Journal of the
American
Statistical
Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Simplex Factor Models for Multivariate Unordered Categorical Data

Anirban Bhattacharya & David B. Dunson

To cite this article: Anirban Bhattacharya & David B. Dunson (2012) Simplex Factor Models for Multivariate Unordered Categorical Data, Journal of the American Statistical Association, 107:497, 362-377, DOI: 10.1080/01621459.2011.646934

To link to this article: http://dx.doi.org/10.1080/01621459.2011.646934

Simplex Factor Models for Multivariate Unordered Categorical Data

Anirban BHATTACHARYA and David B. DUNSON

Gaussian latent factor models are routinely used for modeling of dependence in continuous, binary, and ordered categorical data. For unordered categorical variables, Gaussian latent factor models lead to challenging computation and complex modeling structures. As an alternative, we propose a novel class of simplex factor models. In the single-factor case, the model treats the different categorical outcomes as independent with unknown marginals. The model can characterize flexible dependence structures parsimoniously with few factors, and as factors are added, any multivariate categorical data distribution can be accurately approximated. Using a Bayesian approach for computation and inferences, a Markov chain Monte Carlo (MCMC) algorithm is proposed that scales well with increasing dimension, with the number of factors treated as unknown. We develop an efficient proposal for updating the base probability vector in hierarchical Dirichlet models. Theoretical properties are described, and we evaluate the approach through simulation examples. Applications are described for modeling dependence in nucleotide sequences and prediction from high-dimensional categorical features.

KEY WORDS: Classification; Contingency table; Factor analysis; Latent variable; Nonparametric Bayes; Nonnegative tensor factorization; Mutual information; Polytomous regression.

1. INTRODUCTION

Multivariate unordered categorical data are routinely encountered in a variety of application areas, with interest often in inferring dependencies among the variables. For example, the categorical variables may correspond to a sequence of A, C, G, T nucleotides or responses to questionnaire data on race, religion, and political affiliation for an individual. We shall use $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$ to denote the multivariate observation for the ith subject, with $y_{ij} \in \{1, \ldots, d_j\}$.

Complicated dependence can potentially be expressed in terms of simpler conditional independence relationships via graphical models (Dawid and Lauritzen 1993). Such models have been used for continuous (Lauritzen 1996; Dobra et al. 2004), categorical (Whittaker 1990; Madigan and York 1995), and mixed-scale variables (Pitt, Chan, and Kohn 2006; Dobra and Lenkoski 2011). Although graphical models are popular due to their flexibility and interpretability, computation is daunting since the size of the model space grows exponentially with p. Even with highly efficient search algorithms (Jones et al. 2005; Carvalho and Scott 2009; Dobra and Massam 2010; Lenkoski and Dobra 2011, among others), it is only feasible to visit a tiny subset of the model space even for moderate p. Accurate model selection in this context is difficult when p is moderate to large and the number of samples is not enormous because, in such cases, even the highest posterior probability models receive very small weight and there will typically be a large number of models having essentially identical performance according to any given model selection criteria (Akaike information criterion [AIC], Bayesian information criterion [BIC], etc). Dobra and Lenkoski (2011) advocated model averaging to avoid the inferences to depend explicitly on the choice of the underlying graph.

In parallel to the development of graphical models, factor models (West 2003; Carvalho et al. 2008) have been widely used for modeling of high-dimensional variables and dimension reduction. While Gaussian graphical models work with the precision matrix, factor models provide a framework for regularized covariance matrix estimation. Factor models and generalizations such as structural equation models (Bollen 1989) can accommodate mixtures of continuous, binary, and ordered categorical data through an underlying Gaussian latent factor structure (Muthén 1983). Such models link each observed y_{ij} to an underlying continuous variable z_{ij}, with ordinal y_{ij} arising via thresholding of z_{ij}. For multivariate binary y_i, a multivariate Gaussian distribution on $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^{\mathrm{T}}$ induces the multivariate probit model (Ashford and Sowden 1970; Chib and Greenberg 1998; Ochi and Prentice 1984). Multivariate probit models can accommodate nominal data with $d_j > 2$ by introducing a vector of variables $z_{ij} = (z_{ij1}, \ldots, z_{ijd_j})^{\mathrm{T}}$ underlying y_{ij}, with $y_{ij} = l$ if $z_{ijl} = \max z_{ij}$ (Aitchison and Bennett 1970; Zhang, Boscardin, and Belin 2008). The latent z_i's are usually modeled as $\sum_{j=1}^{p} d_j$-dimensional Gaussian with covariance Σ, with at least p diagonal elements of Σ constrained to be one for identifiability. This constraint makes sampling from the full conditional posterior of Σ difficult. Zhang, Boscardin, and Belin (2006) used a parameter-expanded Metropolis–Hastings algorithm to obtain samples from a correlation matrix for multivariate probit models. Zhang, Boscardin, and Belin (2008) extended their algorithm to multivariate multinomial probit models. An alternative to probit models is to define a generalized linear model for each of the individual outcomes, while including shared normal latent traits to induce dependence (Sammel, Ryan, and Legler 1997; Dunson 2000, 2003; Moustaki and Knott 2000).

For unordered categorical variables, the data could be alternatively presented in the form of a p-way contingency table of dimension $d_1 \times \cdots \times d_p$. There is a vast literature on analysis of

Anirban Bhattacharya is a Ph.D candidate in the Department of Statistical Science, Duke University, NC 27708 (E-mail: *ab179@stat.duke.edu*). David B. Dunson is a Professor in the Department of Statistical Science, Duke University, NC 27708 (E-mail: *dunson@stat.duke.edu*). The authors thank Adrian Dobra for sharing the Rochdale data and providing insightful comments. This research was partially supported by grant number R01 ES017240-01 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH).

Journal of the American Statistical Association
March 2012, Vol. 107, No. 497, Theory and Methods
DOI: 10.1080/01621459.2011.646934

contingency tables dating back to the nineteenth century. Fienberg and Rinaldo (2007) provided an excellent chronological overview of the development of log-linear models, maximum likelihood estimation, and asymptotic tests for goodness of fit. While log-linear models (Bishop, Fienberg, and Holland 1975) have been extensively used to model interactions among related categorical variables, asymptotic tests based on log-linear models face multiple difficulties in the case of sparse contingency tables—refer to the discussion in Section 3 of Fienberg and Rinaldo (2007). In a Bayesian framework, such problems can be avoided by specifying priors of the log-linear model parameters; Massam, Liu, and Dobra (2009) provided a detailed study of a useful class of conjugate priors. Posterior model search in log-linear models using traditional Markov chain Monte Carlo (MCMC) methods tends to bog down quickly as dimensionality increases. Dobra and Massam (2010) proposed a mode-oriented stochastic search method to more efficiently explore high posterior probability regions in decomposable, graphical, and hierarchical log-linear models.

Each of the above-mentioned methods are flexible and have their own distinct advantages. Graphical log-linear models are often preferred for ease of interpretation, while the underlying variable methods are useful for mixed data types, which commonly arise in social science applications. However, these methods face major computational challenges for large contingency tables. Dunson and Xing (2009) developed a nonparametric Bayes approach using Dirichlet process (Ferguson 1973, 1974) mixtures of product multinomials to directly model the joint distribution of multivariate unordered categorical data. They assumed that $(y_{i1}, \dots, y_{ip})^{\mathrm{T}}$ are conditionally independent, given a univariate latent class index $z_i \in \{1, 2, \dots, \infty\}$. The prior specification is completed by assuming a stick-breaking process prior on the distribution of z_i and independent Dirichlet priors for the component-wise position-specific probability vectors. Marginalizing over the distribution of z_i induces dependence among the p variables. This approach extends latent structure analysis (Lazarsfeld and Henry 1968; Goodman 1974) to the infinite mixture case and is conceptually related to nonnegative tensor decompositions (Shashua and Hazan 2005; Kim and Choi 2007). The direct modeling of the joint distribution of the category probabilities in a sparse manner enables efficient posterior computation, thereby allowing their method to efficiently scale up to high dimensions.

Although the Dunson and Xing (2009) approach can handle large contingency tables, the assumption of conditional independence, given a single latent class index, seems restrictive. Although their prior has full support and hence they can flexibly approximate any joint distribution of y_i, in practice, even relatively simple dependence structures may require allocation of individuals to different classes, leading to a large effective number of parameters. Hence, in applications involving moderate to large p and modest sample size n, the Dunson and Xing (2009) approach may face difficulties.

In this article, we propose a new class of simplex factor models for multivariate unordered categorical data in which the dependence among the high-dimensional variables is explained in terms of relatively few latent factors. This is akin to Gaussian factor models, but factors on the simplex are more natural for nominal data. Methods for factor selection are discussed and the proposed approach is shown to have large support and to lead to consistent estimation of joint or conditional distributions for categorical variables. The Dunson and Xing (2009) model is obtained as a particular limiting case of the proposed model, as is the product multinomial model. The joint distribution of the multivariate nominal variables induced from the simplex factor model is shown to correspond to a multilinear singular value decomposition (SVD) (or higher-order SVD [HOSVD]) (De Lathauwer, De Moor and Vandewalle 2000) of probability tensors, which is regarded as a natural generalization of the matrix SVD in the tensor literature. A simple-to-implement data-augmented MCMC algorithm is proposed for posterior computation that scales well to higher dimensions. The methods are illustrated through simulated and real-data examples.

2. MODEL AND PRIOR SPECIFICATION

2.1 The Simplex Factor Model

Let $y_i = (y_{i1}, \dots, y_{ip})^{\mathrm{T}}$ denote a vector of responses and/or predictors. If $y_i \in \Re^p$, then a common approach is to jointly model the y_i's via a normal linear factor model:

$$y_i = \mu + \Lambda\eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Sigma), \quad i = 1, \dots, n, \tag{1}$$

where $\mu \in \Re^p$ is an intercept term, Λ is a $p \times k$ factor loadings matrix, $\eta_i \in \Re^k$ are latent factors, and ϵ_i is an idiosyncratic error with covariance $\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_p^2)$. When $\eta_i \sim \mathrm{N}_k(0, I_k)$, marginally, $y_i \sim N_p(\mu, \Omega)$, with $\Omega = \Lambda\Lambda^{\mathrm{T}} + \Sigma$, a decomposition which uses at most $p(k+1)$ free parameters instead of the $p(p+1)/2$ parameters in an unstructured covariance matrix.

Now consider the case in which $y_{ij} \in \{1, \dots, d_j\}$ for $j = 1, \dots, p$, and the different observations are unordered categorical variables. Let $\eta_i = (\eta_{i1}, \dots, \eta_{ik})^{\mathrm{T}} \in \mathsf{S}_{k-1}$, with S_{k-1} the $(k-1)$-dimensional simplex. The η_i's will play the role of the latent factors but they lie on the simplex instead of being in $\Re^k$. In addition, for each j, let ${}^{(j)}_{h} = (\lambda_{h1}^{(j)}, \dots, \lambda_{hd_j}^{(j)})^{\mathrm{T}}$ be a probability vector for $h = 1, \dots, k$. The ${}^{(j)}_{h}$'s can be interpreted as loadings for factor h and outcome j, but we have a vector instead of a single element, as we would have in the case in which $y_{ij} \in \Re$. With these components, we let:

$$\mathrm{pr}(y_{ij} = c_j \mid \eta_i, 3\,) = \sum_{h=1}^{k} \lambda_{hc_j}^{(j)} \eta_{ih} \tag{2}$$

$$\mathrm{pr}(y_{i1} = c_1, \dots, y_{ip} = c_p \mid \eta_i, 3\,) = \prod_{j=1}^{p} \mathrm{pr}(y_{ij} = c_j \mid \eta_i, 3\,), \tag{3}$$

where $3\; = (\;{}^{(j)}_{h})$. We refer to the model defined in Equations (2) and (3) as a simplex factor model. The formulation is conceptually related to mixed membership models (Pritchard, Stephens, and Donnelly 2000; Barnard et al. 2003; Blei, Ng, and Jordan 2003; Erosheva, Fienberg, and Joutard 2007), which have found widespread applications in text modeling, population genetics, and machine learning. In particular, the latent Dirichlet allocation (LDA) model (Blei et al. 2003) for text modeling arises as a special case of our model when $p = 1$. We can think of ${}^{(j)}_{h}$ as the vector of probabilities of $y_{ij} = 1, 2, \dots, d_j$, respectively, in ancestral population or pure species h, with none of the individuals being pure and η_{ih} being the weight on the hth component for the ith individual. When $k = 1$, the simplex factor model reduces to the product multinomial model representing global independence. As k increases, the complexity of the model increases.

To obtain further insight, we represent the simplex factor model in the following hierarchical form, which will also be used for posterior computation:

$$y_{ij} \sim \text{multinomial}\big(\{1,\ldots,d_j\}, \lambda^{(j)}_{z_{ij}1},\ldots,\lambda^{(j)}_{z_{ij}d_j}\big),$$
$$\text{pr}(z_{ij}=h)=\eta_{ih},\ h=1,\ldots,k. \qquad (4)$$

Clearly, marginalizing out z_{ij} in (4) gives (2). The z_{ij}'s can be considered as local latent class indices for the jth variable and the ith subject. The simplex factor model allows these local indices z_{ij}'s to vary across j for a particular subject, resulting in a more flexible and parsimonious (in terms of number of components k) specification compared with that in Dunson and Xing (2009), where a univariate latent class index is used.

