Biometrics 72, 184–192
March 2016

DOI: 10.1111/biom.12411

Nonparametric Bayes Modeling for Case Control Studies with Many Predictors

Jing Zhou,[1,*] Amy H. Herring,[1,2] Anirban Bhattacharya,[3] Andrew F. Olshan,[2,4] David B. Dunson,[5] and The National Birth Defects Prevention Study

[1]Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.
[2]Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.
[3]Department of Statistics,Texas A&M University, College Station, Texas 77843, U.S.A.
[4]Department of Epidemiology, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.
[5]Departments of Statistical Science, Electrical and Computer Engineering, and Mathematics, Duke University, Durham, North Carolina 27708, U.S.A.
email: amanistat@gmail.com

Summary. It is common in biomedical research to run case-control studies involving high-dimensional predictors, with the main goal being detection of the sparse subset of predictors having a significant association with disease. Usual analyses rely on independent screening, considering each predictor one at a time, or in some cases on logistic regression assuming no interactions. We propose a fundamentally different approach based on a nonparametric Bayesian low rank tensor factorization model for the retrospective likelihood. Our model allows a very flexible structure in characterizing the distribution of multivariate variables as unknown and without any linear assumptions as in logistic regression. Predictors are excluded only if they have no impact on disease risk, either directly or through interactions with other predictors. Hence, we obtain an omnibus approach for screening for important predictors. Computation relies on an efficient Gibbs sampler. The methods are shown to have high power and low false discovery rates in simulation studies, and we consider an application to an epidemiology study of birth defects.

Key words: Bayesian nonparametrics; Big data; Epidemiology; Retrospective likelihood; Sparse parallel factor analysis model; Tensor factorization.

1. Introduction

Retrospective case-control studies are common in epidemiologic research because they are much more cost effective than prospective studies, particularly for rare diseases. However, retrospective studies only model exposure given disease, presenting some challenges in analysis and interpretation of the results. In prospective studies, logistic models are widely used to estimate adjusted odds ratios for each of multiple risk factors. A primary concern when analyzing case-control data is whether prospective inferences can be made. In the frequentist framework, there is a rich literature (Anderson, 1972; Prentice and Pyke, 1979) demonstrating that one can ignore the study design and use estimation and inference based on a logistic regression. That is, it has been shown that odds ratios for prospective and case-control data are equivalent.

Consider the National Birth Defects Prevention Study (NBDPS), the largest case-control study ever conducted in the United States on the etiology of birth defects (Yoon et al., 2001). Data are collected on many different defects along with hundreds of potentially associated factors, including environmental, behavioral, biomedical, and occupational variables. Typically, these variables are categorized because categorization of continuous or ordered exposure variables is widespread in epidemiologic research. This is often done to facilitate simple interpretation of exposure summaries and is valid to the extent that risk is homogeneous within categories and potentially heterogeneous across categories (Rothman et al., 2012). These categorized variables in NBDPS lead to a huge sparse contingency table having mostly zero counts. There is strong prior reason to suspect interactions. Although there is a recent Bayesian literature on analysis of high-dimensional contingency tables (Dunson and Xing (2009), Bhattacharya and Dunson (2011), Zhou et al. (2014)), these methods view the data as multivariate categorical arising from a prospective design. Our focus is on addressing the question of whether we can adapt these approaches to case-control settings.

There is a rich literature on Bayesian analysis of case-control data in low-dimensional settings. Mukherjee et al. (2005) provided an overview of Bayesian methods for case-control studies. Zelen and Parker (1986), Nurminen and Mutanen (1987), Marshall (1988), and Ashby et al. (1993) all considered identical Bayesian formulations of a case-control

model with a binary exposure X. Let ϕ and γ be the probabilities of exposure in control and case populations, respectively. The retrospective likelihood is

$$l(\phi, \gamma) \propto \phi^{n_{01}}(1-\phi)^{n_{00}}\gamma^{n_{11}}(1-\gamma)^{n_{10}}, \quad (1)$$

where n_{01} and n_{00} are the number of exposed and unexposed observations in the control population, whereas n_{11} and n_{10} denote the same for the case population. Independent conjugate prior distributions for ϕ and γ are chosen as $Beta(u_1, u_2)$ and $Beta(\nu_1, \nu_2)$, respectively. After reparametrization, one obtains the posterior distribution of the log odds ratio parameter, $\beta = \log\{\gamma(1-\phi)/\phi(1-\gamma)\}$ as

$$l(\beta|n_{11}, n_{10}, n_{01}, n_{00}) \propto \exp\{(n_{11}+\nu_1)\beta\} \times \int_0^1 \frac{\phi^{n_{11}+n_{01}+\nu_1+u_2-1}(1-\phi)^{n_{10}+n_{00}+\nu_2+u_1-1}}{\{1-\phi+\phi\exp(\beta)\}^{n_{11}+n_{10}+\nu_1+\nu_2}}d\phi. \quad (2)$$

The above references used different methods to approximate the posterior distribution of β shown in (2) as well as discussing different prior elicitations based on historical studies.

An alternative is to induce a retrospective likelihood by starting with a model for the prospective likelihood and using Bayes rule. For each subject i, let y_i be a binary response observed together with covariates X_i. Assume a binary response logistic regression for the conditional likelihood of y_i given covariates, with β the coefficients, and let θ denote parameters in a model for the marginal distribution of X_i. Assuming X_i is continuous, Müller and Roeder (1997) proposed a semiparametric Bayes approach. They factorized the joint posterior as

$$\Pr\left(\beta, \theta|\mathbf{X}, \mathbf{Y}\right) \propto \Pr(\beta, \theta)\prod_{i=1}^{n}\Pr(X_i|y_i, \beta, \theta), \quad (3)$$

where under conditional independence assumptions, they let

$$\Pr(X_i|y_i, \beta, \theta) = \frac{\Pr(y_i|X_i, \beta)\Pr(X_i|\theta)}{\Pr(y_i|\beta, \theta)}. \quad (4)$$

Problems arise in approximating the denominator in (4), as this involves an analytically intractable high-dimensional integral.

