple disease states. The probability of disease development is described by a multinomial logistic regression model. The exposure distribution depends on the disease state and could vary across strata. In such a model, the number of stratum effect parameters grows in direct proportion to the sample size leading to inconsistent MLEs for the parameters of interest even when one uses a retrospective conditional likelihood. We adopt a semiparametric Bayesian framework instead, assuming a Dirichlet process prior with a mixing normal distribution on the distribution of the stratum effects. We also account for possible missingness in the exposure variable in our model. The actual estimation is carried out through a Markov chain Monte Carlo numerical integration scheme. The proposed methodology is illustrated through simulation and an example of a matched study on low birth weight of newborns (Hosmer, D. A. and Lemeshow, S., 2000, *Applied Logistic Regression*) with two possible disease groups matched with a control group.

Key words: Conditional inference; Dirichlet mixture; Exponential family; Gibbs sampling.

1. Introduction

Case–control studies have received a great deal of attention over the last few decades both from statisticians and epidemiologists. The analysis is based on the comparison of persons having a disease (the cases) with those not having the disease (the controls) and assesses the effect of exposure variables on the probability of developing the disease. For efficient use of data, in such studies, one usually implements a matched design, matching one or more controls with a case on the basis of some prognostic factors such as age, family background, etc. In some such situations it is natural to note that the disease state might have more than one category, i.e., we may have subdivisions within the "cases." For example, for patients diagnosed with cancer, they may have cancer of stage I, stage II, or stage III at the time of the diagnosis which is an example of ordinal disease categories. We may also notice nominal categories for disease states such as patients classified into having one eye or both eyes damaged. To our best knowledge there has been hardly any Bayesian and very little frequentist work for analyzing matched data when one has multiple disease states. Along with considering polytomous disease states, we extend the typical methodology for analyzing case–control data in the following way. In analyzing a matched study one usually considers an appropriate conditional likelihood to eliminate the stratum effect parameters involved in determining the disease probability, while nonspecific exposure distributions are assumed to satisfy Prentice–Pyke (Prentice and Pyke, 1979) type constraints. In contrast, we assume the exposure distribution to be a member of the exponential family and allow for the exposure distribution to vary across strata. In such a situation, even the conditional likelihood involves some nuisance parameters that grow in direct proportion to sample size, giving rise to inconsistency of the conditional maximum likelihood estimates (MLEs). We adopt a semiparametric Bayesian approach to circumvent this problem and estimate the parameters of interest through a numerical computation scheme. We also account for possible missingness in the exposure variable through modeling the distribution of the exposure variable.

The example considered in this article involves a matched case–control data set coming from a low–birth weight study conducted by the Baystate Medical Center in Springfield, Massachusetts. The data set is discussed in Hosmer and Lemeshow (2000, Section 1.6.2) and is used as an illustrative example of analyzing a matched case–control study in Chapter 7 of their book. Low birth weight, defined as birth weight less than 2500 g, is a cause of concern for a newborn as infant mortality and birth defect rates are very high for low–birth weight babies. The data were matched according to the age of the mother. A woman's behavior during pregnancy (smoking habits, diet, prenatal care) can greatly alter the chances of carrying the baby to term. The goal of the study was to determine whether these variables were "risk factors" in the clinical population served by Baystate Medical Center. Using the actual birth weight observations we divided the cases, namely the low–birth weight babies, into two categories, *very*

low (weighing less than 2000 g) and low (weighing between 2000 to 2500 g) and tried to assess the impact of smoking habits of mother on the chance of falling in the two low–birth weight categories. Presence of uterine irritability in mother and mother's weight at last menstruation period were considered as relevant covariates. It was noted that smoking mothers had a higher relative risk of having a low–birth weight child when compared to a nonsmoking mother. However, the risk of having a *very* low–birth weight child did not depend on smoking significantly. This observation could not be made without the classification of the data into multiple low–birth weight groups, illustrating the relevance of such a type of analysis in certain situations.

In many real studies, one often does not have access to exposure information on all the subjects under study. In such situations, rather than ignoring the available partial information, one gains in terms of estimation accuracy of the parameters of interest if the distribution of the missing exposure or the missingness process is stochastically modeled. Although the frequentist literature contains a number of articles for matched case–control studies and treatment of missing covariate information (see, e.g., Paik and Sacco, 2000; Satten and Carroll, 2000; Rathouz, Satten, and Carroll, 2002), unified Bayesian methods addressing these problems are needed. Zelen and Parker (1986), Nurminen and Mutanen (1987), and Ashby, Hutton, and McGee (1993) considered case–control problems in a Bayesian formulation when the risk factor was a binary variable, the stratum effect was a constant, and there were no missing covariates. Müller and Roeder (1997) and Müller et al. (1999) considered a semiparametric model for unmatched case–control problems with continuous and possibly missing covariates, and binary disease status. Seaman and Richardson (2001) extended the Müller–Roeder approach for categorical covariates, and brought out the connection between the Zelen–Parker and Müller–Roeder approaches in absence of measurement error. None of these papers considered a multicategory disease status in a matched case–control set-up.

This article intends to develop an approach to case–control studies with a multicategory variable **D** denoting disease states, a completely observed covariate **Z**, and a vector of exposure variables or risk factors **X** that could potentially contain some missing observations. The exposure variables could be discrete or continuous, and we assume that $\mathbf{X} \mid \mathbf{D}, \mathbf{Z}$ has a distribution coming from an exponential family which may have different natural parameters across strata. We will assume a Dirichlet process with a normal base measure on the distribution of the stratum effects and normal priors on the other regression parameters and estimate all the parameters via a Markov chain Monte Carlo (MCMC) computing scheme. This is a major departure from the usual frequentist as well as the Bayesian approach of assuming that the distribution of exposure variable is not affected by any stratum effect except through the measured covariates. This last assumption may not hold in many situations. For example, matching for cancer patients is often done from their family and smoking is a natural exposure to consider. The distribution of the exposure may depend on genetic traits in the family which may affect the disease distribution in different families in different ways and may not be measurable as a covariate. The present Bayesian approach will allow us to model stratum effects on the exposure distribution for highly stratified data and account for missing exposure information.

The outline of the remaining sections is as follows. In Section 2 we introduce notations and model assumptions. Section 3 contains the conditional likelihood and the priors. In Section 4 we analyze the low–birth weight data and discuss the MCMC computation scheme. Section 5 presents the results from a small simulation study comparing the proposed Bayesian semiparametric method with two possible parametric Bayesian alternatives. Section 6 contains concluding remarks. The Appendix contains details of some calculations and computation scheme.

Before concluding this section we highlight some of the new features of this article. First, ours seems to be the first attempt toward an analysis of matched case–control studies with multiple disease states. Second, the introduction of the semiparametric Bayesian method overcomes the difficulties associated with the conventional frequentist procedure when the number of nuisance parameters grows in direct proportion to the sample size even in the retrospective conditional likelihood. Third, we provide a unified analysis for both discrete and continuous exposure variables and account for possible missingness in exposure observations.