Marginalizing out η_i in Equation (3) induces dependence among the y_{ij}'s. Letting $\pi_{c_1\ldots c_p} = \text{pr}(y_{i1}=c_1,\ldots,y_{ip}=c_p \mid 3\,)$, one has

$$\pi_{c_1\ldots c_p} = \int \prod_{j=1}^{p} \text{pr}(y_{ij}=c_j \mid \eta_i, 3\,)\,d\mathsf{Q}(\eta_i)$$
$$= \sum_{h_1=1}^{k}\cdots\sum_{h_p=1}^{k} g_{h_1\ldots h_p}\prod_{j=1}^{p}\lambda^{(j)}_{h_jc_j}, \qquad (5)$$

with Q denoting the distribution of η_i and $g_{h_1,\ldots,h_p} = E_{\mathsf{Q}}(\eta_{ih_1},\ldots,\eta_{ih_p})$.

2.2 Relationship With Tensor Decompositions

It is useful at this point to consider relationships between (5) and the literature on tensor decompositions, which shall be used in particular to illustrate the differences between our model and the Dunson and Xing (2009) model. Let $\mathsf{T}_{d_1\ldots d_p}$ denote the set of all tensors of dimension $d_1 \times \cdots \times d_p$, and $5\ _{d_1\ldots d_p} \subset \mathsf{T}_{d_1\ldots d_p}$ denote the set of probability tensors so that $\ \in 5\ _{d_1\ldots d_p}$ implies

$$= \Big\{\pi_{c_1\ldots c_p} \geq 0,\ c_j = 1,\ldots,d_j,$$
$$j=1,\ldots,p\ :\ \sum_{c_1=1}^{d_1}\cdots\sum_{c_p=1}^{d_p}\pi_{c_1\ldots c_p}=1\Big\}.$$

A decomposed tensor (Kolda 2001) is a tensor $\mathbf{D} \in \mathsf{T}_{d_1\ldots d_p}$ such that $\mathbf{D} = \mathbf{u}^{(1)} \otimes \mathbf{u}^{(2)},\ldots,\otimes\mathbf{u}^{(p)}$, where $\mathbf{u}^{(j)} \in \Re^{d_j}$ and $\otimes$ denotes the outer product so that $D_{c_1,\ldots,c_p} = u^{(1)}_{c_1}u^{(2)}_{c_2},\ldots,u^{(p)}_{c_p}$. One notion of the rank of a tensor is the minimal r such that $\mathbf{D}$ can be expressed as a sum of r decomposed (or rank 1) tensors. Such a decomposition is often referred to as a PARAFAC decomposition (Harshman 1970), which is one way of generalizing the matrix SVD. Tucker (1966) proposed a different decomposition for three-way data, which was later extended to arbitrary tensors by De Lathauwer et al. (2000). The Tucker decomposition or HOSVD aims to decompose a tensor $\mathbf{D} \in \mathsf{T}_{d_1\ldots d_p}$ as

$$D_{c_1\ldots c_p} = \sum_{h_1=1}^{d_1}\cdots\sum_{h_p=1}^{d_p} g_{h_1,\ldots,h_p}u^{(1)}_{h_1c_1},\ldots,u^{(p)}_{h_pc_p}, \qquad (6)$$

where $\mathbf{G} = \{g_{h_1\ldots h_p}\} \in \mathsf{T}_{d_1\ldots d_p}$ is called a core tensor and its entries control interaction between the different components. Wang and Ahuja (2005) and Kim and Choi (2007) empirically noted that the HOSVD achieves better data compression and requires fewer components compared with the PARAFAC model as it uses all combinations of the mode vectors $u^{(j)}_h$'s, $h=1,\ldots,k$.

The product multinomial mixture model of Dunson and Xing (2009) induces a decomposition of a probability tensor as

$$\pi_{c_1\ldots c_p} = \sum_{h=1}^{k}\nu_h\lambda^{(1)}_{hc_1},\ldots,\lambda^{(p)}_{hc_p}, \qquad (7)$$

where $\nu_h = \text{pr}(z_i = h)$ and $\ ^{(j)}_{h} \in \mathsf{S}_{d_j-1}$. Note that (7) is different from a usual PARAFAC decomposition because of the nonnegativity constraints on and the $\ ^{(j)}_{h}$'s. In the subsequent discussion, a nonnegativity matrix/tensor has all entries nonnegativity. The classical nonnegativity matrix factorization (NMF) problem seeks the best approximation of a nonnegative matrix $A \in \Re^{m\times n}_{+}$ as a product of nonnegative matrices $W \in \Re^{m\times k}_{+}$ and $V \in \Re^{k\times n}_{+}$ for some $k \leq \min\{m,n\}$. Gregory and Pullman (1983) were among the first to consider NMF and introduced the notion of nonnegative rank of a matrix, which is the minimal r such that a nonnegative matrix can be written as a sum of rank 1 nonnegative matrices. Cohen and Joel (1993) generalized many properties of the usual rank to the case of nonnegative rank. Along the lines of NMF, one can similarly envision nonnegative versions of the PARAFAC and HOSVD decompositions for tensors (Shashua and Hazan 2005; Kim and Choi 2007).

We note that (7) is a form of nonnegative PARAFAC decomposition, while the simplex factor model in (5) induces a nonnegative HOSVD on the space of probability tensors. Let $\ \in \mathsf{T}^{+}_{d_1\ldots d_p}$ be a nonnegative tensor. Define the nonnegative PARAFAC rank $\text{r}^{+}_{\text{PF}}(\ )$ of to be the minimum k such that admits a decomposition as in (7) with $\ ^{(j)}_{h} \in \mathsf{S}_{d_j-1}$ and $\ \in \Re^{k}$. Similarly, define the nonnegative HOSVD rank $\text{r}^{+}_{\text{HS}}(\ )$ of to be the minimum k such that can be expressed as in (5) with $\ ^{(j)}_{h} \in \mathsf{S}_{d_j-1}$ and $\mathbf{G} \in \mathsf{T}^{+}_{k\ldots k}$. In considering a nonnegative decomposition, assuming $\ ^{(j)}_{h} \in \mathsf{S}_{d_j-1}$ is not restrictive since we can always scale nonnegative weights to lie to the simplex and adjust the scale in or $\mathbf{G}$. In the special case when is a probability tensor, $\ \in \mathsf{S}_{k-1}$ is a probability vector and $\mathbf{G}$ is a probability tensor.

If we start with $k = \text{r}^{+}_{\text{PF}}(\ )$ in (7), then we can clearly express as in (5) using the same k by simply letting $g_{h_1\ldots h_p} = \nu_h 1(h_1 = h,\ldots,h_p = h)$. Conversely, suppose we start with a nonnegative HOSVD of as in (5) with $k = \text{r}^{+}_{\text{HS}}(\ )$ and the core tensor $\mathbf{G} \in 5\ _{k\ldots k}$ having $\text{r}^{+}_{\text{PF}}(\mathbf{G}) = r$. Then, there exist $\mathbf{u}^{(j)}_{l} \in \mathsf{S}_{k-1}$, $l = 1,\ldots,r$ and $\mathbf{q} \in \mathsf{S}_{r-1}$ such that $g_{h_1\ldots h_p} = \sum_{l=1}^{r} q_l u^{(1)}_{lh_1},\ldots,u^{(p)}_{lh_p}$. Substituting this expression for $g_{h_1\ldots h_p}$ in (5), one has

$$\pi_{c_1\ldots c_p} = \sum_{h_1=1}^{k}\cdots\sum_{h_p=1}^{k} g_{h_1\ldots h_p}\lambda^{(1)}_{h_1c_1},\ldots,\lambda^{(p)}_{h_pc_p}$$
$$= \sum_{l=1}^{r} q_l\left\{\prod_{j=1}^{p}\left(\sum_{h=1}^{k}\lambda^{(j)}_{hc_j}u^{(j)}_{lh}\right)\right\} = \sum_{l=1}^{r} q_l v^{(1)}_{lc_1},\ldots,v^{(p)}_{lc_p}, \qquad (8)$$

where $v^{(j)}_{lc_j} = \sum_{h=1}^{k} \lambda^{(j)}_{hc_j} u^{(j)}_{lh}$. Thus, starting with a nonnegative HOSVD of , we have expanded it in nonnegative PARAFAC form. Clearly, $r \ge k$; otherwise, the minimality of k is contradicted. Moreover, very little is known about upper bounds on PARAFAC ranks of tensors. The most general result is for third-order tensors where the upper bound is $\mathrm{O}(k^2)$; hence, r can potentially be much larger than k, requiring very many parameters for the PARAFAC expansion compared with the HOSVD. We summarize the above facts in the following theorem.

Theorem 2.1. Let $\in 5_{d_1\ldots d_p}$ and let $k = \mathrm{r}^{+}_{\mathrm{HS}}(\ )$. Then, $k \le \mathrm{r}^{+}_{\mathrm{PF}}(\ ) \le \mathrm{r}^{+}_{\mathrm{PF}}(\mathbf{G})$, where $\mathbf{G}$ is a core tensor in the minimal HOSVD expansion of . Moreover, among all such minimal expansions in k components, if $\mathbf{G}$ has minimal nonnegative PARAFAC rank, then $\mathrm{r}^{+}_{\mathrm{PF}}(\ ) = \mathrm{r}^{+}_{\mathrm{PF}}(\mathbf{G})$.

From a statistical perspective, the decomposition in (8) shows that the simplex factor model can provide sparser solutions in scenarios where the Dunson and Xing (2009) model may require many components to adequately explain the dependence structure. The Dunson and Xing (2009) model induces a global clustering phenomenon by forcing all the variables for a particular subject to be allocated to the same cluster. This can lead to introduction of too many clusters to accommodate small idiosyncracies within the variables or the subjects might be inappropriately grouped together, obscuring local differences. The simplex factor model instead allows the different variables to be allocated to different clusters via the dependent local cluster indices z_{ij}.

2.3 Prior Specification and Properties

To complete a Bayesian specification of the simplex factor model, a natural choice is to draw the η_i's and the different $^{(j)}_{h}$'s from independent Dirichlet priors. We let $\eta_i \sim \mathrm{Diri}(\alpha\nu_1,\ldots,\alpha\nu_k)$, where $\alpha > 0$ and $= (\nu_1,\ldots,\nu_k)^{\mathrm{T}} = \mathrm{E}(\eta_i) \in \mathsf{S}_{k-1}$ is a vector of probabilities. To obtain a parsimonious representation in which the first few components (subpopulations) tend to be assigned most of the weight, we let $\nu_h = \nu^*_h \Pi_{l<h}(1-\nu^*_l)$ with $\nu^*_k = 1$ and place a beta$(1,\beta)$ prior on ν^*_h, $h = 1,\ldots,k-1$. This corresponds to the stick-breaking formulation (Sethuraman 1994) of the Dirichlet process truncated to the first k terms and is widely used for posterior computation in Dirichlet process mixture models (Ishwaran and James 2001). In this case, k can be viewed as an upper bound on the number of components as higher-indexed components will tend to have $\nu_h \approx 0$, so one obtains $\eta_{ih} \approx 0$ with high probability for all i and the k-component model adaptively collapses on a lower-dimensional model. In practice, such collapsing will be driven by the extent to which the data support a model with fewer components. The model can be expressed in hierarchical form as

$$
\begin{aligned}
&y_{ij} \sim \mathrm{multinomial}\big(\{1,\ldots,d_j\}, \lambda^{(j)}_{z_{ij}1},\ldots,\lambda^{(j)}_{z_{ij}d_j}\big),\\
&\ ^{(j)}_{h} \sim \mathrm{Diri}(a_{j1},\ldots,a_{jd_j}),\ \mathrm{pr}(z_{ij}=h) = \eta_{ih},\\
&\eta_i \sim \mathrm{Diri}(\alpha\ \ ),\ \nu_h = \nu^*_h \prod_{l<h}(1-\nu^*_l),\ \nu^*_h \sim \mathrm{beta}(1,\beta). \qquad (9)
\end{aligned}
$$

The hierarchical prior specification on η_i has similarities to a finite-dimensional version of the hierarchical Dirichlet process (Teh et al. 2006), although the motivation here is slightly different. Essentially, we have a Dirichlet random-effects model with an unknown mean for the subject-specific random effects η_i's, with dependence being induced by marginalizing out the η_i's.

The Dirichlet latent factor distribution allows evaluation of $g_{h_1\ldots h_p} = \mathrm{E}_{\mathsf{Q}}(\eta_{ih_1}\ldots\eta_{ih_p})$ in (5) analytically

$$
g_{h_1\ldots h_p} = \frac{\Gamma(\alpha)}{\Gamma(\alpha+p)} \prod_{h=1}^{k} \frac{\Gamma\{\alpha\nu_h + \tau_h(h_1,\ldots,h_p)\}}{\Gamma(\alpha\nu_h)}, \qquad (10)
$$

where $\tau_h(h_1,\ldots,h_p) = \{\#j : h_j = h\}$ for $h = 1,\ldots,k$. When evident from the context, we shall drop the arguments and use τ_h. Clearly, $\sum_{h=1}^{k}\tau_h = p$.