Seaman and Richardson (2001) extended these two types of models by allowing more than one categorical exposure variable and employing Markov chain Monte Carlo methods to sample the posterior of β. Müller et al. (1999) modeled the retrospective likelihood directly for continuous exposures, also allowing binary covariates via a probit model. Ghosh and Chen (2002), Sinha et al. (2004), and Sinha et al. (2005) developed general Bayesian methods for matched case-control studies in the presence of one or more exposure variables, missing exposures, and multiple disease states. None of the above approaches can accommodate more than a modest number of categorical predictors. As the number of covariates increases, the algorithms either fail to implement or have highly biased estimates.

There has also been research establishing the equivalence of prospective and retrospective Bayesian models. Seaman and Richardson (2004) obtained equivalence through carefully chosen priors. Staicu (2010) extended the class of priors, while still relying on logistic regression. As motivated above, logistic models are too inflexible for our motivating application. Byrne and Dawid (2013) established an equivalence of Bayesian learning of odds ratios via retrospective or prospective likelihoods. However, their result relies on a number of conditions on the model and priors, with the method being impractical for large numbers of covariates.

With this motivation, we develop a nonparametric Bayes method based on directly modeling the retrospective likelihood; this involves novel extensions of recent tensor factorizations for high-dimensional categorical data. The basic framework is proposed in Section 2. Section 3 compares performance with competitors in simulation studies. Section 4 analyzes data from the motivating birth defect study, and Section 5 contains a discussion.

2. Conditional Sparse Parallel Factor Analysis Model

2.1. *Model and Prior*

The general form of the retrospective likelihood is as follows:

$$l(\theta_1, \theta_0) = \prod_{i:y_i=1}\Pr(x_i|y_i=1, \theta_1)\prod_{i:y_i=0}\Pr(x_i|y_i=0, \theta_0), \quad (5)$$

where $\Pr(x_i|y_i = y, \theta_y)$ is the conditional likelihood of the high-dimensional categorical predictors $x_i = (x_{i1}, \ldots, x_{ip})'$, with $x_{ij} \in \{1, \ldots, d_j\}$ for $j = 1, \ldots, p$, given disease status y (0 = control, 1=case). When p is moderate to large (say in the dozens to 100s or more), problems arise in defining a *flexible* model for these high-dimensional categorical predictors. Potentially log-linear models can be used, but unless the vast majority of the interactions are discarded a priori, one obtains an unmanageably enormous number of terms to estimate, store, and process. These bottlenecks are freed by the use of Bayesian low rank tensor factorizations, which have had promising performance in practice ((Dunson and Xing, 2009); (Bhattacharya and Dunson, 2011); (Kunihama and Dunson, 2013); (Zhou et al., 2014)). Johndrow, Bhattacharya, and Dunson (2014) recently showed that a large subclass of sparse log-linear models have low rank tensor factorizations, providing support for the use of tensor factorizations as a computationally convenient alternative.

The tensor factorization methods discussed above are conceptually related to latent structure analysis (Lazarsfeld and Henry, 1968), where the joint distributions of two or more categorical variables are assumed to be conditionally independent given one (or more) latent membership index. For example, if we have two categorical covariates, we can model their joint probability distribution given the disease outcome y as

$$\begin{aligned}\Pr(x_{i1}=c_1, x_{i2}=c_2|y_i=y) &= \sum_{h=1}^{k}\Pr(z_i=h|y_i=y)\prod_{j=1}^{2}\Pr(x_{ij}=c_j|z_i=h, y_i=y)\\ &= \sum_{h=1}^{k}\nu_{yh}\psi^{(1)}_{yhc_1}\psi^{(2)}_{yhc_2}. \quad (6)\end{aligned}$$

With the introduction of the latent class z_i for subjects in outcome group y, covariates x_{i1} and x_{i2} are assumed to be conditionally independent. Marginalizing out the latent index z_i produces a mixture of product multinomial distributions for $x_i = (x_{i1}, x_{i2})'$ characterizing the dependence structure in outcome group y. Any joint probability of x_i for all subjects in each group y can always be decomposed as in (6) for some sufficiently big k (Dunson and Xing, 2009). The extension to the multivariate covariate case is straightforward. A nonparametric Bayes approach can be used to deal with uncertainty in k.

We propose a careful modification of (6) that allows borrowing of information across different disease status, thereby greatly reducing the effective number of parameters and allowing sharper estimation of the features of the distribution of x_i that change with the disease status y_i. Our proposed formulation expresses the joint p.m.f. of x_i conditional on the disease status y_i as

$$\Pr(x_{i1} = c_1, \ldots, x_{ip} = c_p | y_i = y) = \sum_{h=1}^{k} \nu_{yh} \prod_{j=1}^{p} \psi^{(j)}_{yhc_j}, \quad (7)$$

with

$$\psi^{(j)}_{yhc_j} = \begin{cases} \lambda^{(j)}_{yhc_j} = \Pr(x_{ij} = c_j | y_i = y, z_i = h), & \text{if} \quad j \in S_{yh} \\ \lambda^{(j)}_{0c_j} = \Pr(x_{ij} = c_j), & \text{if} \quad j \in S^c_{yh} \end{cases}, \quad (8)$$

where in (7), $\nu_{yh} = \Pr(z_i = h | y_i = y)$ is a mixture probability for latent class variable $z_i \in \{1, \ldots, k\}$ under disease y, and $\sum_{h=1}^{k} \nu_{yh} = 1$. $\boldsymbol{\psi}^{(j)}_{yh} = (\psi^{(j)}_{yh1}, \ldots, \psi^{(j)}_{yhd_j})$ is a vector of the multinomial probabilities of $x_{ij} = 1, \ldots, d_j$ given disease y and latent class component h.