2. Model and Notation

Let D_{ij} denote the disease state of the jth individual in the ith stratum S_i. Suppose that there are $(K+1)$ nominal levels of the disease variable, with $D_{ij} = k$ denoting disease state k, $k = 1, \ldots, K$ and $D_{ij} = 0$ denoting the control group. We assume that there is one case matched with M controls in each stratum and we have n strata in all. For ease of notation, we present our models with a single exposure X, but the model could easily be extended to the case when one has a set of independent multiple exposures, each having a distribution coming from the exponential family. The disease probabilities for each of the K groups are modeled through K logits as in a multinomial logistic regression model:

$$\log \frac{P[D_{ij} = k \mid S_i, \mathbf{Z}_{ij}, X_{ij}]}{P[D_{ij} = 0 \mid S_i, \mathbf{Z}_{ij}, X_{ij}]} = \beta_{0k}(S_i) + \boldsymbol{\beta}_{1k}^{\mathrm{T}} \mathbf{Z}_{ij} + \beta_{2k} X_{ij} \quad \text{for } k = 1, \ldots, K. \tag{1}$$

Each $\boldsymbol{\beta}_{1k}$ is a $p \times 1$ column vector and $\mathbf{Z}_{ij} = (Z_{ij}^{(1)}, \ldots, Z_{ij}^{(p)})^{\mathrm{T}}$ is the vector of p completely observed covariates.

For our example with the low–birth weight data as described in Section 1, we have two disease states with $K = 2$. Then $\exp(\beta_{21})$ signifies the odds of having low–birth weight baby for a mother who smokes relative to one who does not smoke and similarly $\exp(\beta_{22})$ is the risk of a *very* low–birth weight child for a mother who smokes relative to one who does not. The conditional probabilities of the disease variable given the covariate, exposure, and the stratum are given by

$$P(D_{ij} = k \mid S_i, \mathbf{Z}_{ij}, X_{ij}) = \frac{\exp\{\beta_{0k}(S_i) + \boldsymbol{\beta}_{1k}^{\mathrm{T}} \mathbf{Z}_{ij} + \beta_{2k} X_{ij}\}}{1 + \sum_{r=1}^{K} \exp\{\beta_{0r}(S_i) + \boldsymbol{\beta}_{1r}^{\mathrm{T}} \mathbf{Z}_{ij} + \beta_{2r} X_{ij}\}} \quad \text{for } k = 1, \ldots, K \tag{2}$$

and

$$P(D_{ij}=0\,|\,S_i,\mathbf{Z}_{ij},X_{ij}) = \frac{1}{1+\sum_{r=1}^{K}\exp\{\beta_{0r}(S_i)+\boldsymbol{\beta}_{1r}^{\mathrm{T}}\mathbf{Z}_{ij}+\beta_{2r}X_{ij}\}}. \quad (3)$$

To cover both discrete as well as continuous exposures, we assume a general exponential family of distributions for the exposure variable in the control population with respect to some finite dominating measure μ, i.e.,

$$f(X_{ij}\,|\,S_i,\mathbf{Z}_{ij},D_{ij}=0) = \exp[\xi_{ij}\{\theta_{ij}X_{ij}-b(\theta_{ij})\}+c(\xi_{ij},X_{ij})]. \quad (4)$$

The natural parameters are modeled as a regression function of the completely observed covariates, namely, $\theta_{ij}=\gamma_{0i}+\boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{Z}_{ij}$, where $\boldsymbol{\gamma}_1^{\mathrm{T}}=(\gamma_{11},\ldots,\gamma_{1p})$. The dependence of the exposure distribution on the stratum is captured through the varying intercepts γ_{0i}.

3. Likelihood, Priors, and Posteriors

Before writing out the conditional likelihood, we need some technical results stated in the following as lemmas. These lemmas follow by repeating essentially the proofs of Lemmas 1–3 of Sinha et al. (2003). The details are omitted.

LEMMA 1: *Under assumptions* (2)–(4) *the distribution of the exposure variable in a given disease state k, namely* $f(X_{ij}\,|\,S_i,\mathbf{Z}_{ij},D_{ij}=k)$, *is also of general exponential form with scale parameter* ξ_{ij} *and natural parameter* $\theta^*_{ijk}=\theta_{ij}+\xi_{ij}^{-1}\beta_{2k}$ *for* $k=1,\ldots,K$.

LEMMA 2: *Under the same set of assumptions,*

$$\frac{P(D_{ij}=k\,|\,S_i,\mathbf{Z}_{ij})}{P(D_{ij}=0\,|\,S_i,\mathbf{Z}_{ij})} = \exp\{\beta_{0k}(S_i)+\boldsymbol{\beta}_{1k}^{\mathrm{T}}\mathbf{Z}_{ij}\}\times\exp\big[\xi_{ij}\{b(\theta^*_{ijk})-b(\theta_{ij})\}\big]. \quad (5)$$

We need one more lemma to write out the conditional likelihood.

LEMMA 3:

$$\frac{P(D_{is}=k\,|\,S_i,\mathbf{Z}_{is})/P(D_{is}=0\,|\,S_i,\mathbf{Z}_{is})}{\sum_{j=1}^{M+1}P(D_{ij}=k\,|\,S_i,\mathbf{Z}_{ij})/P(D_{ij}=0\,|\,S_i,\mathbf{Z}_{ij})} = \frac{\exp(\boldsymbol{\beta}_{1k}^{\mathrm{T}}\mathbf{Z}_{is})\exp\big[\xi_{is}\{b(\theta^*_{isk})-b(\theta_{is})\}\big]}{\sum_{j=1}^{M+1}\exp(\boldsymbol{\beta}_{1k}^{\mathrm{T}}\mathbf{Z}_{ij})\exp\big[\xi_{is}\{b(\theta^*_{ijk})-b(\theta_{ij})\}\big]} \quad \textit{for } s=1,\ldots,M+1, i=1,\ldots,n. \quad (6)$$