With $\pi^{(j)}_{c_j} = \mathrm{pr}(y_{ij} = c_j \mid \ , 3\ )$ and $\pi^{(jj')}_{c_jc_{j'}} = \mathrm{pr}(y_{ij} = c_j, y_{ij'} = c_{j'} \mid \ , 3\ )$, one has

$$
\pi^{(j)}_{c_j} = \sum_{h=1}^{k} \nu_h \lambda^{(j)}_{hc_j} \qquad (11)
$$

$$
\begin{aligned}
\pi^{(jj')}_{c_jc_{j'}} &= \frac{\alpha}{\alpha+1}\left(\sum_{h=1}^{k}\nu_h\lambda^{(j)}_{hc_j}\right)\left(\sum_{h=1}^{k}\nu_h\lambda^{(j')}_{hc_{j'}}\right)\\
&\quad + \frac{1}{\alpha+1}\sum_{h=1}^{k}\nu_h\lambda^{(j)}_{hc_j}\lambda^{(j')}_{hc_{j'}}. \qquad (12)
\end{aligned}
$$

Now let us consider a few limiting cases; the main results are summarized below.

Proposition 2.2. For any fixed k, (i) in the limit as $\alpha \to \infty$, the simplex factor model (9) simplifies to a product multinomial model with $\mathrm{pr}(y_{ij} = c_j) = \sum_{h=1}^{k}\nu_h\lambda^{(j)}_{hc_j}$ independently for $j = 1,\ldots,p$; and (ii) in the limit as $\alpha \to 0$, model (9) simplifies to the Dunson and Xing (2009) model.

Thus, by putting a hyperprior on α, we can allow the data to inform about α and the posterior to concentrate near either of these two simplifications in cases where the simple structure is warranted. Next, we show that the proposed prior has large support on the space of probability tensors, so any dependence structure can be accurately approximated.

Theorem 2.3. Let $\mathsf{Q}^{(k)}_{\pi}$ denote the prior induced on $5_{d_1\ldots d_p}$ through the k-component simplex factor model in (9) and $\mathsf{N}_{\epsilon}(\ ^0)$ denote an L_1 neighborhood around an arbitrary probability tensor $^0 \in 5_{d_1\ldots d_p}$. Then, for any $_0 \in 5_{d_1\ldots d_p}$ and $\epsilon > 0$, there exists k such that $\mathsf{Q}^{(k)}_{\pi}\{\mathsf{N}_{\epsilon}(\ ^0)\} > 0$.

Since the space of probability tensors is isomorphic to a compact Euclidean space, a straightforward extension of theorem 4.3.1 of Ghosh and Ramamoorthi (2003) ensures that the posterior concentrates in arbitrary small neighborhoods of any true data-generating distribution ^{0} with increasing sample size.

3. POSTERIOR COMPUTATION AND INFERENCE

3.1 MCMC Algorithm for Posterior Computation

Let $\mathbf{y} = (y_{ij})$, $= (\eta_i)$, and $\mathbf{z} = (z_i)$. We use a combination of Gibbs sampling and independence chain Metropolis–Hastings sampling to draw samples from the posterior distribution of $(3\ , \mathbf{z},\ ,\ ^*, \alpha)$ for the hierarchical model specified in (9). We place a gamma(a_α, b_α) prior on α to allow the data to inform more strongly about sparsity in the η_i vectors. In particular, for small α, the tendency will be to assign one element of η_i to a value close to 1, while for larger α, the η_i vectors will be closer to for different subjects. We recommend $a_\alpha = b_\alpha = 1$ as a default value favoring high weights on few components. In addition, we let $a_{j1} = \cdots = a_{jd_j} = 1$, for $j = 1, \ldots, p$, to induce a uniform prior for the category probabilities in each class for each outcome type. This default prior specification can be modified in cases in which one has prior information on the category probabilities and/or the number of subpopulations.

The conditional posteriors for all the parameters other than * and α can be derived in closed form using standard algebra and the sampler cycles through the following steps:

Step 1. For $h = 1, \ldots, k$, update ${}^{(j)}_{h}$ from the following Dirichlet full conditional posterior distribution

$$\pi\left({}^{(j)}_{h} \mid -\right) \sim \text{Diri}\left(a_{j1} + \sum_{i:z_{ij}=h} 1(y_{ij}=1), \ldots, a_{jd_j} + \sum_{i:z_{ij}=h} 1(y_{ij}=d_j)\right).$$

Step 2. Update z_{ij} from the multinomial full conditional posterior distribution, with

$$\text{pr}(z_{ij} = h \mid -) = \frac{\eta_{ih}\lambda^{(j)}_{hy_{ij}}}{\sum_{l=1}^{k} \eta_{il}\lambda^{(j)}_{ly_{ij}}}.$$

Step 3. Update η_i from the Dirichlet full conditional posterior:

$$\pi(\eta_i \mid -) \sim \text{Diri}\left(\alpha\nu_1 + \sum_{j=1}^{p} 1(z_{ij}=1), \ldots, \alpha\nu_k + \sum_{j=1}^{p} 1(z_{ij}=k)\right).$$

Step 4. Update α using a Metropolis random walk on $\log(\alpha)$.

Step 5. Update $\{\nu_h^*\}$ using the following approach. Let $m_{il} = \sum_{j=1}^{p} 1(z_{ij} = l)$, $m_l = \sum_{i=1}^{n} m_{il}$, $m_{l+} = \sum_{i=1}^{n}\sum_{l'>l} m_{il'}$ and $n_{ls} = \{\#i : m_{il} > s\}$ for nonnegative integers s. Letting $m_l^* = \max_{1\le i\le n} m_{il}$, one has $n_{ls} = 0$ for $s \ge m_l^*$. Further, let $\tilde{\nu}_h^{(h)} = \Pi_{l<h}(1-\nu_l^*)$ and $\tilde{\nu}_l^{(h)} = \nu_l/(1-\nu_h^*)$ for $l > h$, and define $c_l^{(h)} = \alpha\tilde{\nu}_l^{(h)}$. The conditional posterior of ν_h^* marginalizing out the η_i's is given by

$$\begin{aligned}\pi(\nu_h^* \mid -) &\propto (1-\nu_h^*)^{\beta-1}\prod_{i=1}^{n}\prod_{l=h}^{k}\frac{\Gamma(\alpha\nu_l + m_{il})}{\Gamma(\alpha\nu_l)}\\ &= (1-\nu_h^*)^{\beta-1}\prod_{l=h}^{k}\prod_{i:m_{il}\neq 0}\prod_{s=0}^{m_{il}-1}(\alpha\nu_l + s). \qquad (13)\end{aligned}$$

We assume the default choice $\beta = 1$. Since the expression for ν_l contains ν_h^* for $l = h$ and $(1-\nu_h^*)$ for $l > h$, the conditional posterior of ν_h^* in (13) is an analytically intractable mixture of beta densities. However, we show that $\pi(\nu_h^* \mid -)$ can be accurately approximated by a single beta distribution. One can thus use an appropriate beta density as a proposal in a Metropolis–Hastings step, with the beta parameters estimated numerically on a fine grid via moment matching. However, the grid-based method is computationally costly, since the expression in (13) needs to be computed at every point on the grid. We propose an approach to provide analytic expressions for the parameters of the approximating beta density. The analytic solution produces high acceptance rates, and there is a dramatic gain in computational time, whose effect is increasingly pronounced with large n and/or p. We mention below the choices of the parameters of the approximating beta density in the different cases, with justification provided in the Appendix.

If $m_h > 0$ and $m_{h+} = 0$, we use a beta$(\hat{a}, 1)$ density with

$$\hat{a} = 1 + \sum_{s=0}^{m_h^*-1} n_{hs}\left\{1 - \frac{s}{c_h^{(h)}}\log\left(1 + c_h^{(h)}/s\right)\right\} \qquad (14)$$

to approximate $\pi(\nu_h^* \mid -)$. Similarly, if $m_h = 0$ and $m_{h+} > 0$, a beta$(1, \hat{b})$ density with

$$\hat{b} = 1 + \sum_{l=h+1}^{k}\sum_{s=0}^{m_l^*-1} n_{ls}\left\{1 - \frac{s}{c_l^{(h)}}\log\left(1 + c_l^{(h)}/s\right)\right\} \qquad (15)$$

is used to approximate $\pi(\nu_h^* \mid -)$.

If $m_h > 0$ and $m_{h+} > 0$, we prove the following fact.

Proposition 3.1. If $m_h > 0$ and $m_{h+} > 0$, then $\pi(\nu_h^* \mid -)$ is unimodal and $\lim_{\nu_h^*\to 0}\pi(\nu_h^* \mid -) = 0$, $\lim_{\nu_h^*\to 1}\pi(\nu_h^* \mid -) = 0$.

We approximate $\pi(\nu_h^* \mid -)$ by a beta$(\hat{a}, \hat{b})$ density in this case, where $\hat{a} = \max(\tilde{a} + 1, 1)$, $\hat{b} = \max(\tilde{b} + 1, 1)$, and $\tilde{a}, \tilde{b}$ are obtained by solving a 2×2 linear system $E(\tilde{a}, \tilde{b})^{\mathrm{T}} = (d_1, d_2)^{\mathrm{T}}$, with $e_{11} = 1/2$, $e_{12} = -1/2$, $e_{21} = 1/6$, $e_{22} = -1/3$ and

$$d_1 = \int_0^1 2\nu_h^* \log\pi(\nu_h^* \mid -)d\nu_h^* - \int_0^1 \log\pi(\nu_h^* \mid -)d\nu_h^*$$

$$d_2 = \int_0^1 3(\nu_h^*)^2 \log\pi(\nu_h^* \mid -)d\nu_h^* - \int_0^1 2\nu_h^* \log\pi(\nu_h^* \mid -)d\nu_h^*.$$

A one-step improvement is obtained next by running a mode search of the log posterior from the estimated $(\tilde{a}, \tilde{b})$ pair above and subsequently adjusting those values to have the right mode. Since $\pi(\nu_h^* \mid -)$ is unimodal in this case, the mode search can be done very efficiently using the Newton–Rapson algorithm.

3.2 Adaptive Selection of the Number of Factors

In practical problems, one typically expects a small number of factors k relative to the number of outcomes p. The stick-breaking prior on induces a sparse formulation a priori, so relatively few components with high weights are encouraged. Accordingly, one can start with a conservative upper bound k^* on the number of factors and the sparsity favoring prior ensures that the posterior will concentrate on a few components if the truth is approximately sparse. However, an overly conservative upper bound wastes substantial computational time. For Gaussian factor models, Lopes and West (2004) compared a number of alternatives to select the number of factors, recommending a reversible-jump MCMC approach that requires a preliminary run for each choice of the number of factors, which is very computationally intensive. Since we are not interested in inference on the number of factors and our sole consideration is computational efficiency, we outline below a simple adaptive scheme to discard the redundant factors in a tuning phase and continue with a smaller number of factors for the remainder of the chain. This results in considerable computational gains in high dimensions; however, for moderate values of p (e.g., $p \leq 20$), one can set p to be the truncation level and ignore the adaptive scheme.

We reserve T_{tune} many iterations at the beginning of the chain for tuning. At iteration t, letting $K_{\text{eff}}^{(t)} \subset \{1, \ldots, k\}$ to denote the unique values among the z_{ij}'s, one clearly has $m_h > 0$ if and only if $h \in K_{\text{eff}}^{(t)}$. We define the effective number of factors $\tilde{k}^{(t)}$ as $|K_{\text{eff}}^{(t)}|$. The inherent sparse structure of the simplex factor model favors small values of $\tilde{k}^{(t)}$. Starting with a conservative guess for k, we monitor the value of $\tilde{k}^{(t)}$ every 50 iterations in the tuning phase. If there are no redundant factors, that is, $m_h > 0$ for all h, we add a factor and initialize the additional parameters for , 3 , * from the prior. Otherwise, we delete the redundant components and retain the elements of , 3 , * corresponding to $h \in K_{\text{eff}}^{(t)}$. In either case, we normalize the samples for and * to ensure they lie on the simplex. We continue the chain with the updated number of factors and the modified set of parameters for the next 50 iterations before making the next adaptation. At the end of the tuning phase, we fix the number of factors for the remainder of the chain at the value corresponding to the last adaptation. In all our examples, we let $T_{\text{tune}} = 5000$ and choose the initial number of factors as 20 or 10.

3.3 Inference

One can estimate the marginal distribution of the y_{ij}'s and conduct inferences on the dependence structure based on the MCMC output and using the expressions for the lower-dimensional marginals in Equations (11) and (12). To conduct inference on dependence between y_{ij} and $y_{ij'}$ for $j \neq j' \in \{1, \ldots, p\}$, we consider the pairwise normalized mutual information matrix $M = (m_{jj'})$, with $m_{jj'} = I_{jj'}/\{H_j H_{j'}\}^{0.5}$ and

$$I_{jj'} = \sum_{c_j=1}^{d_j} \sum_{c_{j'}=1}^{d_{j'}} \pi_{c_j c_{j'}}^{(jj')} \log \left\{ \frac{\pi_{c_j c_{j'}}^{(jj')}}{\pi_{c_j}^{(j)} \pi_{c_{j'}}^{(j')}} \right\},$$

$$H_j = -\sum_{c_j=1}^{d_j} \pi_{c_j}^{(j)} \log \left\{ \pi_{c_j}^{(j)} \right\}.$$

The mutual information $I_{jj'}$ is a general measure of dependence between a pair of random variables $(Y_J, Y_{j'})$, with $I_{jj'} = 0$ if and only if Y_j and $Y_{j'}$ are independent. Using our Bayesian approach, one can obtain samples from the posterior distribution of $m_{jj'}$ for all $j \neq j'$ pairs. In particular, posterior summaries of the $p \times p$ association matrix can be used to infer on the association between pairs of variables accounting for uncertainty in other variables. Dunson and Xing (2009) pointed out that in large model spaces, it is more computationally tractable to consider pairwise marginal dependencies as compared with learning the entire graph of conditional dependencies. Dunson and Xing (2009) and Dobra and Lenkoski (2011) used the pairwise Cramer's V association matrix as a measure of dependence. We also computed the pairwise Cramer's V association matrix for all our examples and obtained similar dependence structures as found by the mutual information criterion. In the analysis of the Rochdale data in Section 4, we present results for both measures of association and in the subsequent simulated and real-data examples, we only provide the results for the mutual information criterion as the conclusions were similar.