The sparsity assumption (8) is key to sharing of information between the disease groups and latent classes. In particular, in each disease group y and component h, we partition the p dimensions of covariates into two mutually exclusive subsets $S_{yh} \cup S^c_{yh} = \{1, \ldots, p\}$, and for the variables within subset S^c_{yh}, we allocate $\psi^{(j)}_{yhc_j}$ to its baseline category $\lambda^{(j)}_{0c_j}$, which is not dependent on the latent class h or the outcome group y. $\boldsymbol{\lambda}^{(j)}_0 = (\lambda^{(j)}_{01}, \ldots, \lambda^{(j)}_{0d_j})$ vectors are *fixed in advance*; one natural choice is as follows: $\boldsymbol{\lambda}^{(j)}_0 = \left(\frac{1}{d_j}, \ldots, \frac{1}{d_j}\right)'$ corresponding to a discrete uniform. This dramatically reduces the number of parameters needed to learn the distribution of x_i by replacing $\boldsymbol{\psi}^{(j)}_{yh}$ with the fixed $\boldsymbol{\lambda}^{(j)}_0$, instead of having to estimate, for a large number of variables, outcome groups and latent classes. A Bayesian approach is used to learn the allocation of the subsets for each variable. Although assuming fixed $\boldsymbol{\lambda}^{(j)}_0$ may seem overly-restrictive, alternative methods that allow fully or empirical Bayes estimation of these parameters have inferior performance to the simple uniform default choice in our experience. This is likely due in part to the fact that the data are not sufficiently abundant to inform about all of the model parameters.

Consider a simple case of three covariates. If we let $\psi^{(3)}_{yhc_3} = \lambda^{(3)}_{0c_3}$ for $h = 1, \ldots, k$ and $y = 0, 1$, and $\psi^{(j)}_{yhc_j} = \lambda^{(j)}_{yhc_j}$ for $j = 1, 2$, we have

$$\begin{aligned} \Pr(x_{i1} = c_1, x_{i2} = c_2, x_{i3} = c_3 | y_i = y) &= \lambda^{(3)}_{0c_3} \sum_{h=1}^{k} \nu_{yh} \lambda^{(1)}_{yhc_1} \lambda^{(2)}_{yhc_2} \\ &= \Pr(x_{i3} = c_3) \cdot \Pr(x_{i1} = c_1, x_{i2} = c_2 | y_i = y), \end{aligned}$$

implying the third covariate is independent of the outcome and does not have any interaction with the other two variables. The sparsity assumption also has flexibility in allowing $j \in S^c_{yh}$ for some but not all $h \in \{1, \ldots, k\}$, which leads to some interactions between the j^{th} factor and the other factors. This implicitly indicates the j^{th} covariate can be associated with the disease through the other factors correlated with the disease. Moreover, if a variable j is independent of the other covariates, a marginal association between the j^{th} variable and the outcome can be introduced by having $j \in S^c_{yh}$ for all h but not for all y. In practice, the cardinality of S_{yh} (denoted as $|S_{yh}|$) is unknown but can be estimated by a Bayesian approach, which will be discussed later. In summary, our model has flexibility in allocating j for different combinations of y and h, leading to a flexible and complex dependence structure between covariates and outcome. Furthermore, our model also allows subjects in different outcome groups to have a different mixture probability for each class h (i.e., ν_{yh}), resulting in a more flexible distribution structure for x_i for each outcome group.

Our model has excellent performance in high-dimensional case-control applications due to the combination of flexibility (accounting for arbitrarily complex main effects and interactions), interpretability (in terms of variable selection), and (crucially) two layers of dimensionality reduction. The first layer is from the Bayesian low rank tensor decomposition of Dunson and Xing (2009), equivalent to (7) without (8). This reduces the number of parameters from $2(\prod_{j=1}^{p} d_j - 1)$, obtained by modeling the joint distribution of all covariates nonparametrically within each response category, to $2(k-1) + 2\sum_{j=1}^{p} k(d_j - 1)$. The second layer of dimension reduction reduces the model space further to $2(k-1) + \sum_{y=0}^{1} \sum_{h=1}^{k} \sum_{j \in S_{yh}} (d_j - 1)$. This is achieved through allowing the effective degrees of freedom used in characterizing interactions to vary across the variables by setting $\boldsymbol{\psi}^{(j)}_{yh}$ to a baseline-fixed variable $\boldsymbol{\lambda}^{(j)}_0$ for all $j \in S^c_{yh}$. These two dimensionality reduction steps are adaptive to the true low-dimensional structure in the data, while maintaining the flexibility of nonparametric modeling. The result is similar to placing heavy-tailed shrinkage priors on the coefficients in a saturated logistic regression, with more shrinkage for higher order interactions, but such a logistic shrinkage approach is intractable computationally. Our approach efficiently finds a low-dimensional representation of complex dependence between covariates and the disease outcome.

The model and associated methodology proposed in this article are fundamentally different from previous approaches for categorical data analysis using tensor factorization methods. We are most similar to Zhou et al. (2014), but that

Table 1
True coefficients in log-linear model assuming four mutually correlated variables associated with outcome

	β_2^y	β_4^y	β_{12}^y	β_{14}^y	$\beta_{2,4}^y$	$\beta_{2,12}^y$	$\beta_{2,14}^y$	$\beta_{4,12}^y$	$\beta_{4,14}^y$	$\beta_{12,14}^y$
$y=0$	0.5	−1.5	−2	1	−0.5	0.5	0	0.5	−0.5	−0.5
$y=1$	3	−3	−0.5	4	−0.5	0.5	0	0.5	−0.5	−0.5

	$\beta_{2,4,12}^y$	$\beta_{2,4,14}^y$	$\beta_{2,12,14}^y$	$\beta_{4,12,14}^y$	$\beta_{2,4,12,14}^y$
$y=0$	0.25	0	0	0.5	0
$y=1$	0.25	0	0	0.5	0

approach focuses on joint modeling of multivariate categorical data. We could potentially use the Zhou et al. (2014) approach for the retrospective likelihood. This would place separate models on the covariate distribution within each outcome group, implicitly assuming that all covariates are related to the response, not allowing variable selection, and not exploiting similarities in the structure across the groups. Our model modifies Zhou et al. (2014) to borrow information across the groups, in order to conduct variable selection and allow learning of which parameters are common in the factorizations of the two conditional distributions. This borrowing of information reduces the number of parameters to $2(k-1)+\sum_{y=0}^{1}\sum_{h=1}^{k}\sum_{j\in S_{yh}}(d_j-1)$ instead of the $2(k-1)+2\sum_{h=1}^{k}\sum_{j\in S_h}(d_j-1)$ that would be obtained in applying separate Zhou et al. (2014) factorizations.