Without loss of generality we may assume that the first subject in each stratum is a case, and if we denote the disease state (type of case) in stratum i as k_i (k_i could assume any of the values $1,\ldots,K$), then the conditional likelihood given that there is one case in each stratum will be of the form

$$\begin{aligned}L_c &\propto \prod_{i=1}^{n}P\Big\{D_{i1}=k_i, D_{ij}=0(j=2,\ldots,M+1), X_{ij}(j=1,\ldots,M+1)\,|\,S_i,\mathbf{Z}_{ij},\sum_{j=1}^{M+1}D_{ij}=k_i\Big\}\\ &= \prod_{i=1}^{n}\Big\{f(X_{i1}\,|\,S_i,\mathbf{Z}_{i1},D_{i1}=k_i)\prod_{j=2}^{M+1}f(X_{ij}\,|\,S_i,\mathbf{Z}_{ij},D_{ij}=0)\Big\}\\ &\quad\times\prod_{i=1}^{n}\frac{P(D_{i1}=k_i\,|\,S_i,\mathbf{Z}_{i1})\prod_{j=2}^{M+1}P(D_{ij}=0\,|\,S_i,\mathbf{Z}_{ij})}{\sum_{l=1}^{M+1}P(D_{il}=k_i\,|\,S_i,\mathbf{Z}_{il})\prod_{j\neq l}^{M+1}P(D_{ij}=0\,|\,S_i,\mathbf{Z}_{ij})}\\ &= \prod_{i=1}^{n}\Big\{f(X_{i1}\,|\,S_i,\mathbf{Z}_{i1},D_{i1}=k_i)\prod_{j=2}^{M+1}f(X_{ij}\,|\,S_i,\mathbf{Z}_{ij},D_{ij}=0)\Big\}\\ &\quad\times\prod_{i=1}^{n}\frac{P(D_{i1}=k_i\,|\,S_i,\mathbf{Z}_{i1})/P(D_{i1}=0\,|\,S_i,\mathbf{Z}_{i1})}{\sum_{j=1}^{M+1}P(D_{ij}=k_i\,|\,S_i,\mathbf{Z}_{ij})/P(D_{ij}=0\,|\,S_i,\mathbf{Z}_{ij})}.\end{aligned} \quad (7)$$

In many situations, one may not have all observations recorded on the exposure variable. For example, in the famous endometrial cancer data set as discussed in Breslow and Day (1980), 16% of the observations are missing on a possible risk factor of obesity. In such situations, a typical conditional frequentist matched analysis loses the entire information on a subject with a single missing exposure. Modeling the exposure distribution in such situations leads to more efficient estimation of parameters of interest as compared to completely ignoring the partial information that is still available (Satten and Kupper, 1993a,b; Satten and Carroll, 2000; Sinha et al., 2003).

We modify the above likelihood appropriately for the situations when the exposure variable may contain some missingness. Let δ_{ij} be an indicator variable for missing exposures defined in the following manner:

$$\delta_{ij}=\begin{cases}1 & \text{if } X_{ij} \text{ is observed,}\\ 0 & \text{if } X_{ij} \text{ is missing,}\\ & i=1,\ldots,n \text{ and } j=1,\ldots,M+1.\end{cases}$$

We also assume that the distribution of δ_{ij} does not involve the parameters $\boldsymbol{\beta}_{1k}$, β_{2k}, $\boldsymbol{\gamma}_1$, and $\boldsymbol{\gamma}_0^{\mathrm{T}}=(\gamma_{01},\ldots,\gamma_{0n})$. The conditional likelihood including missingness of the exposure variable is seen to be

$$\begin{aligned}L_c &\propto \prod_{i=1}^{n}\Big\{f(X_{i1}\,|\,S_i,\mathbf{Z}_{i1},D_{i1}=k_i)^{\delta_{i1}}\times\prod_{j=2}^{M+1}f(X_{ij}\,|\,S_i,\mathbf{Z}_{ij},D_{ij}=0)^{\delta_{ij}}\Big\}\\ &\quad\times\prod_{i=1}^{n}\frac{P(D_{i1}=k_i\,|\,S_i,\mathbf{Z}_{i1})/P(D_{i1}=0\,|\,S_i,\mathbf{Z}_{i1})}{\sum_{j=1}^{M+1}P(D_{ij}=k_i\,|\,S_i,\mathbf{Z}_{ij})/P(D_{ij}=0\,|\,S_i,\mathbf{Z}_{ij})}.\end{aligned} \quad (8)$$

This likelihood involves the parameters $\boldsymbol{\beta}_1^{p\times K} = (\boldsymbol{\beta}_{11},\ldots,\boldsymbol{\beta}_{1K})$, $\boldsymbol{\beta}_2^{\mathrm{T}} = (\beta_{21},\ldots,\beta_{2K})$, $\boldsymbol{\gamma}_1$, and $\boldsymbol{\gamma}_0$.

We consider mutually independent normal priors: $\boldsymbol{\beta}_{1k} \sim N_p(\boldsymbol{\mu}_{\beta_{1k}}, \sigma^2_{\beta_{1k}}\boldsymbol{I}_p)$ for $k = 1,\ldots,K$. $\boldsymbol{\beta}_2 \sim N_K(\boldsymbol{\mu}_{\beta_2}, \sigma^2_{\beta_2}\boldsymbol{I}_K)$ $\boldsymbol{\gamma}_1 \sim N_p(\boldsymbol{\mu}_{\gamma_1}, \sigma^2_{\gamma_1}\boldsymbol{I}_p)$.

One of the key features of our approach is to retain the nuisance parameters γ_{0i} in the model and to put a Dirichlet with normal base measure on them, i.e., $\gamma_{0i} \,|\, G \overset{\text{i.i.d.}}{\sim} G$, where $G \sim DP(\alpha G_0)$ and G_0 is $N(\zeta_0, \sigma_0^2)$. Using the classical result of Antoniak (1974), it follows that

$$\gamma_{0i} \,|\, \gamma_{0k}(k \neq i) \sim \frac{\alpha}{\alpha+n-1} N\left(\zeta_0, \sigma_0^2\right) + \frac{1}{\alpha+n-1}\sum_{k=1,k\neq i}^{n} I_{\gamma_{0k}}(\gamma_{0i}), \qquad (9)$$

where I is the indicator function. The result allows the possibility of equal values of some γ_{0i} as well. As we will notice in our simulations, the fact that it can allow for equal values of γ_{0i} plays an important role in making the procedure robust over a wide spectrum of scenarios, from widely varying γ_{0i}'s to the case when they are all equal.

For our data analysis and simulations, we compared our Bayesian semiparametric (BSP) method with two possible parametric Bayesian alternatives. The first one is a parametric Bayesian analogue of the method proposed by Satten and Carroll (2000) for matched case–control studies with a single disease state. Satten and Carroll (2000) assumed a constant stratum effect on the exposure distribution, i.e., $\gamma_{0i} \equiv \gamma_0$. In the parametric Bayesian analogue of their method (denoted by PBC, C standing for constant stratum effect) in the context of multiple disease states, we consider a normal distribution as a prior on this common stratum effect parameter γ_0 and carry out Bayesian analysis. The other parametric Bayesian alternative (denoted by PBV, V standing for varying stratum effects) allows for possibly varying γ_{0i} and assumes i.i.d. normal prior on each γ_{0i}.

Remark 1. It follows from (9) that for very large values of α the BSP method is equivalent to the PBV method, whereas for very small values of α it amounts to assuming a completely discrete prior on γ_{0i}. In our numerical work we assumed a Gamma prior on α and resampled from the full conditional distribution of α using a latent beta variable as prescribed in Escobar and West (1995).

