Published in final edited form as:
J Am Stat Assoc. 2012 January 1; 104(487): 1042–1051. doi:10.1198/jasa.2009.tm08439.

Nonparametric Bayes Modeling of Multivariate Categorical Data

David B. Dunson [Professor] and
Department of Statistical Science, Duke University, Durham, NC 27705

Chuanhua Xing [Postdoctoral Associate]
Department of Biology, Duke University, Durham, NC 27705

David B. Dunson: dunson@stat.duke.edu; Chuanhua Xing: chuanhua.xing@duke.edu

Abstract

Modeling of multivariate unordered categorical (nominal) data is a challenging problem, particularly in high dimensions and cases in which one wishes to avoid strong assumptions about the dependence structure. Commonly used approaches rely on the incorporation of latent Gaussian random variables or parametric latent class models. The goal of this article is to develop a nonparametric Bayes approach, which defines a prior with full support on the space of distributions for multiple unordered categorical variables. This support condition ensures that we are not restricting the dependence structure a priori. We show this can be accomplished through a Dirichlet process mixture of product multinomial distributions, which is also a convenient form for posterior computation. Methods for nonparametric testing of violations of independence are proposed, and the methods are applied to model positional dependence within transcription factor binding motifs.

Keywords

Bayes factor; Dirichlet process; Goodness-of-fit test; Latent class; Mixture model; Motif data; Product multinomial; Unordered categorical

1. INTRODUCTION

Our goal is to develop methods for Bayesian nonparametric modeling of multivariate unordered categorical data. The application that motivated this work is the modeling of dependence of nucleotides within a transcription factor binding motif, which is a short DNA sequence involved in gene regulation. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ denote a sequence of nucleotides within positions $1, \ldots, p$ of a known motif of interest. We let $x_{ij} \in \{1, \ldots, d_j\}$, for $d_j = d = 4$, with values 1, 2, 3, 4 corresponding to the A, C, G, T nucleotides, respectively. In the motif application, the most common model for $\mathbf{x}_i$ is the product-multinomial distribution, which characterizes the elements of $\mathbf{x}_i$ as independent draws from multinomial distributions having a position-specific weight matrix (Lawrence and Reilly 1990; Liu 1994; Liu, Neuwald, and Lawrence 1995). However, experimental data suggest independence is often violated (Bulyk, Johnson, and Church 2002; Tomovic and Oakeley 2007).

Barash et al. (2003) proposed using a Bayesian network (BN). Unless p is small, the number of possible BNs is so large that it is only feasible to visit a small proportion of the models. Even computationally expensive search algorithms can miss the best models. In large model spaces, there are typically many BNs that are consistent with the data, so the best BN according to any criteria is very unlikely to be the true model. In addition, the selected BN will be sensitive to the criteria chosen for inclusion of an edge, and there is a tendency for overfitting in large networks due to the massive number of potential edges. For a recent

article developing an efficient recursive method for BN search, refer to Xie and Geng (2008). Zhou and Liu (2004) noted that the Barash et al. (2003) approach is very time consuming to implement and is subject to problems in prior choice and overfitting, motivating a simpler model based on pairs of correlated positions. Nikolajewa et al. (2007) focused on tree-augmented BNs to reduce computational time and problems with poor performance of BNs in small datasets.

Ideally, it would be possible to accommodate the fact that we have no idea what the true dependence structure is, while favoring a sparse structure that allows efficient computation without facing problems with overfitting. The dependence is of substantial interest biologically, and more realistic models of variability within motifs open the door for improved approaches for motif discovery. With this motivation in mind, our goal is to develop a nonparametric Bayes approach to allow the distribution of multiple unordered categorical variables to be unknown. We show that mixtures of product multinomial distributions have full support, with Dirichlet process (DP) (Ferguson 1973, 1974) mixtures providing a convenient and flexible framework. For posterior computation, we can rely on efficient and simple-to-implement methods developed for conjugate DP mixture models. Hence, we bypass the complexities involved in building a Bayes network for unordered categorical data without inducing any restrictions, while also favoring a sparse formulation of the heterogeneity structure. A formal test of lack of fit for the product multinomial model against a nonparametric alternative can be constructed directly from the Gibbs sampling output.