In many practical examples, one routinely encounters a high-dimensional vector of nominal predictors y_i along with nominal response variables u_i, with interest in building predictive models for u_i given y_i. For example, y_i might correspond to a nucleotide sequence, with u_i being the existence/nonexistence of some special feature within the gene sequence. By using a simplex factor model for the joint distribution of the response and predictors, one obtains a very flexible approach for classification from categorical predictors that may have higher-order interactions. Such a joint model also trivially allows imputation of missing values under the missing at-random assumption or any other informative missingness.

4. ANALYSIS OF ROCHDALE DATA

Dobra and Lenkoski (2011) used copulas to extend traditional Gaussian graphical models to allow mixed outcomes. They applied their general class of copula Gaussian graphical models (CGGM) to analyze the Rochdale data, a 2^8 contingency table popular in the social science literature. In this section, we illustrate various aspects of making inference with the simplex factor model on this well-known dataset and compare our results with the CGGM.

The Rochdale dataset was previously analyzed in Whittaker (1990). It is a social survey dataset aimed to assess the relationship among factors influencing women's economic activity. The dataset consists of eight related binary variables coded $a, b, \ldots, h$ for $n = 665$ women. A detailed account of the different variables can be found in Dobra and Lenkoski (2011). The resulting 2^8 contingency table is sparse, with 165 cell counts of zero. The top 10 cell counts are all greater than 20.

Whittaker (1990) used log-linear models to analyze this dataset and argued against using higher-order interactions involving more than two variables. He considered two log-linear models, one being the all two-way interaction model, and the minimal sufficient statistics for the other one consisted of 14 two-way marginals in equation 5.1 of Dobra and Lenkoski (2011). Dobra and Lenkoski (2011) analyzed this dataset using their proposed CGGM and also compared it with a copula full model where the underlying graph was not updated and was fixed at the full graph.

The simplex factor model was run for 50,000 iterations, with the first 30,000 iterations discarded as burn-in and every fifth sample post burn-in was collected. We started with 10 factors and the adaptive algorithm selected four factors, with more than 85% acceptance rate for all elements of *. The posterior mean of α was 0.10, with a 95% credible interval of (0.03–0.20).

According to Whittaker (1990), the strongest pairwise interaction in this dataset is for the pair (b, d), followed by (b, h), (e, f), and (a, g). Dobra and Lenkoski (2011) obtained strongest associations for the pairs (b, d), (b, h), (a, g), (e, f), and (c, g) according to Cramer's V statistic. They also noted that conditioning on the full graph in the copula full model leads to severe underestimation of all the pairwise Cramer's V associations. From the MCMC output, we obtained posterior samples for the pairwise normalized mutual information and the pairwise Cramer's V. Figure 1 shows posterior summaries of the

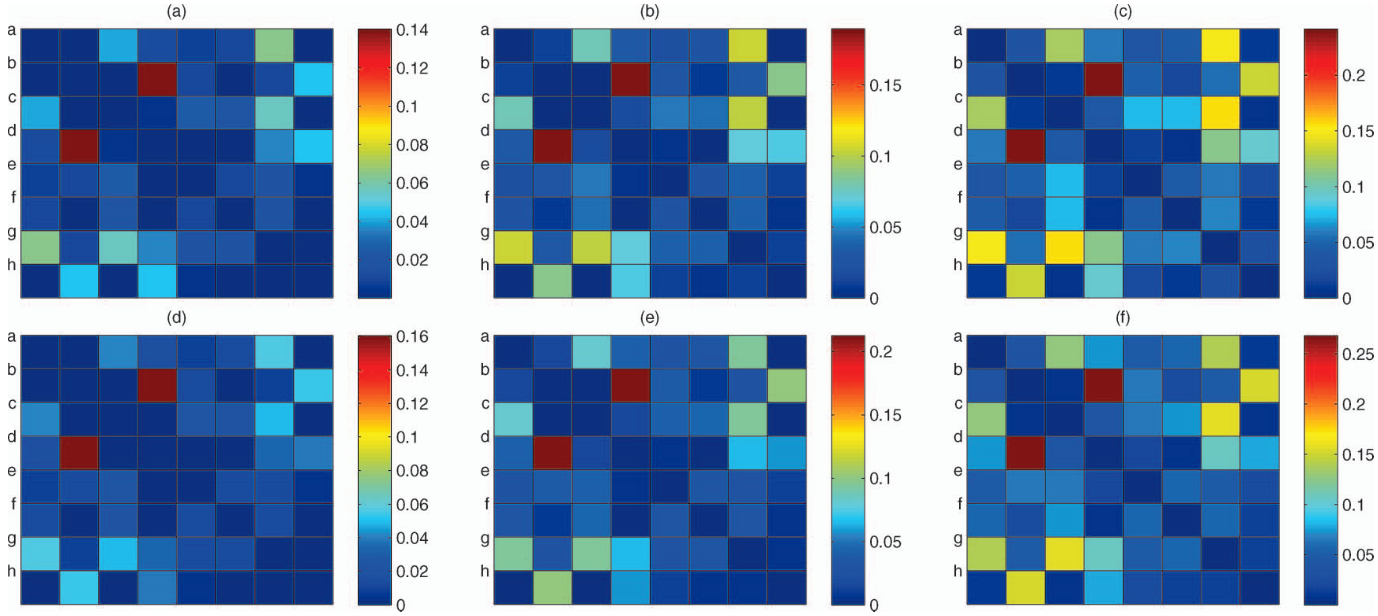

Figure 1. Results for Rochdale data—posterior means (second column), and 2.5 and 97.5 percentiles (first and third column, respectively) of the pairwise normalized mutual information matrix (top row) and the pairwise Cramer's V association matrix (bottom row). (The online version of this figure is in color.)

pairwise associations for these two measures. From Figure 1, it is evident that both measures obtained a similar dependence structure and the top four pairs according to either of the two measures were (b, d), followed by (b, h), (a, g), and (c, g). The posterior means of the Cramer's V values in Figure 1(e) for these four pairs were 0.21, 0.10, 0.09, and 0.09, respectively, which are very similar to those obtained by the CGGM (Dobra and Lenkoski 2011, table 5.5). The variable a denotes wife's economic activity and is of interest in determining the variables that share association with a. The variables having largest Cramer's V association to a were g, c, and d, with the posterior means for the pairs (a, g), (a, c), and (a, d) given by 0.09, 0.08, and 0.04, respectively. Again, the same ordering was discovered by the CGGMs. Overall, our results were very much in agreement with those obtained by the CGGMs, with the only notable difference being (a, c) ranking over (e, f) for both the measures in our case.

The MCMC was also run with the set of 10 factors for the entire length of the chain with a beta$(1, 1)$ prior on β. The results were robust, with the same ordering of the pairwise Cramer V's obtained as in the previous case. As discussed before, the stick-breaking prior on the ν_h^* drives the ν_h's for the redundant components close to zero, thereby making the procedure robust with respect to the choice of the number of factors as long as there are sufficiently many factors.

5. SIMULATION STUDY

We considered two simulation scenarios to assess the performance of the simplex factor model. We simulated $y_{ij} \in \{A, C, G, T\}$ at $p = 50$ locations and a nominal response u_i having two and three levels in the two simulation cases, respectively. We considered two pure species $(k = 2)$ and simulated the local subpopulation indices z_{ij} as in (9) to induce dependence among the response and a subset of locations $J = (2, 4, 12, 14, 32, 34, 42, 44)^{\mathrm{T}}$. The eight locations in J had different A, C, G, T probabilities in the two pure species, while the phenotype probabilities at the remaining 42 locations were chosen to be the discrete uniform distribution on four points in each species.

In the first simulation scenario, we considered $n = 100$ sequences and two randomly chosen subpopulations of sizes 60 and 40, respectively. All the z_{ij}'s were assigned a value of 1 in the first subpopulation, and 2 in the second one. Within each subpopulation, the nucleotides were drawn independently across locations, with the jth nucleotide having phenotype probabilities $(\lambda_{z_{ij}A}^{(j)}, \lambda_{z_{ij}C}^{(j)}, \lambda_{z_{ij}G}^{(j)}, \lambda_{z_{ij}T}^{(j)})^{\mathrm{T}}$. The binary response $(u_i \in \{1, 2\})$ had category probabilities $(0.92, 0.08)$ and $(0.08, 0.92)$ in the two subpopulations, respectively.

The second scenario had a more complicated dependence structure. We considered 200 sequences and three subpopulations of sizes 80, 80, and 40, respectively, with all the local indices z_{ij} assigned a value of 1 and 2, respectively, in the first two subpopulations. In the third subpopulation, the z_{ij}'s for the first 30 locations were assigned a value of 1 $(z_{ij} = 1, j = 1, \ldots, 30)$, and the remaining 20 locations a value of 2 $(z_{ij} = 2, j = 31, \ldots, 50)$. The response variable had three categories in this case, with category probabilities $(0.90, 0.05, 0.05)$, $(0.05, 0.90, 0.05)$, and $(0.05, 0.05, 0.90)$ in the three subpopulations, respectively. The third subpopulation is biologically motivated as a rare group that has local similarities with each of the other groups, and thus, is difficult to distinguish from the other two.

For each case, we generated 50 simulation replicates and the simplex factor model was fitted separately to each dataset using the MCMC algorithm mentioned in Section 3.1. The sampler was run for 30,000 iterations, with a burn-in of 10,000 and every fifth sample was collected. We obtained good mixing and convergence for the elements of the pairwise mutual information matrix based on examination of trace plots. Figure 2 shows the

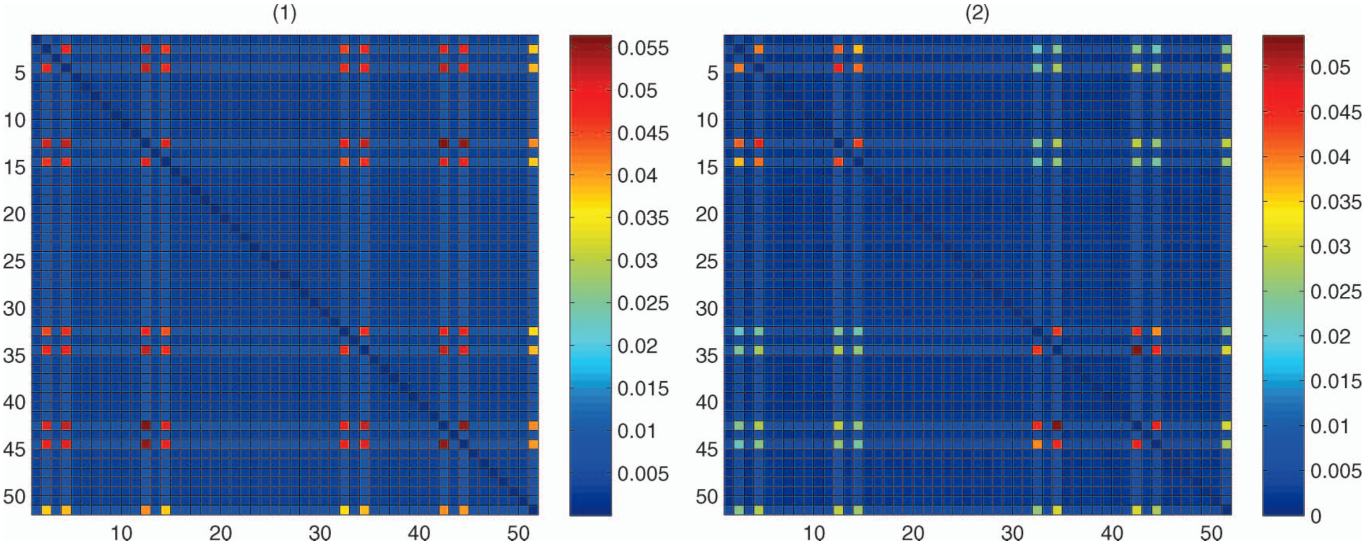

Figure 2. Simulation results—posterior means of the pairwise normalized information matrix of the nominal variables in simulation cases 1 and 2. Dependence between locations (2, 4, 12, 14, 32, 34, 42, 44) and the response (last row/column). (The online version of this figure is in color.)

posterior means of the pairwise normalized mutual information matrix $m_{jj'}$ averaged across simulation replicates in the two simulation scenarios, with the last row/column corresponding to the response variable. The posterior means corresponding to all $\binom{9}{2} = 36$ dependent location pairs are clearly well separated from the remaining nondependent pairs. We also provide kernel density plots of the posterior means and upper 97.5 percentiles of the $m_{jj'}$ across the 50 replicates for the two simulation cases in Figure 3. For the first simulation case (top row), the density plot for the posterior means in Figure 3(a) is bimodal, with a tall spike near zero and a very heavy right tail, thus showing a clear separation between the dependent and the nondependent pairs. The second simulation also had a very heavy tail for the posterior means in Figure 3(c), and the second mode is visible for the upper quantile in Figure 3(d).