Remark 2. The entire analysis carries through if each stratum contains varying number of controls, with equations (5)–(9) remaining essentially the same with M replaced by M_i in the ith stratum.

The estimation of the parameters is done by the Markov chain Monte Carlo numerical integration scheme. To generate random numbers from the posterior distributions of the parameters we use a componentwise Metropolis Hastings scheme. We describe the computation scheme along with the analysis of the low–birth weight study in the following section.

4. Example and Computing Scheme

In the previous sections, we discussed the general methodology which we now apply to the matched case–control study for low–birth weight data as described in Section 1. The matched data contain 29 strata, and each stratum has one case and three controls. We denote the low–birth weight and the *very* low–birth weight group as disease states 1 and 2, respectively. One can possibly think of many different models for explaining the disease in terms of the possible covariates recorded in the data set. We consider smoking status of mother as a single exposure variable. Two other covariates, a binary variable denoting presence of uterine irritability (UI) in mother and weight of the mother at last menstrual period (LWT) are also included in the model.

As a starting point, we separated the 29 strata into two groups depending on whether the case belonged to low–birth weight category 1 or 2. We formed two cross-classification tables of birth weight category versus smoking status of mother during pregnancy period for these two separate matched samples and noted that $\widehat{\mathrm{OR}}(1,0) = 3.4$ and $\widehat{\mathrm{OR}}(2,0) = 1.917$, where $\widehat{\mathrm{OR}}(k,0)$ denotes the odds ratio of maternal smoking habits for birth weight group $= k$ versus normal birth weight group $= 0$. The two odds ratios demonstrate that the odds of having a low–birth weight baby for a smoking mother as opposed to a nonsmoking mother are higher in category 1, whereas for category 2 we notice a relatively weaker behavior in the same direction. The difference in the odds ratios in these two tables led us to use these data as a testbed example to illustrate our methods.

For our proposed analysis we have a stochastic distribution on the exposure variable that belongs to the exponential family. The binary variable smoking status is assumed to follow a Bernoulli distribution:

$$f(X_{ij} \,|\, D_{ij} = 0, \mathbf{Z}_{ij}, S_i) = p_{ij}^{X_{ij}}(1-p_{ij})^{1-X_{ij}}; \qquad (10)$$

so here $\xi_{ij} = 1$, $\theta_{ij} = \ln(p_{ij}/(1-p_{ij}))$, $b(\theta_{ij}) = \ln(1+\exp(\theta_{ij}))$, and $c(X_{ij}, \xi_{ij}) = 0$. Also here $K = 2$, $p = 2$, $n = 29$, and $M = 3$. Using Lemmas 1–3 we obtain the conditional likelihood for the whole data:

$$L_c \propto \prod_{i=1}^{n} \Bigg\{ \exp\left\{\theta_{i1}^* X_{i1} - \ln\left(1+\exp\left(\theta_{i1}^*\right)\right)\right\} \times \exp\left[\sum_{j=2}^{M+1}\{\theta_{ij}X_{ij} - \ln(1+\exp(\theta_{ij}))\}\right] \times \frac{\exp\left(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\beta}_{11}Z_{i1}^{(1)} + \boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\beta}_{12}Z_{i1}^{(2)}\right) \times \frac{\left(1+\exp\left(\theta_{i1}^*\right)\right)}{(1+\exp(\theta_{i1}))}}{\sum_{j=1}^{M+1}\exp\left(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\beta}_{11}Z_{ij}^{(1)} + \boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\beta}_{12}Z_{ij}^{(2)}\right) \times \frac{\left(1+\exp\left(\theta_{ij}^*\right)\right)}{(1+\exp(\theta_{ij}))}} \Bigg\}, \qquad (11)$$

Table 1

Analysis of low–birth weight data using full data set. BSP stands for Bayesian semiparametric method, whereas PBC and PBV stand for parametric Bayes methods assuming constant and varying stratum effects, respectively.

		BSP			PBC			PBV		
Logit	Parameter	Mean	SD	HPD region	Mean	SD	HPD region	Mean	SD	HPD
1	SMOKE	1.42	0.60	(0.33, 2.72)	1.26	0.56	(0.25, 2.50)	1.48	0.65	(0.26, 2.08)
	LWT	−0.86	1.39	(−3.78, 1.81)	−1.03	1.35	(−3.58, 1.86)	−0.73	1.36	(−3.40, 2.01)
	UI	0.15	0.67	(−1.27, 1.46)	0.10	0.67	(−1.19, 1.52)	0.18	0.67	(−1.14, 1.52)
2	SMOKE	0.37	0.83	(−1.35, 2.05)	0.23	0.66	(−1.10, 1.54)	0.38	0.85	(−1.30, 2.17)
	LWT	−0.52	1.61	(−3.76, 2.52)	−0.55	1.59	(−3.73, 2.41)	−0.55	1.62	(−3.65, 2.79)
	UI	1.81	0.83	(0.18, 3.51)	1.78	0.83	(0.30, 3.59)	1.81	0.87	(0.27, 3.72)

where $\theta^*_{ij} = \theta_{ij} + \boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{\beta}_2$. Since the value of k (i.e., disease type) is completely determined by knowing the stratum, we omit the subscript k for θ^*_{ijk} in the above expression; $Z^{(1)}_{is}$ and $Z^{(2)}_{is}$ denote the observed value of the two covariates UI and LWT for the sth subject in the ith stratum, respectively, and $\boldsymbol{h}_i$ is defined as $\boldsymbol{h}_i = (h_{i1}, \ldots, h_{iK})^{\mathrm{T}}$ where

$$h_{ir} = \begin{cases} 1 & \text{if } D_{i1} = r \\ 0 & \text{otherwise}, i = 1, \ldots, n \text{ and } r = 1, \ldots, K. \end{cases}$$

Our analysis is based on normal priors centered at zero with large variances for all the regression parameters. In instances (such as ours) when prior elicitation is not possible, these priors usually lead to posteriors relying more heavily on the data and protect against model failures. In many real applications, the practitioner may have a more precise knowledge about the sign and magnitude of the relative risk parameters and can suitably change the prior if necessary. We conducted a sensitivity analysis with several choices of prior parameters. For the regression parameters and the normal base measure of the Dirichlet process, we experimented with normal priors centered at zero and with variances 2, 4, 5, 6, and 9. We used a gamma prior on the concentration parameter α of the Dirichlet process and ran our analysis with both shape and size parameter set at 0.5, 1, 2, 4, 10, 40, 100, and 200, and with many other possible pairs like $G(0.5, 4)$, $G(1, 10)$, $G(10, 40)$, $G(100, 40)$, and $G(200, 10)$. We noted that the ultimate numerical estimates are reasonably stable over a varying range of prior parameters.

The results are reported for independent $N(0, 5)$ prior on each component of $\boldsymbol{\beta}_{11}$, $\boldsymbol{\beta}_{12}$, and $\boldsymbol{\beta}_2$, $N(0, 6)$ as the normal base measure for the Dirichlet process prior, and a Gamma(2, 2) prior for the concentration parameter α.