Our proposed model (7) with assumptions (8) can be expressed in a hierarchical form with priors specified for the unknown parameter vectors: for $y=0$ or 1,

$$x_{ij}|y_i=y, z_i=h \sim \text{Mult}\Big(\{1,\ldots,d_j\}; \psi_{yh1}^{(j)},\ldots,\psi_{yhd_j}^{(j)}\Big),$$

$$\boldsymbol{\psi}_{yh}^{(j)} \equiv \Big(\psi_{yh1}^{(j)},\ldots,\psi_{yhd_j}^{(j)}\Big) \sim (1-\tau_{yh})\delta_{\boldsymbol{\lambda}_0^{(j)}} + \tau_{yh}\text{Diri}(a_{j1},\ldots,a_{jd_j}), \tag{9}$$

$$\Pr(z_i=h|y_i=y)=\nu_{yh}=V_{yh}\prod_{l<h}(1-V_{yl}),$$

$$V_{yh}\sim\text{Beta}(1,\alpha),\quad \alpha\sim\text{Gamma}(a_\alpha,b_\alpha),\quad \tau_{yh}\sim\text{Beta}(1,\gamma).$$

Expression (9) is equivalent to letting the subset-size $|S_{yh}|\sim\text{Binom}(p,\tau_{yh})$ and drawing a random subset S_{yh} uniformly from all subsets of $\{1,\ldots,p\}$ of size $|S_{yh}|$ in (8). Although we could potentially choose a prespecified rank for the factorization, using a nonparametric Bayes approach through a stick-breaking representation of the Dirichlet process prior (Sethuraman, 1994) allows for uncertainty in rank selection. A hyperprior on the concentration parameter α allows the data to inform more strongly about the component weights. The probability of allocation τ_{yh} to the *active* (nonbaseline) category is chosen as beta$(1,\gamma)$, with $\gamma>1$ favoring allocation of many of the $\boldsymbol{\psi}_{yh}^{(j)}$s to the baseline category $\boldsymbol{\lambda}_0^{(j)}$ in both outcome groups. Posterior computation proceeds via a simple Gibbs sampler, with the steps shown in a Web Supplement.

3. Simulation Studies

3.1. *Simulation from Log-Linear Models*

We first conduct a replicated simulation study mimicking a case-control design to assess the performance using the proposed model compared with logistic regression with and without the Benjamini and Hochberg correction (Benjamini and Hochberg, 1995), CART (Breiman et al., 1984), random forest (Breiman, 2001), and lasso (Tibshirani, 1996). For 50 cases and 50 control subjects, we simulated p binary covariates $x_{ij}\in\{0,1\}$, $j=1,\ldots,p$, under two scenarios: (i) $p=20$, and (ii) $p=100$, among which four variables ($j=2,4,12,14$) were assumed dependent and generated from a saturated log-linear model with coefficients varying by outcome y:

$$\log\left(\frac{\pi_{c_2,c_4,c_{12},c_{14}}^y}{\pi_{0,0,0,0}^y}\right)=\sum_{s=1}^{4}\sum_{S^*\subset\{2,4,12,14\}:|S^*|=s}\beta_{S^*}^y 1_{(c_{S^*}=1)}, \tag{10}$$

where $\pi_{c_2,c_4,c_{12},c_{14}}^y=\Pr(x_{i2}=c_2,x_{i4}=c_4,x_{i,12}=c_{12},x_{i,14}=c_{14}\mid y_i=y)$. If $S^*=\{2,4\}$, for example, then $\beta_{S^*}^y=\beta_{2,4}^y$ and $1_{(c_{S^*}=1)}=1_{(c_2=1,c_4=1)}$ with $1_{(\cdot)}$ denoting the indicator function. Different values of c_j, $j=2,4,12,14$, will lead to different coefficients in the model. One illustration is if $c_2=1,c_4=1,c_{14}=1$, model (10) becomes

$$\log\left(\frac{\pi_{1,1,0,1}^y}{\pi_{0,0,0,0}^y}\right)=\beta_2^y+\beta_4^y+\beta_{14}^y+\beta_{2,4}^y+\beta_{2,14}^y+\beta_{4,14}^y+\beta_{2,4,14}^y. \tag{11}$$

All the true coefficients are set as in Table 1.

Having different main effects given disease outcome in the log-linear model result in association between the outcome and those four variables. All the remaining null variables $j\in\{1,\ldots,p\}$, $j\neq\{2,4,12,14\}$ were independently generated from a discrete uniform distribution. This data generating mechanism induces dependence among the variables in S^* and their impact on outcome, while rendering the other variables marginally independent.