Next, we aim to assess out-of-sample predictive performance for the simplex factor model. Since the subpopulations were chosen in a random order, we chose the first 20 samples in simulation case 1 and the first 30 samples in simulation case 2 as training sets within each replicate. We compared our

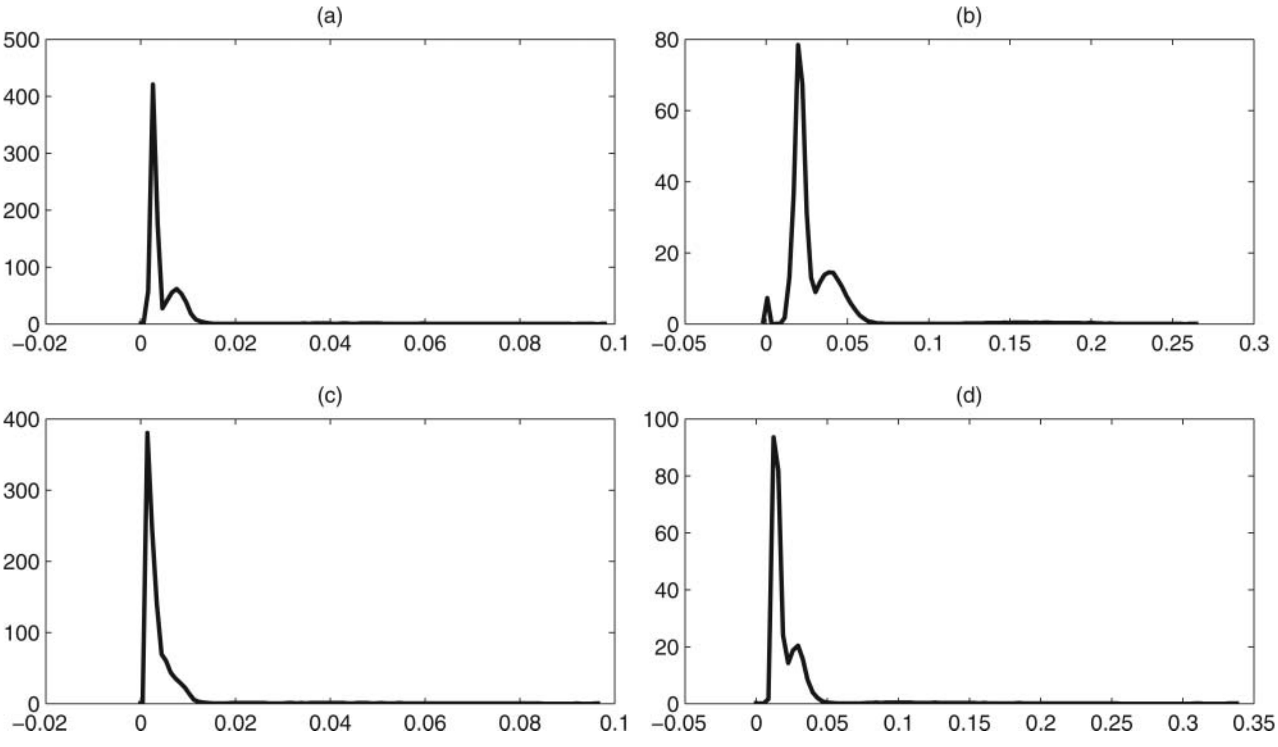

Figure 3. Simulation results—density plots of the posterior means (first column) and upper 97.5 percentiles (second column) of the p.w. normalized mutual information $m_{jj'}$'s across simulation replicates in the two simulation cases.

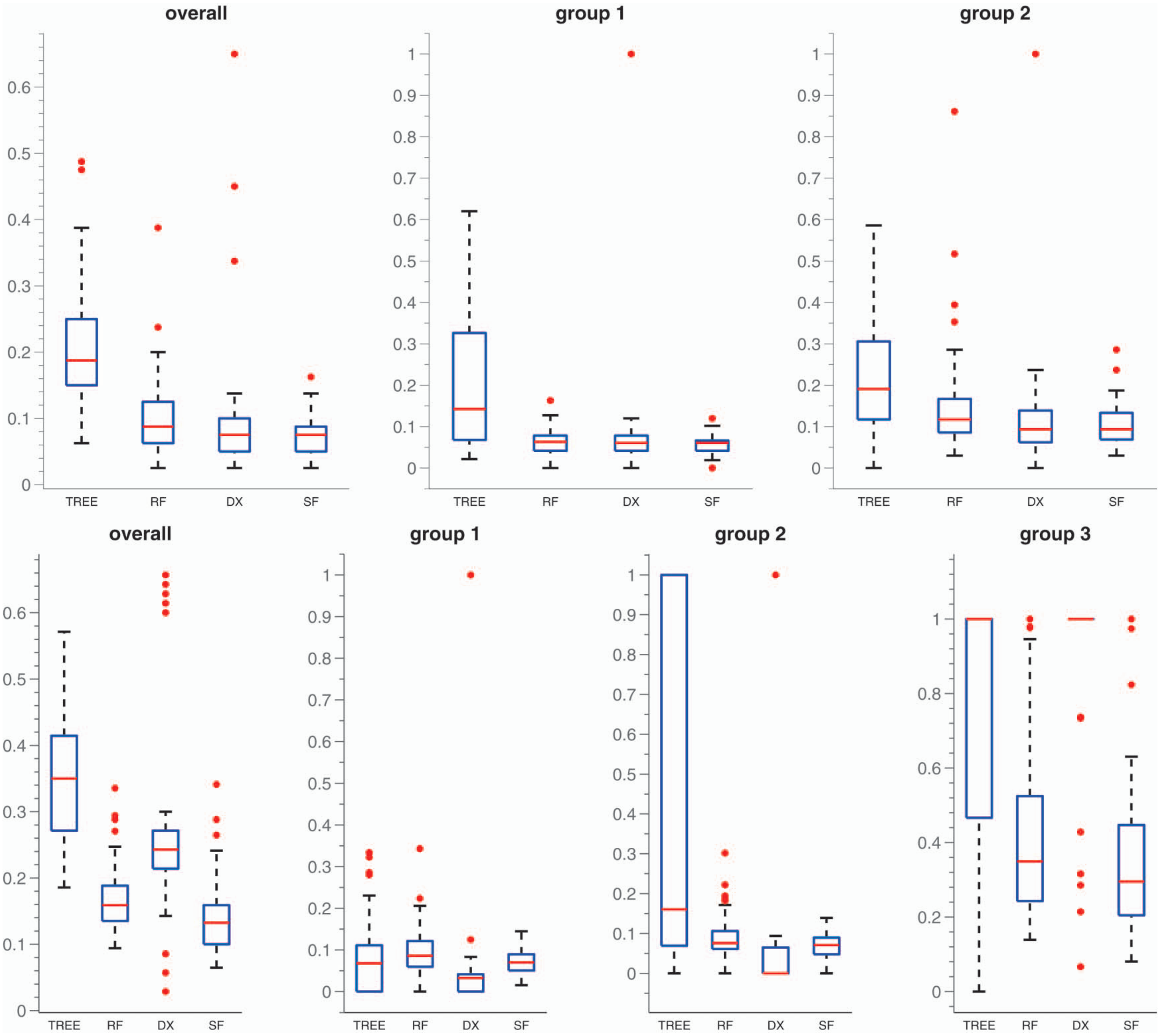

Figure 4. Box plots for misclassification proportions across simulation replicates for the different methods. TREE = tree-based classifier, RF = random forest classifier, DX = Dunson–Xing method, SF = simplex factor model. The top row corresponds to simulation case 1, bottom row to simulation case 2. The first column corresponds to overall misclassification proportion, the remaining columns indicate group-specific misclassification proportions. (The online version of this figure is in color.)

approach with the Dunson and Xing (2009) method, a tree classifier built in MATLAB, and the random forest ensemble classifier (Breiman 2001), which was implemented using the `RandomForest` package (Liaw and Wiener 2002) in R. We did not consider a fully Bayesian graphical modeling approach, such as Dobra and Lenkoski (2011), because such methods do not scale computationally to the sized contingency tables we are considering. Figure 4 shows box plots of the overall and group-specific misclassification proportions for the different methods across the simulation replicates. As expected, the misclassification rates for simulation case 2 are higher than simulation case 1. The misclassification percentages corresponding to the second category are slightly larger than those for the first category in simulation case 1, which is explained by the relative sizes of the two subpopulations. It is clear from Figure 4 that the simplex factor model had better performance than the tree classifier and the random forest classifier in both cases. The overall misclassification percentage and category-specific misclassification percentages for the simplex factor model and random forest in the two simulation scenarios are provided in Table 1.

Simulation case 1 was designed to comply with the Dunson and Xing (2009) model, since all of the variables for a particular subject were assigned to the same subpopulation. Accordingly, the Dunson and Xing (2009) method had very similar performance compared with the simplex factor model in the first case and did better than the other two methods. However, the performance of the Dunson and Xing (2009) method deteriorated in simulation case 2. In this case, the subjects in the third group

Table 1. Misclassification percentages in the two simulation cases for the simplex factor model and random forest

	Simulation case 1		Simulation case 2	
	Simplex factor	Random forest	Simplex factor	Random forest
Best	(2.50, 0, 3.03)	(2.50, 0, 3.03)	(6.47, 1.54, 0, 8.10)	(8.82, 0, 0, 13.89)
Average	(7.55, 5.45, 10.39)	(9.90, 5.84, 15.51)	(14.08, 7.15, 6.80, 37.07)	(16.67, 9.30, 9.18, 41.03)
Worst	(16.25, 12.00, 28.57)	(35.00, 16.32, 77.77)	(34.12, 14.49, 13.89, 100)	(34.70, 27.14, 31.74, 100)

NOTE: Each vector represents overall and category-specific misclassification percentages, with the categories arranged according to their index. Best-, average-, and worst-case performances across replicates are reported.

had mixed membership for the different variables and the misclassification rates for this group were particularly high for the Dunson and Xing (2009) method.

6. NUCLEOTIDE SEQUENCE APPLICATIONS

We first applied our method to the p53 transcription factor-binding motif data (Wei et al. 2006). The data have $n = 574$ DNA sequences consisting of A, C, G, T nucleotides ($d_j = 4$) at $p = 20$ positions. Transcription factors are proteins that bind to specific locations within DNA sequences and regulate copying of genetic information from the DNA to the mRNA. p53 is a widely known tumor suppressor protein that regulates expression of genes involved in a variety of cellular functions. It is of substantial biological interest to discover positional dependence within such DNA sequences.

We ran the MCMC algorithm for 25,000 iterations after the tuning phase, with a burn-in of 10,000 and collected every fifth sample. The adaptive algorithm selected four factors and we obtained acceptance rates greater than 90% for all the elements of * using our proposed independence sampler. Using a gamma(0.1, 0.1) prior for α, the posterior mean of α was 0.11, with a 95% credible interval of (0.08–0.16). From Proposition 2.2, the small value of α indicates that the model favors a simpler dependence structure, as in Dunson and Xing (2009). We actually obtained very similar results to the Dunson and Xing (2009) method for this particular dataset. Figure 5(a)–(c) shows the posterior means and quantiles for the simplex factor model. Figure 5(d)–(f) shows the same for the Dunson and Xing (2009) method. Clearly, the dependence structure is sparse and the strongest dependence are found near the center of the sequence, with position pairs (11, 12) and (9, 11) having the largest normalized mutual information using both methods. The Xie and Geng (2008) approach flagged all 190 pairs as dependent using a p-value of 0.01 or 0.05 for edge inclusion; such overfitting can typically occur for Bayes Nets unless the threshold on edge inclusion is very carefully chosen.