We used componentwise Metropolis Hastings algorithm to generate random numbers from the full conditionals of the parameters. For generating observations from the posterior distribution of γ_{0i}, $i = 1, \ldots, n$, we used an algorithm proposed by Neal (2000) for simulating observations from posteriors of Dirichlet mixtures for nonconjugate cases. The details of the computation scheme are given in Sinha et al. (2003). The full conditional distributions for the parameters are presented in the Appendix. We ran the chain typically from 7000 to 10,000 iterations and calculated the diagnostic proposed by Gelman and Rubin (1992) as a measure of convergence.

Table 1 contains the posterior means, posterior standard deviations, and 95% HPD credible intervals for the parameters of interest under the proposed Bayesian semiparametric method (BSP) and the parametric Bayes (PBC and PBV) methods as discussed before. In the parametric Bayes methods, we used an $N(0, 6)$ prior on the constant γ_0 (in PBC) and i.i.d. $N(0, 6)$ prior on varying γ_{0i}'s (in PBV). To illustrate the methods in presence of missingness, we deleted 40% of exposure values completely at random (in the sense of Little and Rubin, 1987) and reran the three analyses. The results are presented in Table 2.

The full data analysis indicates that smoking of mother is a significant risk factor for low–birth weight category (category 1) and is not very significant in the *very* low–birth weight

Table 2

Analysis of low–birth weight data after deleting 40% *observations on smoking completely at random. BSP stands for Bayesian semiparametric method, whereas PBC and PBV stand for parametric Bayes methods assuming constant and varying stratum effects, respectively.*

		BSP			PBC			PBV		
Logit	Parameter	Mean	SD	HPD region	Mean	SD	HPD region	Mean	SD	HPD
1	SMOKE	0.86	0.88	(−0.88, 2.55)	0.55	0.78	(−0.84, 2.18)	0.56	0.86	(−0.71, 2.49)
	LWT	−0.92	1.31	(−3.63, 1.51)	−1.03	1.34	(−3.50, 1.83)	−1.01	1.33	(−3.57, 1.75)
	UI	0.19	0.69	(−1.11, 1.49)	0.21	0.67	(−1.10, 1.55)	0.20	0.69	(−1.31, 1.48)
2	SMOKE	0.54	1.04	(−1.46, 2.07)	0.13	0.93	(−1.85, 1.88)	0.12	0.93	(−1.10, 1.54)
	LWT	−0.43	1.62	(−3.98, 2.38)	−0.59	1.63	(−3.75, 2.59)	−0.54	1.54	(−3.62, 2.45)
	UI	1.82	0.86	(0.34, 3.65)	1.80	0.83	(0.24, 3.46)	1.87	0.82	(0.43, 3.63)

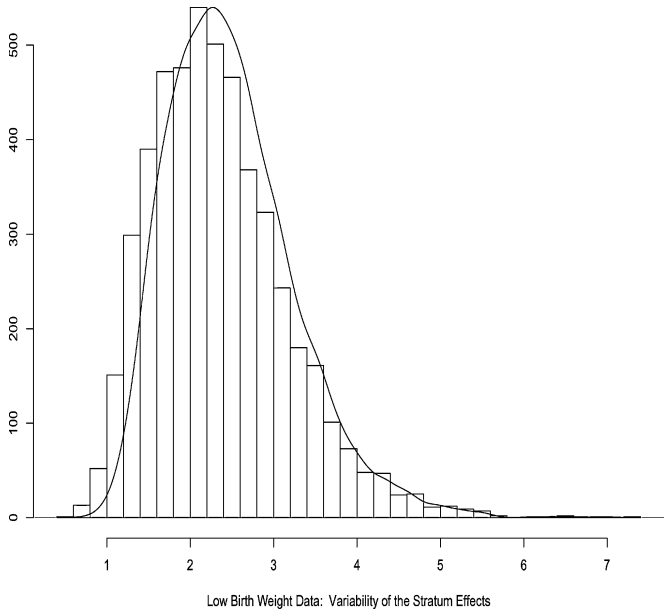

Figure 1. Plot of the variability of γ_{0i}'s for the low–birth weight study example. Estimates of the 29 γ_{0i}'s were collected for each of the last 3000 MCMC samples. Variances of these 29 values were then calculated for each run. The histogram is of these 3000 variance values with a kernel density estimate overlayed on it.

category (category 2). UI, on the other hand, shows an opposite association, showing significance in category 2 and almost no significance in category 1. LWT does not seem to be a significant covariate in any of the categories. The BSP and the PBV methods are in closer agreement, whereas the PBC estimates show some numerical differences.

Figure 1 shows a plot of the variance of the 29 stratum effects in the last 3000 MCMC samples. The average variance is approximately 2.3, showing that there indeed exists variability in the stratum effects. As a result, the BSP and PBV methods that account for this variability are in close agreement, whereas the PBC method assuming constant stratum effect differs numerically from these two methods.

For the analysis with 40% missing observations on smoking, one notices that the estimates corresponding to smoking in the BSP method come closer to their full data counterparts even though the inferences are the same in all three methods. As one might expect, with 40% missingness, the parameter estimates for smoking lose precision and the effect of smoking now appears to be not significant in both categories 1 and 2. Inferences on the other two covariates remain essentially unchanged when compared to full data inferences.

We also analyzed the data after collapsing categories 1 and 2 into only one category (birth weight less than 2500 g). We carried out the Bayesian analysis for a simple matched case–control data (Sinha et al., 2003) and the usual conditional logistic regression (CLR) analysis (Breslow and Day, 1980). Table 3 shows that both BSP and PBV methods assuming varying stratum effects bring out the effect of mother's smoking on having low–birth weight newborns and produce very similar results. The PBC and the CLR methods, assuming constant stratum effect, are in closer agreement with each other and they do demonstrate the effect of smoking but not as precisely as the other two methods that allow varying stratum effects. Figure 1 again demonstrates the differences in results between these two classes of models. Obviously, without the finer classification into two weight categories, the fact that smoking is not so significant for category 2 and UI is appreciably significant for category 2 cannot be concluded from looking at the overall analysis. Thus, the multicategory analysis may render some useful additional information to the practitioner.

5. Simulation Study

In the low–birth weight data we noticed appreciable variability in the stratum effects. In practice, the experimenter may not have a prior idea about the nature of variability among stratum effects and there could be situations where the standard model assumption of constant stratum effect, i.e., $\gamma_{0i} \equiv \gamma_0$, hold. Sinha et al. (2003) contain an example from equine epidemiology where the stratum effects have very small variability. We conducted a simulation study to ascertain the robustness of the BSP method even when variability in the stratum effects is negligible.