Simulations were conducted based on 1000 data replicates for each scenario. In each replicate, the posterior marginal odds ratio for each variable j can be computed according to

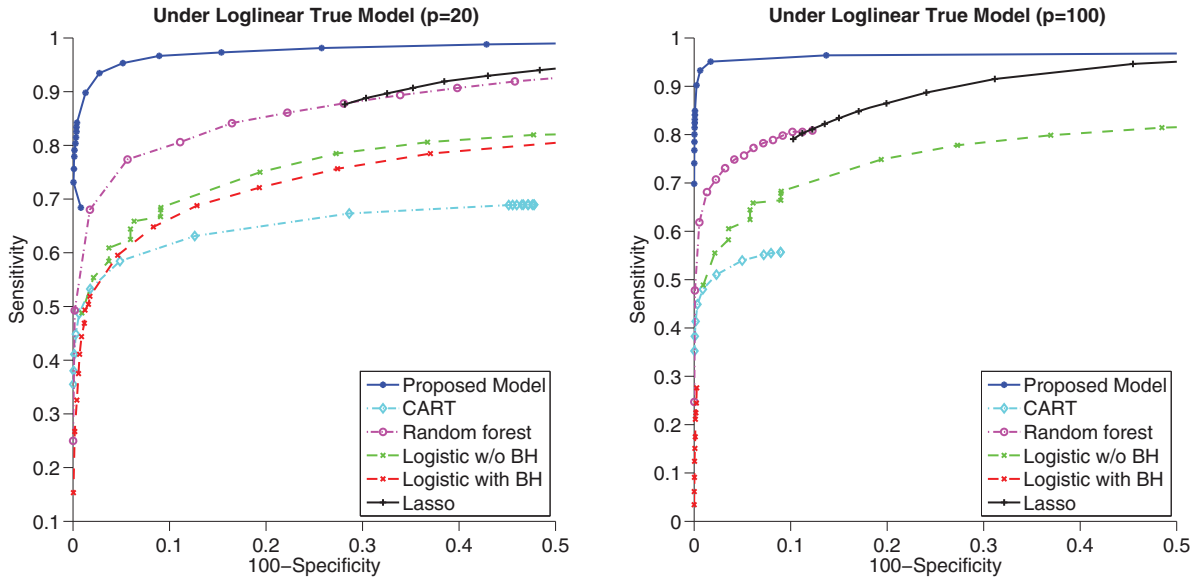

Figure 1. ROC curves comparison under log-linear true models ($n = 100$)—Left: $p = 20$; Right: $p = 100$.

the following:

$$\mathrm{OR}^{(j)} = \frac{\Pr(x_{ij} = 1|y_i = 1)}{\Pr(x_{ij} = 0|y_i = 1)} \Big/ \frac{\Pr(x_{ij} = 1|y_i = 0)}{\Pr(x_{ij} = 0|y_i = 0)}, \quad (12)$$

where

$$\Pr(x_{ij} = c_j|y_i = y) = \begin{cases} \sum_{h=1}^{k} \nu_{yh} \lambda_{yhc_j}^{(j)}, & \text{if} \quad j \in S_{yh} \\ \sum_{h=1}^{k} \nu_{yh} \lambda_{0c_j}^{(j)} = \lambda_{0c_j}^{(j)}, & \text{if} \quad j \in S_{yh}^{c} \end{cases}. \quad (13)$$

The corresponding credible interval of the odds ratio was used to identify whether the variable j was significant. For each data replicate, we ran the Gibbs sampler described in the Web Supplement for 25, 000 iterations, discarding the first 10, 000 iterations as burn-in, and collecting every fifth sample post burn-in to thin the chain. Mixing and convergence rates were good based on examination of trace plots; in fact, substantially shorter chains would have produced essentially identical results.

Receiver Operating Characteristic (ROC) curves were plotted to compare the performance among methods under the two p scenarios, respectively. ROC is a plot of the true positive rate (sensitivity) against the false positive rate (100-specificity) for different possible cut-off points. In our case, we define sensitivity as the combined power for all true variables, while 100-specificity is the combined type-I error rate for all null variables. To aggregate the results from simulation replicates, we computed the percentage of significance for each predictor over 1000 replicates, and then averaged these percentages for the four true variables as the sensitivity. Likewise, we averaged over these percentages for the remaining $p - 4$ null variables as the 100-specificity.

Each point on the ROC curve represents a sensitivity/100-specificity pair corresponding to a particular decision threshold. A wide and fine grid of thresholds was chosen for each method to produce a full curve in the figure. We chose $\{10\%, 20\%, \ldots, 90\%, 91\%, \ldots, 99\%\}$ as the credible interval thresholds for our proposed approach. As for logistic regression, $\{0.01, 0.02, \ldots, 0.1, 0.2, \ldots, 0.9\}$ were set as the p-value cut-off points. Deviance thresholds $\{0.001, 0.002 \ldots, 0.01, 0.02, \ldots, 0.1\}$ were used for CART for splitting the tree. In random forest, the ROC curve was derived by selecting the $\{1, 2, \ldots, 15\}$ highest importance scores as the significant variables. 0 to 1 penalty with an interval of 0.1 was chosen for lasso.

As illustrated in Figure 1, the proposed Bayesian case-control method is obviously the best among the six approaches for both p cases. Our method tends to have much smaller combined type-I error and provides better power. Note that the x axis scale is only shown within [0,0.5] for display purposes and because large type-I error is not acceptable. Some ROC curves are cut off due to the scale limit.