We next applied our method to the promoter data (Frank and Asuncion 2010) publicly available at the UCI Machine Learning Repository. The data consist of A, C, G, T nucleotides at $p = 57$ positions for $n = 106$ sequences, along with a binary response indicating instances of promoters and nonpromoters. There are 53 promoter sequences and 53

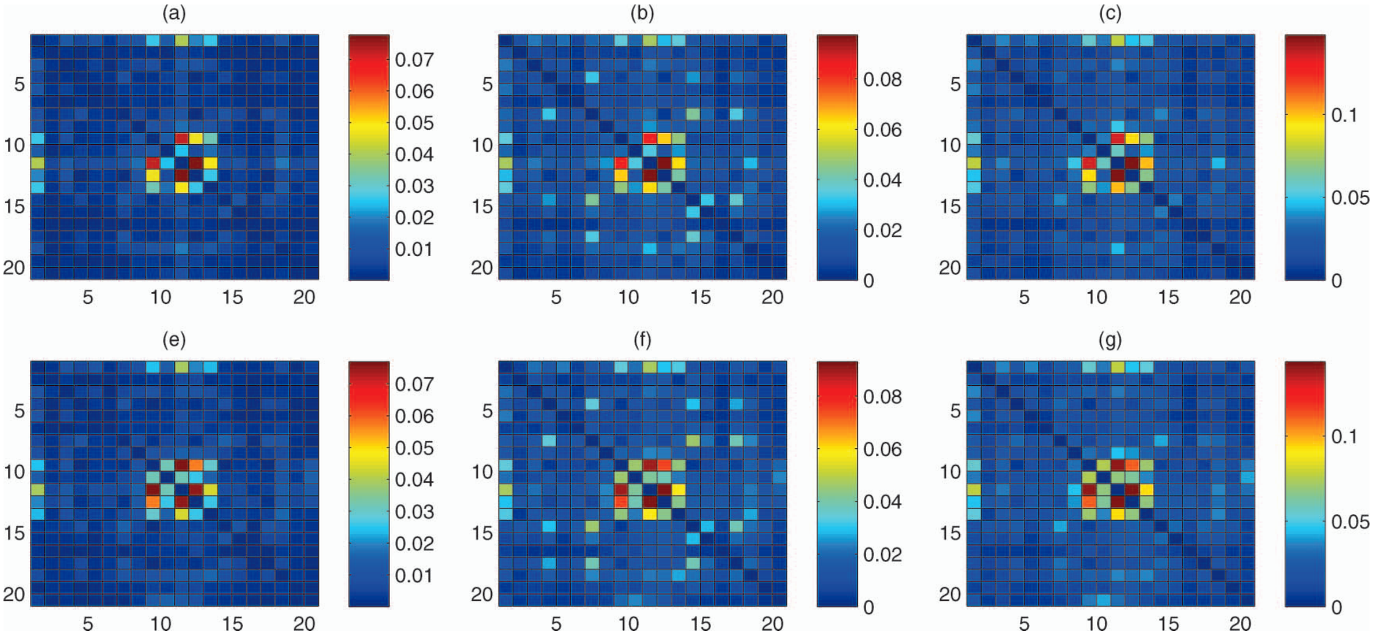

Figure 5. Results for the p53 data—posterior means (second column), and 2.5 and 97.5 percentiles (first and third column, respectively) of the normalized mutual information matrix. The top row corresponds to the simplex factor model, the bottom row is for the Dunson and Xing (2009) method. (The online version of this figure is in color.)

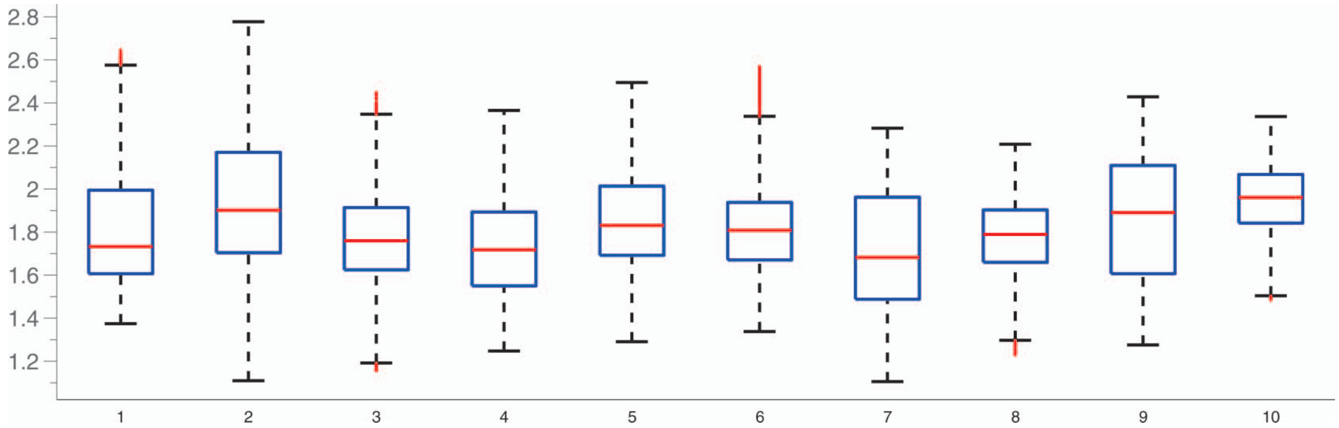

Figure 6. Results for the promoter data—box plots for posterior samples of α for the 10 replicates in the first case, where training size is 20% of the sample size. The red line denotes the posterior median, and the edges of the box denote posterior 25th and 75th percentiles. (The online version of this figure is in color.)

nonpromoter sequences in the dataset. We selected six training sizes as 20%, 30%, 40%, 50%, 60%, and 70% of the sample size n. For each training size, we randomly selected 10 training samples and evaluated the overall misclassification percentage and misclassification percentages specific to the promoter and nonpromoter groups. We compared out-of-sample predictive performance with the Dunson and Xing (2009) method, random forest, and a tree classifier. The MCMC algorithm for the simplex factor model was run for 30,000 iterations, with five factors selected. Our proposed approach for updating the ν_h^*'s again produced high acceptance rates; the mean acceptance rates averaged across all replicates for $\nu_1^* - \nu_4^*$ were $0.85, 0.92, 0.96,$ and 0.97, respectively. The posterior means for α across all replicates within each training size was greater than 1; we have provided box plots for the posterior samples of α across the 10 replicates corresponding to the smallest training size in Figure 6.

The simplex factor model had superior performance compared with the other three methods across all training sizes. In particular, for the training size $= 20$, the average misclassification percentages (overall, promoters, nonpromoters) for the Dunson and Xing (2009) method and random forest were $(23.41, 26.24, 19.69)$ and $(25.53, 29.92, 19.30)$, respectively, while the same for our method were $(14.35, 15.41, 12.16)$. Table 2 provides the average-, best-, and worst-case misclassification percentages for the simplex factor model and random forest corresponding to the smallest and largest training sizes. From Table 2, the misclassification percentage for the nonpromoter group was smaller compared with that of the promoter group. For the smallest training size of 21, the simplex factor model provides an improvement of more than 14% in terms of the misclassification percentages for the promoter group. We also plot the average-, best-, and worst-case overall misclassification proportions across different training sizes for the competing methods in Figure 7 and the average misclassification proportions specific to the promoter and nonpromoter groups in Figure 8. It is clearly seen from Figures 7 and 8 that the simplex factor model provides the best performance across all training sizes.

We performed sensitivity analysis for the prior on α by choosing a gamma$(1, 1)$ prior instead of gamma$(0.1, 0.1)$, with the results unchanged. We also multiplied and divided the prior mean by a factor of 2 and did not observe any notable changes. For ν_h^*, the uniform prior was found to be a reasonable default choice, as in most practical cases, one expects to have few dominant components in the case of contingency tables. That said, one can alternatively place a beta$(1, \beta)$ prior on ν_h^*, with a gamma prior assigned to β.

The additional flexibility of our model over the nonparametric Bayes method of Dunson and Xing (2009) produces improved performance in classification for the promoter data, and our approach also does better than sophisticated frequentist methods such as the tree classifier and random forest. We also applied our classification method to the splice data (publicly available at the UCI Machine Learning Repository) and obtained similar conclusions.

Table 2. Results for the promoter data—misclassification percentages (overall, among promoters, among nonpromoters) for the smallest and largest training sizes for the simplex factor model and random forest

	Training size = 21		Training size = 74	
	Simplex factor	Random forest	Simplex factor	Random forest
Best	(8.24, 0, 0)	(15.29, 0, 0)	(0, 0, 0)	(6.25, 0, 0)
Average	(14.35, 15.41, 12.16)	(25.53, 29.92, 19.30)	(7.81, 10.14, 4.79)	(12.50, 17.84, 6.37)
Worst	(30.59, 53.32, 35.56)	(44.71, 80.85, 64.44)	(15.62, 21.05, 17.65)	(15.62, 33.33, 23.53)

NOTE: Best-, average-, and worst-case performances across 10 training samples for each training size are reported.

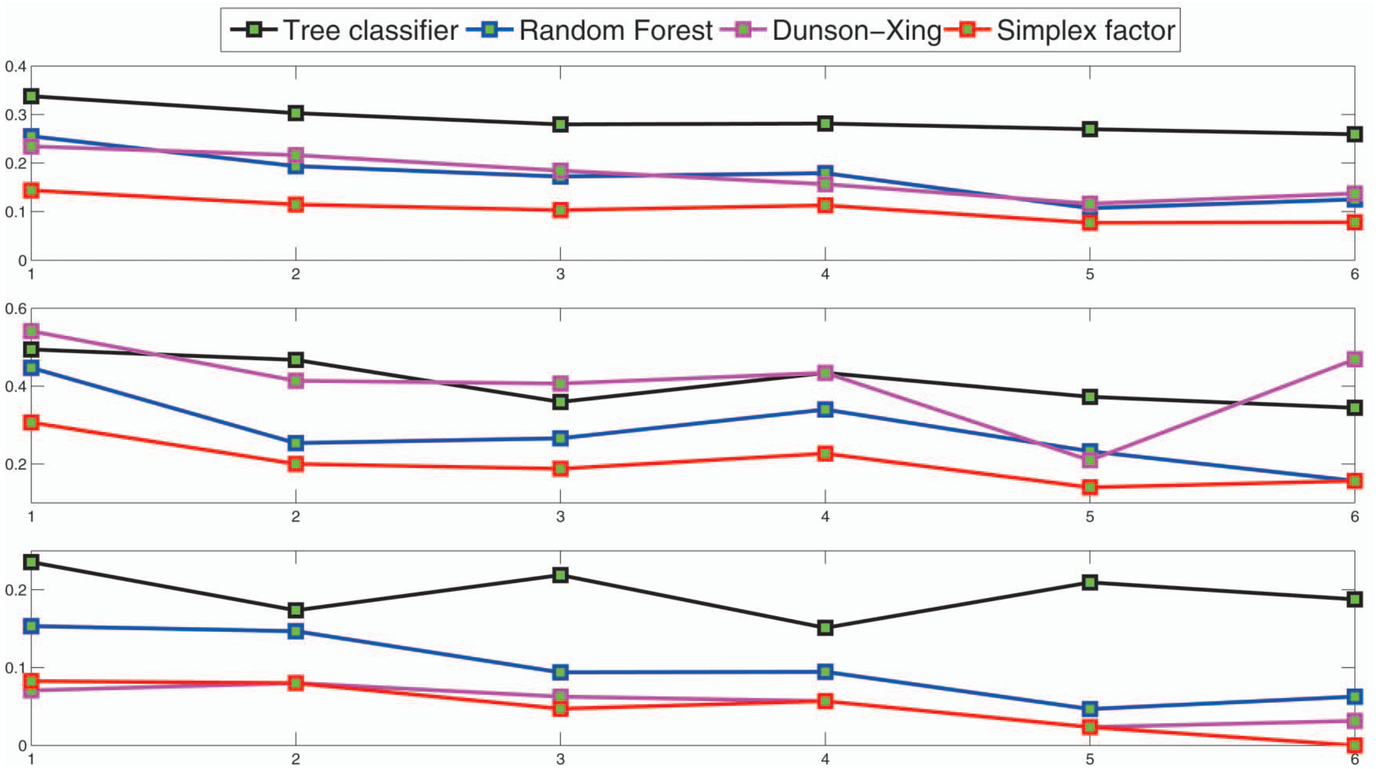

Figure 7. Results for the promoter data—misclassification proportions (overall) for the different methods versus training size (ranging from 20% to 70% of the sample size $n = 106$). The rows correspond to average-, worst- and best-case performances, respectively, across 10 training sets for each training size. (The online version of this figure is in color.)

7. DISCUSSION

In a variety of problems, one now encounters data where the dimensionality of the outcome is comparable or even larger than the number of subjects. In such scenarios, one needs to make sparsity assumptions for meaningful inference. For continuous outcomes, one might consider sparse modeling of the covariance matrix via factor models or alternatively use Gaussian graphical models for sparse modeling of the precision matrix. However, the scope of either of these two frameworks is not solely limited to continuous variables. In more general terms, graphical models aim to model conditional dependencies among the variables, while factor models model marginal dependence

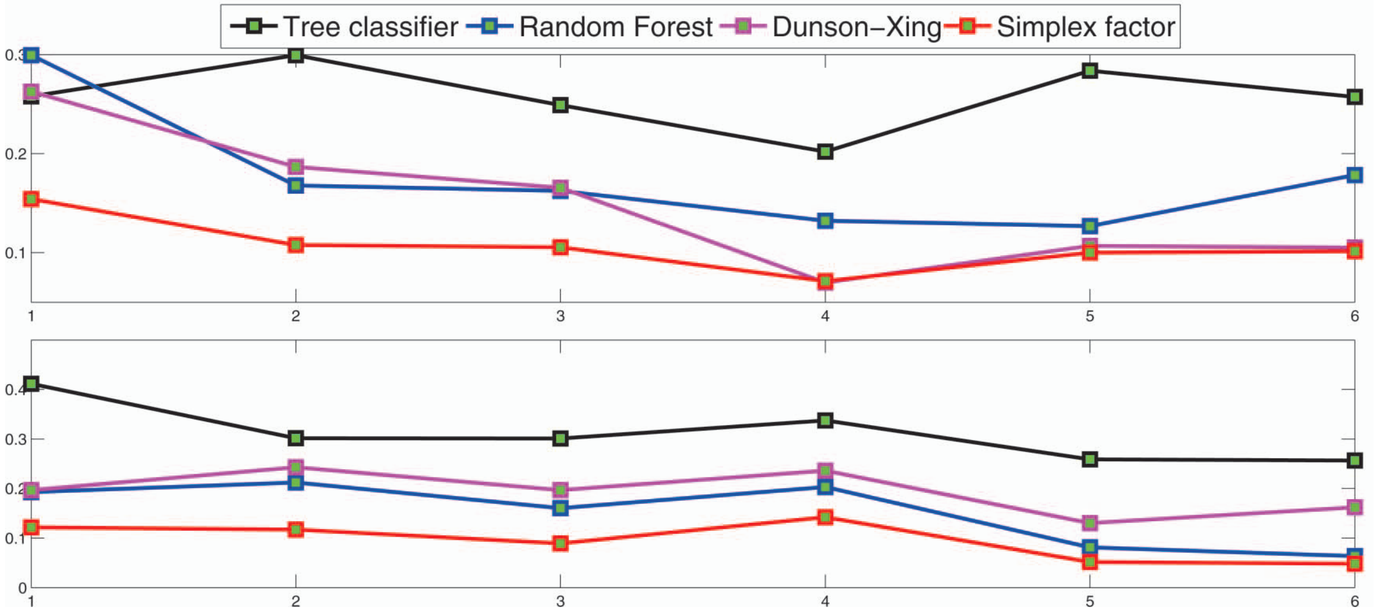

Figure 8. Results for the promoter data—average misclassification proportions (for the promoter and nonpromoter group, respectively) for the different methods versus training size (ranging from 20% to 70% of the sample size $n = 106$). (The online version of this figure is in color.)

relationships. In this article, we have proposed a sparse Bayesian factor modeling approach for multivariate nominal data that aims to explain dependence among high-dimensional nominal variables in terms of few latent factors that reside on a simplex. Posterior computation is straightforward and scales linearly with n and p. The proposed method can be thus used in high-dimensional problems, which is an advantage over graphical model-based approaches, which face computational challenges in scaling up to high dimensions.