Although our initial motivation was the motif application, our methodology provides a general approach for nonparametric Bayes modeling of multivariate unordered categorical data. The current literature on parametric models focuses primarily on underlying continuous variable specifications. For example, the multinomial probit model (Ashford and Sowden 1970; Ochi and Prentice 1984; Chib and Greenberg 1998) incorporates a latent multivariate Gaussian random variable, which is linked to the categorical observations through thresholding. Zhang, Boscardin, and Belin (2008) developed a Bayes approach to posterior computation for the multivariate multinomial probit model. Such models are quite flexible, but also tend to be difficult to implement due to the need to estimate a correlation matrix in underlying variables that have a complex relationship with the measured outcomes. An alternative approach is to use a finite mixture model. Such an approach has historically been referred to as latent structure analysis (Lazarsfeld and Henry 1968), though the term latent class modeling is more commonly used in recent years. Refer to Formann (2007) for a recent article applying latent class modeling to multivariate categorical data with missing values.

The nonparametric Bayes literature on multivariate unordered categorical data analysis for more than two categories is essentially nonexistent. A number of articles have focused on nonparametric Bayes methods for multiple binary outcomes (Quintana and Newton 2000; Hoff 2005; Jara, Garcia-Zattera, and Lesaffre 2007). In addition, Kottas, Müller, and Quintana (2005) propose a nonparametric Bayes modification of the multivariate probit model for ordinal data, with the normal distribution on the underlying variables replaced with a Dirichlet process mixture (DPM) of normals. Ansari and Iyengar (2006) proposed a semiparametric Bayes approach to repeated unordered categorical choice data, again following the approach of using DPMs to relax normality assumptions on latent variables. Quintana (1998) instead considered a nonparametric Bayes approach to assess homogeneity in contingency tables having fixed right marginals. His approach borrowed information across the rows of the contingency table by assuming the row vectors of probabilities are drawn from a common distribution, which is assigned a DP prior. This approach is

appropriate for univariate unordered categorical data collected for subjects in different groups.

Section 2 describes the proposed nonparametric Bayes mixture model, providing theoretical support. Section 3 describes a simple approach for posterior computation and goodness-of-fit testing. Section 4 contains a simulation study. Section 5 applies the method to assess dependence within the p53 motif, and Section 6 contains a discussion.

2. PRODUCT MULTINOMIAL MIXTURE MODELS

2.1 Latent Structure Analysis and Properties

We focus initially on the case in which $p = 2$, so that data for subject i consist of a pair of categorical variables, $\mathbf{x}_i = (x_{i1}, x_{i2})'$, resulting in a $d_1 \times d_2$ contingency table with cell (c_1, c_2) containing the count $\sum_{i=1}^{n} 1(x_{i1}=c_1, x_{i2}=c_2)$, for $c_1 = 1, \ldots, d_1$ and $c_2 = 1, \ldots, d_2$. Our focus is on sparse non-parametric modeling of the cell probabilities, $\boldsymbol{\pi} = \{\pi_{c_1 c_2}\}$ with $\pi_{c_1 c_2} = \Pr(x_{i1} = c_1, x_{i2} = c_2)$.