Another case considered is including covariates which are correlated, but not associated with outcome, in addition to the outcome-dependent correlated variables mentioned above. We added another four variables using a saturated log-linear model similar to (10) but having the same coefficients in both disease groups instead. The coefficients are listed in Table 2. The new mechanism results in the extra four covariates correlated to each other but not impacting disease. All the other

Table 2
True coefficients for additional four correlated variables independent of outcome

β_1	β_3	β_{11}	β_{13}	$\beta_{1,3}$	$\beta_{1,11}$	$\beta_{1,13}$	$\beta_{3,11}$
1.5	−0.5	2	−1	−0.5	0.25	0	−0.5
$\beta_{3,13}$	$\beta_{11,13}$	$\beta_{1,3,11}$	$\beta_{1,3,13}$	$\beta_{1,11,13}$	$\beta_{3,11,13}$	$\beta_{1,3,11,13}$	
0.5	−0.5	−0.25	0.25	−0.25	0	0	

Table 3
Power and type-I error using 95% credible intervals including 1 as the significance threshold

		Log-linear model		Latent class model	
		$p = 20$	$p = 100$	$p = 20$	$p = 100$
Power	Four correlated covariates	0.791	0.801	0.697	0.664
	Eight correlated covariates	0.792	0.801	0.701	0.673
Type-I error	Four correlated covariates	0.002	3.75E-04	0.003	0.003
	Eight correlated covariates	0.016	0.004	0.022	0.008

$p - 8$ variables are generated independently from a discrete uniform distribution. We then created two new ROC curves for both p cases shown in Figure S1 in the Web Supplement. Compared with Figure 1, we obtained a slightly inflated false positive rate. However, our approach still performs much better than the other methods.

3.2. *Simulation from Latent Class Models*

We now perform another simulation study with data generated from a latent class model rather than a log-linear model. We again had 50 cases and 50 control subjects having p binary predictors with (i) $p = 20$ and (ii) $p = 100$ for each data replicate. We assume four predictors are associated with the outcome in the true model, whereas those four variables ($j = 2, 4, 12, 14$) are correlated by introducing the multiple latent classes $z = 1, \ldots, k$, with each latent class having different marginal probabilities for those four variables. We assumed 80% of individuals fell into the first latent class, with the remaining individuals in a second latent class ($k = 2$). Furthermore, the variable dependence on disease outcome can be induced by letting the marginal probabilities under each latent class vary by disease outcome. In particular,

$$Pr(x_{ij} = c_j \mid y_i = y, z_i = h) = \psi^{(j)}_{yhc_j}, \quad h = 1, 2; \; y = 0, 1 \tag{14}$$

where vector $\boldsymbol{\psi}^{(j)}_{yh}$ only varies by y and h for $j = 2, 4, 12, 14$. All the remaining predictors were generated from a discrete uniform distribution with $\boldsymbol{\psi}^{(j)}_{yh} = \boldsymbol{\lambda}^{(j)}_0 \equiv (\frac{1}{2}, \frac{1}{2})'$. Within a latent class and the disease group, all the variables are conditionally independent. However, marginalizing out the latent class indicator, one obtains dependence in those variables that have different marginal probabilities across the latent classes, conditional on the disease outcome. Additionally, it becomes clear that having the probability of those correlated covariates differing by the disease group implies the association between the outcome and those four variables.

We generated 1000 simulated datasets, then ran the Gibbs sampler in the Web Supplement with satisfactory mixing and convergence rates. We also computed the posterior samples of the odds ratios to assess the ROC performance. As in Figure S2 shown in the Web Supplement, our approach outperforms the other methods for both $p = 20$ and 100 with much better combined power and lower type-I error.

To examine the performance of all methods when including correlated covariates not associated with outcome, we generated another four correlated variables for both p cases, with different marginal probabilities in each latent class but not varying by disease group (i.e., $Pr(x_{ij} = c_j \mid z_i = h) = \lambda^{(j)}_{hc_j}, h = 1, 2$). The corresponding ROC curves, based on 1000 simulated datasets, are provided in Figure S3 in the Web Supplement. It shows that the performance is similar to Figure S2, but more complex data adding correlated covariates not related to outcome slightly affects the false positive rate.

The power and type-I error rates for screening based on 95% credible intervals for odds ratios not including 1 are provided in Table 3. In both simulation scenarios, the power does not appear to be affected with respect to adding more correlated covariates. The type-I error, as we observed in the ROC curves, is slightly inflated but remains well below 0.05 in a more correlated covariate structure.

4. Application to the National Birth Defects Prevention Study

Our method is motivated by the analysis of multiple diverse exposures and birth defects using data from the National Birth Defects Prevention Study (NBDPS). NBDPS is a multisite etiologic case-control study with approximately 30,000 cases and 10,000 controls that was started in 1997 (Yoon et al., 2001). The study was designed to evaluate environmental, behavioral, biomedical, sociodemographic, genetic, and occupational factors and their association with the prevalence of 30 major structural birth defects. There are nine states currently participating in this study: Arkansas, California, Georgia, Iowa, Massachusetts, New York, North Carolina, Texas, and Utah. Control infants without congenital anomalies were randomly selected from either birth certificates or hospital records, depending on the state center's protocol.

For this analysis, we investigated the association of 172 potential risk factors for conotruncal heart defects, which comprise a subgroup of congenital heart defects (CHDs) that are malformations of cardiac outflow tracts and great arteries. Common types of conotruncal heart defects include truncus arteriosus, transposition of the great arteries, double outlet right ventricle, and tetralogy of Fallot. CHDs are the most prevalent structural birth defect, occurring in 8–11 of every 1000 live births, while conotruncal heart defects account for approximately 20 to 30% of all CHDs (Hobbs et al., 2014).

We employed our case-control Bayesian method to investigate associations using odds ratios (OR) as the measure of association and compared the results with standard univariate logistic regression. The analyses under our proposed method were conducted in two different ways: (i) all cases and con-