An interesting extension of our proposed approach is joint modeling of a vector of nominal predictors and a continuous response, and more generally mixtures of different data types, as such situations are often encountered in biological and social sciences. To elaborate, let $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$ denote a vector of observations as before, where the y_{ij}'s now are allowed to be of different types, such as binary, count, ordinal, continuous, etc. Letting $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{ip})^{\mathrm{T}} \in \{1, 2, \ldots, \infty\}^p$ denote a multivariate latent class index for subject i, one can let

$$y_{ij} \sim \mathsf{K}_j(\theta_{\gamma_{ij}j}), \quad j = 1, \ldots, p, \quad \gamma_i \sim G, \quad G \sim \Pi, \quad (16)$$

where G is the joint distribution of the multivariate categorical variable γ_i. The different y_{ij}'s are assumed to be conditionally independent, given the latent class index γ_i, and a prior Π on the distribution G of the latent class indices induces dependence among the y_{ij}'s. In a Dirichlet process mixture modeling framework, one usually has a single cluster index γ_i for the different data types, which forces individuals to be allocated to the same cluster across all data types. This often leads to blowing up of the number of clusters and degraded performance; see, for example, Dunson (2009). We can instead allow for separate but dependent clustering across the different domains by letting Π correspond to the simplex factor prior. One can also include covariate information by stacking together the covariates x_i and the response y_i in a vector $z_i = (y_i, x_i)$ and jointly model z_i as above, with inference based on the induced conditional distribution of $y_i \mid x_i$ obtained from the joint model. Müller, Erkanli, and West (1996) considered such joint models in a nonparametric Bayes frame work using Dirichlet process mixtures.

There has been a recent surge of interest in developing flexible Bayesian density regression models where the entire conditional distribution of the response y, given the predictors x, is allowed to change flexibly with x; see, for example, Griffin and Steel (2006); Dunson, Pillai, and Park (2007); Dunson and Park (2008); Chung and Dunson (2009); Rodriguez and Dunson (2011). It would be interesting to consider extensions of these models for multivariate categorical response variables by allowing predictor-dependent weights, for example, using the probit stick-breaking process (Chung and Dunson 2009; Rodriguez and Dunson 2011). Pati and Dunson (2011) developed theoretical tools for studying posterior consistency with a broad class of predictor-dependent stick-breaking priors; see also Norets and Pelenis (2011). Along those lines, one can envision extensions of our baseline posterior consistency results to the uncountable collection of probability tensors $\{\ \ (x) \ : \ x \in \mathsf{X}\}$.

APPENDIX

Proof of Proposition 2.2

From Equation (10), one has

$$g_{h_1\ldots h_p} = \frac{\prod_{h:\tau_h \neq 0}(\alpha\nu_h)(\alpha\nu_h + 1), \ldots, (\alpha\nu_h + \tau_h - 1)}{\alpha(\alpha + 1), \ldots, (\alpha + p - 1)}.$$

Dividing the numerator and the denominator in the above expression by α^p, it is evident that $\lim_{\alpha\to\infty} g_{h_1,\ldots,h_p} = \Pi_{h=1}^k \nu_h^{\tau_h} = \nu_{h_1}, \ldots, \nu_{h_p}$, which corresponds to the product multinomial model.

On the other hand,

$$g_{h\ldots h} = \frac{(\alpha\nu_h)(\alpha\nu_h + 1), \ldots, (\alpha\nu_h + p - 1)}{\alpha(\alpha + 1), \ldots, (\alpha + p - 1)}.$$

Clearly, $\lim_{\alpha\to 0} g_{h\ldots h} = \nu_h$, and thus in the limit, $\sum_{h=1}^k g_{h\ldots h} = 1$. As $\mathbf{G}$ is a probability tensor for every value of α, the nondiagonal elements must converge to 0 as $\alpha \to 0$. Hence, in this limiting case, $\mathbf{G}$ becomes super-diagonal and thus corresponds to the Dunson and Xing (2009) model.

Proof of Theorem 2.3

Fix $\ ^0 \in 5_{d_1\ldots d_p}$ and $\epsilon > 0$. For $\ \in 5_{d_1\ldots d_p}$, the L_1 distance between $\ $ and $\ ^0$ is defined as

$$||\quad - \quad ^0||_1 = \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} |\pi_{c_1\ldots c_p} - \pi^0_{c_1\cdots c_p}|.$$

Suppose $\ ^0$ has nonnegative PARAFAC rank k, so $\ ^0$ can be expressed as:

$$^0 = \sum_{h=1}^k \nu_{0h} 3_{\ 0h}, \quad 3_{\ 0h} = \ _{0h}^{(1)} \otimes \cdots \otimes \ _{0h}^{(p)},$$

where $\ _0 \in \mathsf{S}_{k-1}$ and $\ _{0h}^{(1)}, \ldots, \ _{0h}^{(p)}$ are probability vectors of dimensions $d_1, \ldots, d_p$ for each $h \in \{1, \ldots, k\}$. The prior probability assigned to an ϵ-sized L_1-neighborhood $\mathsf{N}_\epsilon(\ ^0)$ of $\ ^0$ by a k-component simplex factor model is given by:

$$\mathsf{Q}_\pi^{(k)}\{\mathsf{N}_\epsilon(\ ^0)\} = \int 1(||\quad - \quad ^0||_1 < \epsilon)$$
$$d\mathsf{Q}_\pi^{(k)}(\ , \alpha, \ _h^{(j)}, h = 1, \ldots, k; j = 1, \ldots, p), \quad (A.1)$$

where

$$\pi_{c_1\ldots c_p} = \sum_{h_1=1}^k \cdots \sum_{h_p=1}^k g_{h_1\ldots h_p} \prod_{j=1}^p \lambda_{h_j c_j}^{(j)},$$

with $g_{h_1\ldots h_p}$, as in Equation (10). Using Proposition 2.2 and standard algebra, it can be shown that for any $\epsilon > 0$, there exist $\tilde{\alpha} > 0$ and $\tilde{\epsilon} > 0$ such that

$$\alpha < \tilde{\alpha}, \quad ||\quad - \quad _0||_1 < \tilde{\epsilon}, \quad \left|\left| \ _h^{(j)} - \ _{0h}^{(j)} \right|\right|_1 < \tilde{\epsilon} \quad \forall h, j$$

implies that $||\quad - \quad ^0||_1 < \epsilon$. Hence, to prove that (A.1) is positive, it suffices to show that 12501000:

$$\mathsf{Q}_\pi^{(k)}\Big(\alpha < \tilde{\alpha}, \quad ||\quad - \quad _0||_1 < \tilde{\epsilon}, \quad \left|\left| \ _h^{(j)} - \ _{0h}^{(j)} \right|\right|_1$$
$$< \tilde{\epsilon}, \ h = 1, \ldots, k; \ j = 1, \ldots, p\Big) > 0,$$

which immediately follows from the prior specification in (9).

Updating ν

We drop the h superscript in $c_l^{(h)}$ for notational convenience. Let

$$f(\nu_h^*) = \prod_{i=1}^{n}\prod_{l=h}^{k}\frac{\Gamma(\alpha\nu_l + m_{il})}{\Gamma(\alpha\nu_l)} = \left[\prod_{i:m_{ih}\neq 0}\frac{\Gamma(c_h\nu_h^* + m_{ih})}{\Gamma(c_h\nu_h^*)}\right] \times \left[\prod_{l=h+1}^{k}\prod_{i:m_{il}\neq 0}\frac{\Gamma\{c_l(1-\nu_h^*) + m_{il}\}}{\Gamma\{c_l(1-\nu_h^*)\}}\right] \quad (A.2)$$

denote the unnormalized conditional posterior, and define $\phi(\nu_h^*) = \log f(\nu_h^*)$. When $m_h > 0$ and $m_{h+} = 0$,

$$\pi(\nu_h^* \mid -) \propto \prod_{i:m_{ih}\neq 0}\prod_{s=0}^{m_{ih}-1}(c_h\nu_h^* + s),$$

so $\lim_{\nu_h^*\to 0}\pi(\nu_h^* \mid -) = 0$, $0 < \lim_{\nu_h^*\to 1}\pi(\nu_h^* \mid -) < \infty$, and $\pi(\nu_h^* \mid -)$ is convex. Similarly, if $m_h = 0$ and $m_{h+} > 0$, then $0 < \lim_{\nu_h^*\to 0}\pi(\nu_h^* \mid -) < \infty$, $\lim_{\nu_h^*\to 1}\pi(\nu_h^* \mid -) = 0$, and $\pi(\nu_h^* \mid -)$ is convex. In both these cases, the conditional posterior of ν_h^* can be approximated by a single beta density.

When $m_h > 0$ and $m_{h+} > 0$,

$$\begin{aligned}\phi(\nu_h^*) &= \sum_{l=h}^{k}\sum_{i:m_{il}\neq 0}\sum_{s=0}^{m_{il}-1}\log(\alpha\nu_l + s)\\ &= \sum_{l=h}^{k}\sum_{s=0}^{\infty}\log(\alpha\nu_l + s)n_{ls} = \sum_{s=0}^{\infty}n_{hs}\log(c_h\nu_h^* + s)\\ &\quad + \sum_{s=0}^{\infty}\sum_{l=h+1}^{k}n_{ls}\log\{c_l(1-\nu_h^*) + s\}.\end{aligned}$$

As $\nu_h^* \to 0$, $\sum_{s=1}^{\infty}n_{hs}\log(c_h\nu_h^* + s) + \sum_{s=0}^{\infty}\sum_{l=h+1}^{k}n_{ls}\log\{c_l(1-\nu_h^*) + s\}$ converges to a finite limit. Since $m_h > 0$, $n_{h0} > 0$ and thus $\lim_{\nu_h^*\to 0}n_{h0}\log(c_h\nu_h^*) = -\infty$, implying $\lim_{\nu_h^*\to 0}\phi(\nu_h^*) = -\infty$. Similarly, $\lim_{\nu_h^*\to 1}\phi(\nu_h^*) = -\infty$, since $m_{h+} > 0$ implies that there exists $h' > h$ such that $n_{h's} > 0$. Hence, $\lim_{\nu_h^*\to 0,1}\phi(\nu_h^*) = -\infty$ and thus $\lim_{\nu_h^*\to 0,1}\pi(\nu_h^* \mid -) = 0$.

We now show that $\pi(\nu_h^* \mid -)$ has a unique mode by considering the first and second derivatives of ϕ. We have

$$\phi'(\nu_h^*) = \sum_{s=0}^{\infty}n_{hs}\frac{c_h}{(c_h\nu_h^* + s)} - \sum_{s=0}^{\infty}\sum_{l=h+1}^{k}n_{ls}\frac{c_l}{\{c_l(1-\nu_h^*) + s\}}, \quad (A.3)$$

and

$$\phi''(\nu_h^*) = -\sum_{s=0}^{\infty}n_{hs}\left\{\frac{c_h}{(c_h\nu_h^* + s)}\right\}^2 - \sum_{s=0}^{\infty}\sum_{l=h+1}^{k}n_{ls}\left\{\frac{c_l}{\{c_l(1-\nu_h^*) + s\}}\right\}^2. \quad (A.4)$$

Once again, using the fact that $n_{h0} > 0$ and there exists $h' > h$ such that $n_{h'0} > 0$, one can prove that $\lim_{\nu_h^*\to 0}\phi'(\nu_h^*) = \infty$ and $\lim_{\nu_h^*\to 1}\phi'(\nu_h^*) = -\infty$. One can then find $\epsilon > 0$ such that $\phi'(\nu_h^*) > 0$ for $\nu_h^* < \epsilon$ and $\phi'(\nu_h^*) < 0$ for $\nu_h^* > 1-\epsilon$. Since ϕ' is a continuous function, by the intermediate value theorem, there exists $\nu_{0h}^* \in (0,1)$ such that $\phi'(\nu_{0h}^*) = 0$. Since $\phi''(\nu_h^*) < 0$ on $(0,1)$, ϕ' is monotonically decreasing and hence ν_{0h}^* is the unique mode of $\pi(\nu_h^* \mid -)$.