In order to simulate a realistic data set for comparing the BSP, PBC, and PBV methods, we decided to use the low–birth weight data themselves. We generated a hypothetical 1:1 matched data set with 50 strata with one binary exposure variable (corresponding to X, smoking status of mother), and one binary covariate (corresponding to Z, presence of uterine irritability). The true values for β_{11}, β_{12}, β_{21}, and β_{22} are chosen to be 0.20, 1.80, 1.40, and 0.4, respectively, close to the estimates obtained by analyzing the low–birth weight data by the BSP method as presented in Table 1.

In order to elicit values for γ_1, the coefficient of Z in the natural parameter of the exposure distribution, namely θ_{ij} on Z, we ran a logistic regression of X(SMOKE) on Z(UI)

Table 3

Analysis of low–birth weight data after collapsing categories 1 *and* 2 *into a single low–birth weight category* (*less than* 2500 g). *BSP stands for Bayesian semiparametric method, whereas PBC and PBV stand for parametric Bayes methods assuming constant and varying stratum effects, respectively. CLR stands for usual conditional logistic regression for analyzing matched data.*

	BSP			PBC			PBV			CLR	
Parameter	Mean	SD	HPD region	Mean	SD	HPD region	Mean	SD	HPD region	Estimate	SE
SMOKE	1.106	0.52	(0.15, 2.17)	0.89	0.45	(−0.03, 1.78)	1.105	0.54	(0.02, 2.25)	0.86	0.45
LWT	−1.07	1.14	(−3.29, 1.17)	−1.11	1.14	(−3.28, 1.12)	−0.98	1.15	(−3.25, 1.30)	−1.13	1.36
UI	0.85	0.51	(−0.20, 1.78)	0.84	0.50	(−0.03, 2.03)	0.87	0.49	(−0.13, 1.82)	0.85	0.51

using the control sample and the fitted model turned out to be $\text{logit}(\Pr(X_{ij})) = -0.734 + 0.579Z_{ij}$. Accordingly, in the simulation study, we used $\gamma_1 = 0.60$.

We followed the structure of our models as described before for simulating Z, X, and the trinomial variable D, indicating the birth weight category. For the low–birth weight data, occurrence of uterine irritability was reported in approximately 20% of the patients. Thus, the completely observed covariate Z was generated first as a Bernoulli variable with success probability 0.20. Second, we generated the trinomial disease variable D. For the ith stratum, one should note that

$$\Pr(D_{i1} = 0 \mid Z_{i1}, Z_{i2}, D_{i1} + D_{i2} = 1 \text{ or } 2, S_i) = \frac{1}{1 + \dfrac{Q_{i1}}{Q_{i2}}},$$

where, for $j = 1, 2$,

$$Q_{ij} = \exp(\beta_{11}Z_{ij})\frac{1 + \exp(\theta_{ij} + \beta_{21})}{1 + \exp(\theta_{ij})} + \exp(\beta_{12}Z_{ij})\frac{1 + \exp(\theta_{ij} + \beta_{22})}{1 + \exp(\theta_{ij})}. \tag{12}$$

We generated a Bernoulli random variable with the above success probability and if this variable assumed a value 1, the simulated value for D_{i1} was taken to be 0, implying that the first subject in the ith stratum is a member of the control population. Let, for $k = 1, 2$,

$$p^*_{ik} = \frac{1}{1 + \exp[z_{ik}(\beta_{12} - \beta_{11})]\dfrac{1 + \exp(\theta_{ik} + \beta_{22})}{1 + \exp(\theta_{ik} + \beta_{21})}}. \tag{13}$$

If $D_{i1} = 0$, we determine D_{i2} according to a Bernoulli draw with success probability p^*_{i2}; set $D_{i2} = 1$ if it results in a success, otherwise set $D_{i2} = 2$. If $D_{i1} \neq 0$, set $D_{i2} = 0$ and we determine the value D_{i1} according to a Bernoulli draw with success probability p^*_{i1}; if this results in a success $D_{i1} = 1$, otherwise $D_{i1} = 2$.

Conditional on the value of D, we proceed to simulate the exposure X. If $D_{ij} = 0$, we generated a binary exposure X_{ij} with success probability given by $\text{logit}(p_{ij}) = \gamma_{0i} + \gamma_1 Z_{ij}$. If $D_{ij} = k$, we generate the binary exposure variable with success probability, $\text{logit}(p_{ij}) = \gamma_{0i} + \gamma_1 Z_{ij} + \beta_{2k}$, $j, k = 1, 2$.

We performed two sets of simulations, one with a constant value of γ_{0i}, namely -1.00, the other with a relatively varying set of γ_{0i}'s simulated from a normal distribution with mean -0.5 and standard deviation 1.5. We assumed $N(0, 5)$ prior on all the relative risk parameters and a Gamma(2, 2) prior for α. In all our simulations, we used identical parameters for the normal distribution which is assumed to be prior on γ_0 in PBC and the i.i.d. prior on γ_{0i} in PBV and also as the mixing distribution in the Dirichlet process prior for BSP ($N(0, 6)$ in this case). We replicated the simulation 50 times, generating 50 different data sets, and obtained the parameter estimates by above-mentioned methods and computed their average and MSE. For each replication we also generated data with 30% exposure values missing completely at random and recalculated all the estimates.

The simulation results presented in Tables 4 and 5 illustrate that for constant stratum effect, the three methods are comparable with PBV estimates being furthest from the true

Table 4

Results of the simulation study. Here "Mean" is the simulated mean, while MSE is the mean squared error ×1000. The true parameter values are $\beta_{11} = 0.20$, $\beta_{12} = 1.80$, $\beta_{21} = 1.40$, $\beta_{22} = 0.40$, and $\gamma_1 = 0.60$. BSP stands for Bayesian semiparametric method, whereas PBC and PBV stand for parametric Bayes methods assuming constant and varying stratum effects, respectively.

Method	β_{11}	β_{12}	β_{21}	β_{22}	γ_1
Full data, fixed $\gamma_{0i} \equiv -1.00$					
BSP					
Mean	0.23	1.88	1.44	0.40	0.43
MSE	0.22	6.57	12.50	6.55	36.54
PBC					
Mean	0.19	1.92	1.43	0.43	0.51
MSE	0.57	14.59	14.42	10.41	20.26
PBV					
Mean	0.20	1.88	1.43	0.32	0.41
MSE	0.53	10.56	9.71	15.50	48.36
30% missing data, fixed $\gamma_{0i} \equiv -1.00$					
Bayes semiparametric					
Mean	0.20	1.90	1.45	0.36	0.44
MSE	0.17	31.78	50.07	37.83	81.31
Bayes parametric					
Mean	0.20	1.89	1.47	0.46	0.47
MSE	4.79	10.05	22.09	11.33	23.98
i.i.d. parametric					
Mean	0.20	1.90	1.35	0.28	0.38
MSE	0.56	9.89	57.05	67.88	83.19

Table 5

Results of the simulation study. Here "Mean" is the simulated mean, while MSE is the mean squared error ×1000. The true parameter values are $\beta_{11} = 0.20$, $\beta_{12} = 1.80$, $\beta_{21} = 1.40$, $\beta_{22} = 0.40$, and $\gamma_1 = 0.60$. BSP stands for Bayesian semiparametric method, whereas PBC and PBV stand for parametric Bayes methods assuming constant and varying stratum effects, respectively.