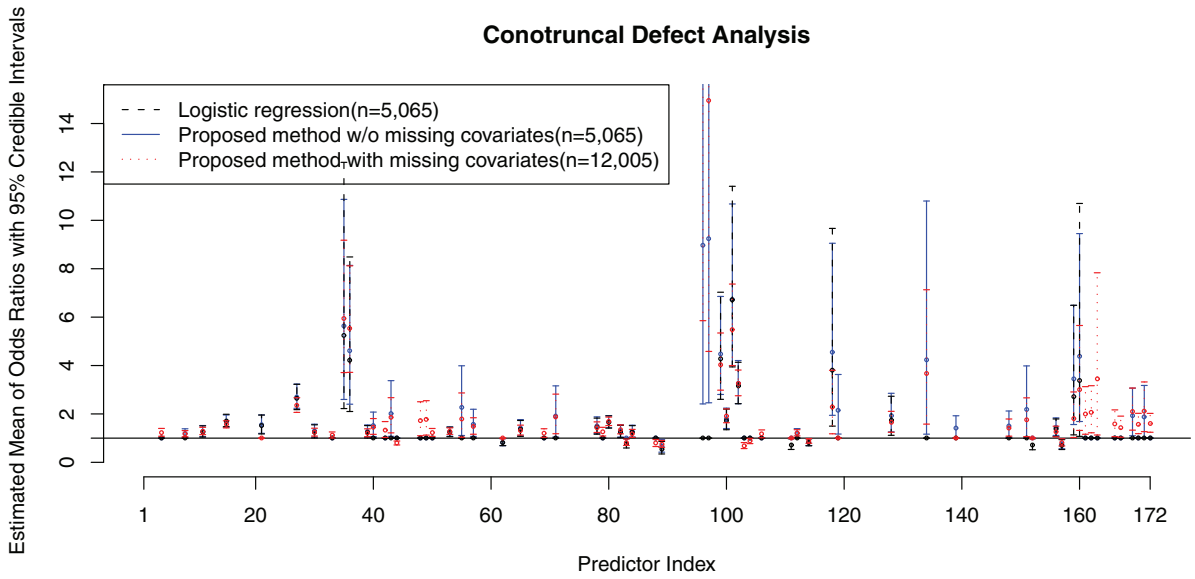

Figure 2. Significant odds ratio results for conotruncal heart defect comparing univariate logistic regression and proposed method omitting or retaining missing data.

trols with missing covariates were removed before analysis to facilitate comparison with previous NBDPS publications; (ii) keeping all the missing data. We note that our method can easily accommodate missing data. In particular, due to the form of the model, it is not even necessary to impute the missing observations from their full conditional distributions; the forms of the Gibbs sampler conditionals remain the same with some additional book-keeping. Thus, compared with competing approaches, we have another advantage of handing missing data.

We ran 20,000 iterations with the first 5000 as a burn-in, collecting every fifth sample post burn-in to thin the chain. The effective sample sizes for marginal odds ratios range from 2428 to 3000, which suggests the thinned samples are close to independent. Trace plot examination indicates satisfactory mixing and convergence. We estimated a 95% credible interval (CI) for the marginal odds ratio for each factor. Figure 2 shows the estimated odds ratios and 95% CIs for factors having 95% intervals not including 1 at least in one of the three models. For clarity, only the significant CIs are displayed, meaning the nonsignificant CIs are denoted as a point at 1 in the figure.

In this analysis, there are no conflicting associations detected in the three approaches. Of 172 factors examined, all three approaches agreed that the factors were not significant predictors (95% interval estimates contained the null value 1) in 106 cases, and all three approaches agreed the factors were significant predictors (in the same direction) in 25 cases. In the remaining 41 cases, point estimates were all in the same direction, and interval estimates simply differed in inclusion of the null value 1. In 18 of these cases, both variants of our proposed methods had interval estimates that did not contain the null value 1, while the estimates from logistic regression were not statistically significant. In another 16 cases, only the proposed method with missing data as in (ii) produced interval estimates that did not contain the null value 1. This often occurred for predictors with higher levels of missing data, for example paternal exposures, which were not available for all participants due to lack of data on all fathers. In only two cases were effects identified using logistic regression but not the proposed methods. The remaining five cases were more heterogeneous (e.g., significance only in logistic regression and in (i) or (ii) above).

Due to space limitations, we are unable to describe all associations but concentrate on a few interesting findings here. All models identified many well-established predictors of conotruncal defects. Unless otherwise indicated, posterior means and 95% credible intervals presented are from the model (ii) that accounts for missing data; estimates from model (i) and logistic regression were similar in nature to those presented. In general, results from our proposed method had shorter credible interval width than confidence interval from logistic regression due to borrowing of information, and results from method (ii) had shorter credible interval width than those from method (i) due to accommodation of missing data (e.g., on paternal exposures).

Specifically, women with type-I (OR = 5.95, 95% CI = 3.71, 9.18) and type-II (OR = 5.54, 95% CI = 3.72, 8.12) diabetes were at much greater risk than their counterparts without preexisting diabetes, consistent with the effect estimates reported in Correa et al. (2012). Women whose pregravid body mass index (BMI) exceeded 30 kg/m^2 had 1.29 (1.14, 1.45) times the odds of having a baby with a conotruncal defect than women in the preferred range with BMI 18.5–25 kg/m^2, consistent with Waller et al. (2007), who analyzed all heart defects combined. In an earlier case group from NBDPS, Malik et al. (2007) noted a significant association between conotruncal defects and small-for-gestational age births, with an OR = 2.41 (95% CI = 1.89, 3.08), which was similar to the association seen in our analysis with OR = 2.36 (95% CI = 2.06, 2.67).

As stated above, our analysis results are generally consistent with previous findings from the NBDPS (Dawson et al.,

2012, 2013). However, some associations had not been previously identified or suffered from wide interval estimates in earlier analyses and will now be followed-up with more detailed analysis. For example, Alwan et al. (2007) estimated an OR = 1.3 (0.8, 2.1) for maternal selective serotonin reuptake inhibitor (SSRI) use during pregnancy. In our analysis, which involved additional years of NBDPS recruitment, we estimated an OR = 1.48 (95% CI = 1.16, 1.84) for these medications, typically used to treat depressive symptoms. In a simple logistic regression model fit to our data, the point estimate was similar, but the interval estimate of (0.96, 2.09) included the null value.