Next, we discuss an approach to avoid the grid approximation and obtain analytic expressions for the parameters of the approximating beta distribution. When $m_h > 0$ and $m_{h+} = 0$, we want to find $a \geq 1$ such that a beta$(a, 1)$ approximates $\pi(\nu_h^* \mid -)$. Define $\tilde f(\nu_h^*) = f(\nu_h^*)/f(1)$ so that $\tilde f(1) = 1$. Hence, the above problem can be equivalently posed as approximating $\tilde f(\nu_h^*)$ by $(\nu_h^*)^{a-1}$. Since $\int_0^1\log\{(\nu_h^*)^{a-1}\}d\nu_h^* = -(a-1)$, we let $\hat a = 1 - \int_0^1\tilde\phi(\nu_h^*)d\nu_h^*$, where $\tilde\phi(x) = \log\tilde f(x)$, which leads to the expression in (14). Observe that $\log(1 + c_h/s) \leq c_h/s$; hence, $\hat a \geq 1$. The analysis for the case where $m_h = 0$ and $m_{h+} > 0$ proceeds along similar lines, where we define $\tilde f(\nu_h^*) = f(\nu_h^*)/f(0)$ and find $\hat b$ so that $\tilde f(\nu_h^*) \approx (1-\nu_h^*)^{b-1}$.

When $m_h > 0$ and $m_{h+} > 0$, the analysis proceeds slightly differently. We want $a, b \geq 1$ such that $\pi(\nu_h^* \mid -) \approx \{1/\text{beta}(a,b)\}(\nu_h^*)^{a-1}(1-\nu_h^*)^{b-1}$. Comparing the first three moments of $\log\pi(\nu_h^* \mid -)$ to the log beta(a, b) density, we build the 2×2 linear system mentioned in Section 3.1 to estimate a, b. The formulas for d_1, d_2 can be obtained from the following expressions

$$\int_0^1\log\pi(\nu_h^* \mid -)d\nu_h^* = \log c + \sum_{s=0}^{\infty}\sum_{l=h}^{k}n_{ls}\psi_0(c_l, s)$$

$$\int_0^1 2\nu_h^*\log\pi(\nu_h^* \mid -)d\nu_h^* = \log c + 2\sum_{s=0}^{\infty}n_{hs}\psi_1(c_h, s) + 2\sum_{s=0}^{\infty}\sum_{l=h+1}^{k}n_{ls}\{\psi_0(c_l, s) - \psi_1(c_l, s)\}$$

$$\int_0^1 3(\nu_h^*)^2\log\pi(\nu_h^* \mid -)d\nu_h^* = \log c + 3\sum_{s=0}^{\infty}\sum_{l=h}^{k}n_{ls}\psi_2(c_l, s) + 3\sum_{s=0}^{\infty}\sum_{l=h+1}^{k}n_{ls}\{\psi_0(c_l, s) - 2\psi_1(c_l, s)\},$$

where

$$\psi_0(c, s) = \int_0^1\log(cx + s)dx = \frac{s}{c}\log(c + s) + \log(c + s) - 1 - \frac{s}{c}\log(s)$$

$$\psi_1(c, s) = \int_0^1 x\log(cx + s)dx = -\frac{s^2}{2c^2}\log(c + s) + \frac{1}{2}\log(c + s) + \frac{s}{2c} - \frac{1}{4} - \frac{s^2}{2c^2}\log(s)$$

$$\psi_2(c, s) = \int_0^1 x^2\log(cx + s)dx = \frac{s^3}{3c^3}\log(c + s) + \frac{1}{3}\log(c + s) - \frac{s^2}{3c^2} + \frac{s}{6c} - \frac{1}{9} - \frac{s^3}{3c^3}\log(s).$$

[Received October 2010. Revised May 2011.]

REFERENCES

Aitchison, J., and Bennett, J. (1970), "Polychotomous Quantal Response by Maximum Indicant," *Biometrika*, 57(2), 253–262. [362]

Ashford, J. R., and Sowden, R. R. (1970), "Multivariate Probit Analysis," *Biometrics*, 26, 535–546. [362]

Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D., and Jordan, M. (2003), "Matching Words and Pictures," *Journal of Machine Learning Research*, 3, 1107–1135. [363]

Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, New York: Springer. [362]

Blei, D., Ng, A., and Jordan, M. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [363]

Bollen, K. (1989), *Structural Equations With Latent Variables*, New York: Wiley. [362]

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45(1), 5–32. [369]

Carvalho, C., Lucas, J., Wang, Q., Nevins, J., and West, M. (2008), "High-Dimensional Sparse Factor Modelling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456. [362]

Carvalho, C., and Scott, J. (2009), "Objective Bayesian Model Selection in Gaussian Graphical Models," *Biometrika*, 96(3), 1–16. [362]

Chib, S., and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 347–361. [362]

Chung, Y., and Dunson, D. (2009), "Nonparametric Bayes Conditional Distribution Modeling With Variable Selection," *Journal of the American Statistical Association*, 104(488), 1646–1660. [374]

Cohen, J. E., and Rothblum, U. G. (1993), "Nonnegative Ranks, Decompositions, and Factorizations of Nonnegative Matrices," *Linear Algebra and Its Applications*, 190, 149–168. [364]

Dawid, A., and Lauritzen, S. (1993), "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *The Annals of Statistics*, 21(3), 1272–1317. [362]

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000), "A Multilinear Singular Value Decomposition," *SIAM Journal on Matrix Analysis and Applications*, 21(4), 1253–1278. [363,364]

Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G., and West, M. (2004), "Sparse Graphical Models for Exploring Gene Expression Data," *Journal of Multivariate Analysis*, 90(1), 196–212. [362]

Dobra, A., and Lenkoski, A. (2011), "Copula Gaussian Graphical Models," *The Annals of Applied Statistics*, 5, 969–993. [362,367,369]

Dobra, A., and Massam, H. (2010), "The Mode Oriented Stochastic Search (MOSS) Algorithm for Log-Linear Models With Conjugate Priors," *Statistical Methodology*, 7(3), 240–253. [362]

Dunson, D. B. (2000), "Bayesian Latent Variable Models for Clustered Mixed Outcomes," *Journal of the Royal Statistical Society*, Series B, 62(2), 355–366. [362]

—— (2003), "Dynamic Latent Trait Models for Multidimensional Longitudinal Data," *Journal of the American Statistical Association*, 98(463), 555–563. [362]

—— (2009), "Nonparametric Bayes Local Partition Models for Random Effects," *Biometrika*, 96(2), 249. [374]

Dunson, D. B., and Park, J. (2008), "Kernel Stick-Breaking Processes," *Biometrika*, 95(2), 307. [374]

Dunson, D. B., Pillai, N., and Park, J. (2007), "Bayesian Density Regression," *Journal of the Royal Statistical Society*, Series B, 69(2), 163–183. [374]

Dunson, D. B., and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, 104(487), 1042–1051. [363,364,365,367,369,370,371,372,374]

Erosheva, E., Fienberg, S., and Joutard, C. (2007), "Describing Disability Through Individual-Level Mixture Models for Multivariate Binary Data," *The Annals of Applied Statistics*, 1(2), 502–537. [363]

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [363]

—— (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629. [363]

Fienberg, S., and Rinaldo, A. (2007), "Three Centuries of Categorical Data Analysis: Log-Linear Models and Maximum Likelihood Estimation," *Journal of Statistical Planning and Inference*, 137(11), 3430–3445. [362]

Frank, A., and Asuncion, A. (2010), "UCI Machine Learning Repository," available at *http://archive.ics.uci.edu/ml*. Irvine, CA: School of Information and Computer Science, University of California. [371]

Ghosh, J., and Ramamoorthi, R. (2003), *Bayesian Nonparametrics*, New York: Springer-Verlag. [365]

Goodman, L. A. (1974), "Explanatory Latent Structure Assigning Both Identifiable and Unidentifiable Models," *Biometrika*, 61, 215–231. [363]

Gregory, D., and Pullman, N. (1983), "Semiring Rank: Boolean Rank and Nonnegative Rank Factorizations," *Journal of Combinatorics, Information & System Sciences*, 8(3), 223–233. [364]

Griffin, J., and Steel, M. (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101(473), 179–194. [374]

Harshman, R. (1970), "Foundations of the PARAFAC Procedure: Models and Conditions for an 'Explanatory' Multi-Modal Factor Analysis," *UCLA Working Papers in Phonetics*, 16 (1), 84, Los Angeles, CA: UCLA. [364]

Ishwaran, H., and James, L. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96(453), 161–173. [365]

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005), "Experiments in Stochastic Computation for High-Dimensional Graphical Models," *Statistical Science*, 20(4), 388–400. [362]

Kim, Y., and Choi, S. (2007), "Nonnegative Tucker Decomposition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007—CVPR'07*, pp. 1–8. [363,364]

Kolda, T. (2001), "Orthogonal Tensor Decompositions," *SIAM Journal on Matrix Analysis and Applications*, 23(1), 243–255. [364]

Lauritzen, S. (1996), *Graphical Models*, Oxford: Oxford University Press. [362]

Lazarsfeld, P., and Henry, N. (1968), *Latent Structure Analysis*, Boston, MA: Houghton Mifflin. [363]

Lenkoski, A., and Dobra, A. (2011), "Computational Aspects Related to Inference in Gaussian Graphical Models With the G-Wishart Prior," *Journal of Computational and Graphical Statistics*, 20(1), 140–157. [362]

Liaw, A., and Wiener, M. (2002), "Classification and Regression by Random Forest," *R News*, 2(3), 18–22. [369]

Lopes, H., and West, M. (2004), "Bayesian Model Assessment in Factor Analysis," *Statistica Sinica*, 14(1), 41–68. [366]

Madigan, D., and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63(2), pp. 215–232. [362]

Massam, H., Liu, J., and Dobra, A. (2009), "A Conjugate Prior for Discrete Hierarchical Log-Linear Models," *The Annals of Statistics*, 37(6A), 3431–3467. [362]

Moustaki, I., and Knott, M. (2000), "Generalized Latent Trait Models," *Psychometrika*, 65(3), 391–411. [362]

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian Curve Fitting Using Multivariate Normal Mixtures," *Biometrika*, 83(1), 67. [374]

Muthén, B. (1983), "Latent Variable Structural Equation Modeling With Categorical Data," *Journal of Econometrics*, 22(1–2), 43–65. [362]

Norets, A., and Pelenis, J. (2011), "Posterior Consistency in Conditional Density Estimation by Covariate Dependent Mixtures," Technical report, Princeton, NJ: Princeton University. [374]

Ochi, Y., and Prentice, R. (1984), "Likelihood Inference in a Correlated Probit Regression Model," *Biometrika*, 71(3), 531–543. [362]

Pati, D., and Dunson, D. (2011), "Posterior Consistency in Conditional Distribution Estimation," Working Paper, Durham, NC: Department of Statistical Science, Duke University. [374]

Pitt, M., Chan, D., and Kohn, R. (2006), "Efficient Bayesian Inference for Gaussian Copula Regression Models," *Biometrika*, 93(3), 537–554. [362]

Pritchard, J., Stephens, M., and Donnelly, P. (2000), "Inference of Population Structure Using Multilocus Genotype Data," *Genetics*, 155(2), 945. [363]

Rodriguez, A., and Dunson, D. (2011), "Nonparametric Bayesian Models Through Probit Stick-Breaking Processes," *Bayesian Analysis*, 6, 145–178. [374]

Sammel, M., Ryan, L., and Legler, J. (1997), "Latent Variable Models for Mixed Discrete and Continuous Outcomes," *Journal of the Royal Statistical Society*, Series B, 59(3), 667–678. [362]

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4(2), 639–650. [365]

Shashua, A., and Hazan, T. (2005), "NonNegative Tensor Factorization With Applications to Statistics and Computer Vision," in *Proceedings of the 22nd International Conference on Machine Learning*, p. 799. [363,364]

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101(476), 1566–1581. [365]

Tucker, L. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31(3), 279–311. [364]

Wang, H., and Ahuja, N. (2005), "Rank-R Approximation of Tensors: Using Image-as-Matrix Representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005—CVPR'05*, pp. 346–353. [364]

Wei, C., Wu, Q., Vega, V., Chiu, K., Ng, P., Zhang, T. et al. (2006), "A Global Map of p53 Transcription-Factor Binding Sites in the Human Genome," *Cell*, 124(1), 207–219. [371]

West, M. (2003), "Bayesian Factor Regression Models in the 'Large p, Small n' Paradigm," *Bayesian Statistics*, 7(2003), 723–732. [362]

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, New York: Wiley. [362,367]

Xie, X., and Geng, Z. (2008), "A Recursive Method for Structural Learning of Directed Acyclic Graphs," *Journal of Machine Learning Research*, 9, 459–483. [371]

Zhang, X., Boscardin, W. J., and Belin, T. R. (2006), "Sampling Correlation Matrices in Bayesian Models With Correlated Latent Variables," *Journal of Computational and Graphical Statistics*, 15(4), 880–896. [362]

—— (2008), "Bayesian Analysis of Multivariate Nominal Measures Using Multivariate Multinomial Probit Models," *Computational Statistics & Data Analysis*, 52, 3297–3708. [362]