Method	β_{11}	β_{12}	β_{21}	β_{22}	γ_1
Full data, varying $\gamma_{0i} \sim N(-0.5, 2.25)$					
BSP					
Mean	0.19	1.89	1.43	0.42	0.50
MSE	0.94	8.20	8.58	10.56	13.47
PBC					
Mean	0.20	1.89	1.52	0.48	0.53
MSE	0.70	8.48	27.98	13.36	97.75
PBV					
Mean	0.19	1.92	1.44	0.43	0.51
MSE	0.65	12.84	11.15	9.07	13.25
30% missing data, varying $\gamma_{0i} \sim N(-0.5, 2.25)$					
BSP					
Mean	0.20	1.91	1.50	0.44	0.50
MSE	0.59	11.86	16.96	11.58	19.56
PBC					
Mean	0.20	1.92	1.54	0.49	0.53
MSE	0.46	14.04	30.28	14.68	15.54
PBV					
Mean	0.20	1.94	1.49	0.44	0.50
MSE	0.76	14.31	16.15	14.42	19.35

parameters of interest, whereas for varying stratum effect the BSP and PBV methods have a clear edge over the PBC method. Overall, BSP seems to be the more robust choice as at the onset of a study one is not sure about the nature of variability in the stratum effects.

6. Conclusion

In this article, we proposed a semiparametric Bayesian method to analyze matched case–control data with more than one disease state and illustrate the methods with a real example. The simulation results indicate that in presence of stratum variability and missing data, the Bayesian semiparametric method is superior to the parametric Bayesian alternatives. All three methods perform comparably with constant stratum effects.

Our proposed model considers a nondeterministic exposure variable having a probability distribution belonging to the exponential family. Moreover, the distribution of the exposure could be different in different strata. Our model takes into account both discrete and continuous exposure along with possible missingness in the exposure variable. The growing number of stratum effect parameters are modeled in a semiparametric Bayesian way to overcome the inconsistency problems arising out of classical analysis of such matched data. The computations involving a Dirichlet process prior with a normal base measure are done through a suitable MCMC scheme and enable us to obtain estimates of the parameters of interest. The method could be extended to multiple exposures having an underlying association pattern as well. The general framework is extremely flexible for being used in unorthodox data situations involving missingness and measurement error as well as incorporating widely different types of exposure variables that one may come across in practice.

Acknowledgements

The research of Drs. Sinha and Ghosh was partially supported by NIH grant R01-85414. Dr. Mukherjee's research was partially supported by the New Researchers' scholarship sponsored through Stanford University. The software needed to carry out the data analysis is available at `http://stat.ufl.edu/~mukherjee`. We are grateful to the associate editor for his/her valuable comments.

Résumé

Nous présentons une approche bayésienne pour l'analyse des données "cas-témoins" appariées avec plusieurs états d'une maladie. La probabilité de développement de la maladie est décrite par un modèle de régression logistique multinomiale. La distribution de la variable d'exposition dépend de l'état de la maladie, et peut varier entre strates. Dans ce modèle, le nombre de paramètres d'effets de la strate croît proportionnellement à l'effectif de l'échantillon, ce qui rend inconsistants les estimateurs du maximum de vraisemblance, même quand on utilise une vraisemblance conditionnelle rétrospective. Nous adoptons donc un contexte semi-paramétrique bayésien, en supposant un processus de Dirichlet pour la distribution *a priori*, et un mélange de distributions normales pour la distribution des effets de l'état. Nous tenons aussi compte dans ce modèle d'éventuelles valeurs manquantes pour la variable d'exposition. L'estimation proprement dite se fait par intégration numérique de type Monte Carlo sur chaîne de Markov. La méthodologie proposée est illustrée par des simulations et par l'application à des données appariées de nouveaux-nés de petit poids, où les groupes correspondants à deux maladies possibles sont appariés à un groupe témoin.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to non-parametric problems. *Annals of Statistics* **2,** 1152–1174.

Ashby, D., Hutton, J. L., and McGee, M. A. (1993). Simple Bayesian analyses for case–controlled studies in cancer epidemiology. *Statistician* **42,** 385–389.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research,* Volume 1. Lyon: International Agency for Research on Cancer.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90,** 577–588.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7,** 457–472 (Disc: 483–501, 503–511).

Hosmer, D. A. and Lemeshow, S. (2000). *Applied Logistic Regression,* 2nd edition. New York: Wiley.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case–control studies with errors in variables. *Biometrika* **84,** 523–537.

Müller, P., Parmigiani, G., Schildkraut, J., and Tardella, L. (1999). A Bayesian hierarchical approach for combining case–control and prospective studies. *Biometrics* **55,** 858–866.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9,** 249–265.

Nurminen, M. and Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics* **14,** 67–77.

Paik, M. C. and Sacco, R. (2000). Matched case–control data analyses with missing covariates. *Applied Statistics* **49,** 145–156.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika* **66,** 403–411.

Rathouz, P. J., Satten, G. A., and Carroll, R. J. (2002). Semiparametric inference in matched case–control studies with missing covariate data. *Biometrika* **89,** 905–916.

Satten, G. A. and Carroll, R. J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics* **56,** 384–388.

Satten, G. A. and Kupper, L. (1993a). Conditional regression analysis of the odds ratio between two binary variables when one is not measured with certainty: A method for epidemiologic studies. *Biometrics* **49,** 429–440.

Satten, G. A. and Kupper, L. (1993b). Inferences about exposure–disease associations using probability-of-exposure information. *Journal of the American Statistical Association* **88,** 200–208.

Seaman, S. R. and Richardson, S. (2001). Bayesian analysis of case–control studies with categorical covariates. *Biometrika* **88,** 1073–1088.

Sinha, S., Mukherjee, B., Mallick, B. K., Ghosh, M., and Carroll, R. (2003). Semiparametric Bayesian analysis of matched case–control studies with missing exposure. Preprint.

Zelen, M. and Parker, R. A. (1986). Case–control studies and Bayesian inference. *Statistics in Medicine* **5,** 261–269.

Received May 2003. *Revised September* 2003.
Accepted September 2003.