5. Discussion

In this article, a new method utilizing a sparse parallel factor analysis model has been proposed for case-control designs. It has been shown through simulations that it has exceptional performance in identifying true predictors while keeping the type-I error rate very small. The strong performance, compared to existing methods, is due to flexible distribution modeling for the retrospective likelihood and borrowing information among variables in our model. This method can be applied to any case-control study that has many categorical covariates with an interest in investigating the marginal associations. Our method also has the flexibility of allowing outcomes to be multicategorical, not necessarily case and control. In such a case, the Gibbs sampling steps shown in the Web Supplement stay the same except for minor updates to include all the category probabilities of y. Our article is focused on developing flexible nonparametric methods for improving inferences on marginal associations, motivated in particular by birth defects case-control studies. An important next step is to develop approaches for inferences on conditional associations.

6. Supplementary Materials

Web Supplement referenced in Sections 2, 3, and 5 is available with this paper at the *Biometrics* website on Wiley Online Library. Web Supplement also contains a template matlab code implementing the proposed method for case-control data with many binary covariates.

Acknowledgements

This study was supported by NIEHS R01ES020619; a cooperative agreement from the Centers for Disease Control and Prevention under PA 96043, PA 02081, and FOA DD09-001; and contract 200-2000-08018 from the Centers for Disease Control and Prevention and the National Institute for Occupational Safety and Health. Coding of drug information in the NBDPS used the Slone Drug Dictionary under license from the Slone Epidemiology Center of Boston University. The authors wish to thank the NBDPS study participants for their dedication in facilitating the study of congenital malformations. In addition, Dr. Bhattacharya acknowledges support for this project from the Office of Naval Research (ONR BAA 14-0001).

References

Alwan, S., Reefhuis, J., Rasmussen, S. A., Olney, R. S., and Friedman, J. M. (2007). Use of selective serotonin-reuptake inhibitors in pregnancy and the risk of birth defects. *New England Journal of Medicine* **356**, 2684–2692.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Ashby, D., Hutton, J. L., and McGee, M. A. (1993). Simple Bayesian analyses for case-control studies in cancer epidemiology. *The Statistician* **42**, 385–397.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.

Bhattacharya, A. and Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98,** 291–306.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. New York: CRC press.

Byrne, S. P. and Dawid, A. P. (2013). Retrospective-prospective symmetry in the likelihood and Bayesian analysis of case-control studies. *Biometrika* 050.

Correa, A., Gilboa, S. M., Botto, L. D., Moore, C. A., Hobbs, C. A., Cleves, M. A., Riehle-Colarusso, et al. (2012). Lack of periconceptional vitamins or supplements that contain folic acid and diabetes mellitus–associated birth defects. *American journal of obstetrics and gynecology* **206**, 218–e1.

Dawson, A., Flak, A., and Reefhuis, J. (2013). National birth defects prevention study matrix of results published 2004-July 2012. In *National Birth Defects Prevention Network (NBDPN) 16th Annual Meeting. February 25–27, 2013. Atlanta, Georgia.*

Dawson, A., Flak, A., Reefhuis, J., and the National Birth Defects Prevention Study (2012). National birth defects prevention study results matrix 1997–2011. In *National Birth Defects Prevention Network (NBDPN) 15th Annual Meeting. February 27–29, 2012. Arlington, Virginia.*

Dunson, D. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**, 1042–1051.

Ghosh, M. and Chen, M.-H. (2002). Bayesian inference for matched case-control studies. *Sankhyā: The Indian Journal of Statistics, Series B* **64**, 107–127.

Hobbs, C. A., Cleves, M. A., MacLeod, S. L., Erickson, S. W., Tang, X., Li, J., et al. (2014). Conotruncal heart defects and common variants in maternal and fetal genes in folate, homocysteine, and transsulfuration pathways. *Birth Defects Research Part A: Clinical and Molecular Teratology* **100**, 116–126.

Kunihama, T. and Dunson, D. B. (2013). Bayesian modeling of temporal dependence in large sparse contingency tables. *arXiv:1205.2816*.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton, Mifflin.

Malik, S., Cleves, M. A., Zhao, W., Correa, A., Hobbs, C. A., et al. (2007). Association between congenital heart defects and small for gestational age. *Pediatrics* **119**, e976–e982.

Marshall, R. J. (1988). Bayesian analysis of case-control studies. *Statistics in Medicine* **7**, 1223–1230.

Mukherjee, B., Sinha, S., and Ghosh, M. (2005). Bayesian analysis of case-control studies. *Handbook of Statistics* **25**, 793–819.

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.

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

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

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

Rothman, K. J., Greenland, S., and Lash, T. L. (2012). *Modern Epidemiology, 3rd Edition*. Philadelphia, PA: Lippincott Williams & Wilkins.

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

Seaman, S. R. and Richardson, S. (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika* **91**, 15–25.

Sethuraman, J. (1994). A constructive definition of Dirichlet measures. *Statistica Sinica* **4,** 639–650.

Sinha, S., Mukherjee, B., and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case–control studies with multiple disease states. *Biometrics* **60**, 41–49.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B. K., and Carroll, R. J. (2005). Semiparametric Bayesian analysis of matched case-control studies with missing exposure. *Journal of the American Statistical Association* **100**, 591–601.

Staicu, A.-M. (2010). On the equivalence of prospective and retrospective likelihood methods in case-control studies. *Biometrika* **97**, 990–996.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288.

Waller, D. K., Shaw, G. M., Rasmussen, S. A., Hobbs, C. A., Canfield, M. A., Siega-Riz, A.-M., et al. (2007). Prepregnancy obesity as a risk factor for structural birth defects. *Archives of Pediatrics & Adolescent Medicine* **161**, 745–750.

Yoon, P. W., Rasmussen, S. A., Lynberg, M., Moore, C., Anderka, M., Carmichael, S., et al. (2001). The national birth defects prevention study. *Public Health Reports* **116**, 32.

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

Zhou, J., Bhattacharya, A., Herring, A., and Dunson, D. (2014). Bayesian factorizations of big sparse tensors. *arXiv:1306.1598*.

Received November 2014. *Revised June* 2015.
Accepted August 2015.