Appendix

Full Conditional Distributions for the Parameters

The following are the forms of the full conditional distributions of the parameters:

As stated before, here we have assumed the following set priors $\boldsymbol{\beta}_{1k} \sim N_2(\boldsymbol{\mu}_{\beta_{1k}}, \sigma^2_{\beta_{1k}}\boldsymbol{I})$, for $k = 1, 2$, $\boldsymbol{\beta}_2 \sim N_2(\boldsymbol{\mu}_{\beta_2}, \sigma^2_{\beta_2}\boldsymbol{I})$ and $\boldsymbol{\gamma}_1 \sim N_2(\boldsymbol{\mu}_{\gamma_1}, \sigma^2_{\gamma_1}\boldsymbol{I})$:

$$\pi(\boldsymbol{\beta}_{1k} \,|\, \cdot) \propto \frac{\exp\left[-\frac{1}{2\sigma^2_{\beta_{1k}}} \left(\boldsymbol{\beta}_{1k} - \boldsymbol{\mu}_{\beta_{1k}} - \sigma^2_{\beta_{1k}} \sum_{i=1}^{n} \boldsymbol{h}_i Z^{(k)}_{i1} \right)^{\mathrm{T}} \left(\boldsymbol{\beta}_{1k} - \boldsymbol{\mu}_{\beta_{1k}} - \sigma^2_{\beta_{1k}} \sum_{i=1}^{n} \boldsymbol{h}_i Z^{(k)}_{i1} \right) \right]}{\prod_{i=1}^{n} \left\{ \sum_{j=1}^{M+1} \exp\left(\boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{\beta}_{11} Z^{(1)}_{ij} + \boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{\beta}_{12} Z^{(2)}_{ij} \right) \times \frac{\left(1 + \exp\left(\theta^{\star}_{ij}\right)\right)}{\left(1 + \exp\left(\theta_{ij}\right)\right)} \right\}}, \tag{A.1}$$

for $k = 1, 2$

$$\pi(\boldsymbol{\beta}_2 \,|\, \cdot) \propto \frac{\exp\left[-\frac{1}{2\sigma^2_{\beta_2}} \left(\boldsymbol{\beta}_2 - \boldsymbol{\mu}_{\beta_2} - \sigma^2_{\beta_2} \sum_{i=1}^{n} \delta_{i1} \boldsymbol{h}_i X_{i1} \right)^{\mathrm{T}} \left(\boldsymbol{\beta}_2 - \boldsymbol{\mu}_{\beta_2} - \sigma^2_{\beta_2} \sum_{i=1}^{n} \delta_{i1} \boldsymbol{h}_i X_{i1} \right) \right]}{\prod_{i=1}^{n} \left\{ \sum_{j=1}^{M+1} \exp\left(\boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{\beta}_{11} Z^{(1)}_{ij} + \boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{\beta}_{12} Z^{(2)}_{ij} \right) \times \frac{\left(1 + \exp\left(\theta^{\star}_{ij}\right)\right)}{\left(1 + \exp\left(\theta_{ij}\right)\right)} \right\}} \times \prod_{i=1}^{n} \left\{ 1 + \exp\left(\theta^{*}_{i1}\right) \right\}^{1-\delta_{i1}}, \tag{A.2}$$

$$\pi(\boldsymbol{\gamma}_1 \,|\, \cdot) \propto \frac{\exp\left[-\frac{1}{\sigma^2_{\gamma_1}} \left(\boldsymbol{\gamma}_1 - \boldsymbol{\mu}_{\gamma_1} - \sigma^2_{\gamma_1} \sum_{i=1}^{n} \sum_{j=1}^{M+1} \delta_{ij} \mathbf{Z}_{ij} X_{ij} \right)^{\mathrm{T}} \left(\boldsymbol{\gamma}_1 - \boldsymbol{\mu}_{\gamma_1} - \sigma^2_{\gamma_1} \sum_{i=1}^{n} \sum_{j=1}^{M+1} \delta_{ij} \mathbf{Z}_{ij} X_{ij} \right) \right]}{\prod_{i=1}^{n} \left\{ \sum_{j=1}^{M+1} \exp\left(\boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{\beta}_{11} Z^{(1)}_{ij} + \boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{\beta}_{12} Z^{(2)}_{ij} \right) \times \frac{\left(1 + \exp\left(\theta^{\star}_{ij}\right)\right)}{\left(1 + \exp\left(\theta_{ij}\right)\right)} \right\}} \times \prod_{i=1}^{n} \left\{ \frac{\left(1 + \exp\left(\theta^{*}_{i1}\right)\right)^{1-\delta_{i1}}}{\left(1 + \exp\left(\theta_{i1}\right)\right)} \prod_{j=2}^{M+1} \left(1 + \exp\left(\theta_{ij}\right)\right)^{-\delta_{ij}} \right\}, \tag{A.3}$$

$$\pi(\gamma_{0i} \,|\, \cdot) \propto \sum_{k \neq i} \frac{1}{\alpha + n - 1} w_k(\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \gamma_{0s}, s \neq i) I(\gamma_{0k}) L_c + w_i(\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \gamma_{0s}, s \neq i) f_i(\gamma_{0i} \,|\, \cdot), \tag{A.4}$$

where,

$$f_i(\gamma_{0i} \,|\, \cdot) \propto \frac{\exp\left[-\frac{1}{2\sigma_0^2} \left(\gamma_{0i} - \xi_0 - \sigma_0^2 \sum_{i=1}^{n} \sum_{j=1}^{M+1} \delta_{ij} X_{ij} \right)^2 \right]}{\prod_{i=1}^{n} \left\{ \sum_{j=1}^{M+1} \exp\left(\boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{\beta}_{11} Z^{(1)}_{ij} + \boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{\beta}_{12} Z^{(2)}_{ij} \right) \times \frac{\left(1 + \exp\left(\theta^{\star}_{ij}\right)\right)}{\left(1 + \exp\left(\theta_{ij}\right)\right)} \right\}} \times \prod_{i=1}^{n} \left\{ \frac{\left(1 + \exp\left(\theta^{*}_{i1}\right)\right)^{1-\delta_{i1}}}{\left(1 + \exp\left(\theta_{i1}\right)\right)} \prod_{j=2}^{M+1} \left(1 + \exp\left(\theta_{ij}\right)\right)^{-\delta_{ij}} \right\} \tag{A.5}$$

and w_r, $r = 1, \ldots, n$ are weight functions such that $\pi(\gamma_{0i} \,|\, \cdot)$ is a proper density. One may note that if X is completely observed $\delta_{ij} = 1 \; \forall i, j$.

Remark. For resampling from the full conditional distribution of α we followed the algorithm suggested by Escobar and West (1995). At each step we counted the number of distinct γ_{0i}, say k, and conditional on the current values of α and k, we simulated a latent beta random variable say, $\eta \sim B(\alpha + 1, n)$. Let $G(a, b)$ denote the prior on α. Using the current value of k and η, α was simulated from the following mixture of Gamma distribution,

$$\pi(\alpha \,|\, \eta, k) = pG(a + k, b - \log(\eta)) + (1 - p)G(a + k - 1, b - \log(\eta)),$$

where $p = (a + k - 1)/[a + k - 1 + n\{b - \log(\eta)\}]$.