One simple and parsimonious model for $\boldsymbol{\pi}$ can be obtained by assuming $\Pr(x_{i1}=c_1)=\psi^{(1)}_{c_1}$ and $\Pr(x_{i2}=c_2)=\psi^{(2)}_{c_2}$ with x_{i1} and x_{i2} independent. In this case, we obtain $\pi_{c_1 c_2}=\psi^{(1)}_{c_1}\psi^{(2)}_{c_2}$, so that instead of $d_1 d_2 - 1$ free parameters characterizing $\boldsymbol{\pi}$ in the saturated case, we have $d_1 + d_2 - 2$ free parameters. In practice, this model may be overly simple, motivating latent structure analysis (Lazarsfeld and Henry 1968; Goodman 1974), which instead relies on the finite mixture specification

$$\Pr(x_{i1}=c_1, x_{i2}=c_2)=\pi_{c_1 c_2}=\sum_{h=1}^{k}\nu_h\psi^{(1)}_{hc_1}\psi^{(2)}_{hc_2}, \quad (1)$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)'$ is a vector of mixture probabilities, $z_i \in \{1, \ldots, k\}$ denotes a latent class index, $\Pr(x_{i1}=c_1|z_i=h)=\psi^{(1)}_{hc_1}$ is the probability of $x_{i1} = c_1$ in class h, $\Pr(x_{i2}=c_2|z_i=h)=\psi^{(2)}_{hc_2}$ is the probability of $x_{i2} = c_2$ in class h, and x_{i1} and x_{i2} are assumed to be conditionally independent given z_i.

There is a rich literature focusing on issues in latent structure modeling, including identifiability of the $\boldsymbol{\nu}$ and $\boldsymbol{\psi}=\{\boldsymbol{\psi}_h\}_{h=1}^{k}$ parameters, methods for selection of k, extensions to multiple group settings, etc. However, our goal is to build Bayesian non-parametric models that favor sparse formulations, while including all joint distributions for x_{i1} and x_{i2} in the support of the prior. With this goal in mind, it is important to verify the following theorem in order for formulation (1) to be sufficiently flexible.

Theorem 1—Any $\boldsymbol{\pi} \in \Pi_{d_1 d_2}$ can be characterized as in (1) for some k, with $\Pi_{d_1 d_2}$ containing all $d_1 \times d_2$ probability matrices with elements $0 \leq \pi_{c_1 c_2} \leq 1$ and $\sum_{c_1=1}^{d_1}\sum_{c_2=1}^{d_2}\pi_{c_1 c_2}=1$.

It follows from Theorem 1 that the joint distribution of two categorical random variables can always be expressed as a finite mixture of product-multinomial distributions. Good (1969) previously noted the similarity between the singular value decomposition (SVD) and the latent structure model without providing a proof that any probability matrix, $\boldsymbol{\pi}$, can be decomposed using formulation (1). Gilula (1979) instead used the SVD to provide

conditions for identifying the distribution of latent variables inducing dependence in two categorical variables.

Focusing on the generalization to the multivariate case, let

$$\pi=\{\pi_{c_1c_2\cdots c_p}, c_j=1,\ldots,d_j, j=1,\ldots,p\}\in\Pi_{d_1\cdots d_p}$$

denote a higher order tensor, with $\Pi_{d_1\cdots d_p}$ denoting the set of all probability tensors of size $d_1\times d_2\times\cdots\times d_p$, where probability tensors have nonnegative elements and

$$\|\pi\|_1=\sum_{c_1=1}^{d_1}\cdots\sum_{c_p=1}^{d_p}|\pi_{c_1c_2\cdots c_p}|=1.$$

Using a tensor generalization of the matrix SVD (Kolda 2001), one can let

$$\pi=\sum_{h=1}^{k}\lambda_h\mathbf{U}_h,\qquad \mathbf{U}_h=\mathbf{u}_h^{(1)}\otimes\mathbf{u}_h^{(2)}\otimes\cdots\otimes\mathbf{u}_h^{(p)},\quad (2)$$

where $\lambda_1\geq\cdots\geq\lambda_k>0$, $\mathbf{U}_h$ is a decomposed tensor, $\mathbf{u}_h^{(j)}\in\Re^{d_j}$, and $\otimes$ denotes the outer product, so that $\pi_{c_1\cdots c_p}=\sum_{h=1}^{k}\lambda_h\prod_{j=1}^{p}u_{hc_j}^{(j)}$. If k is chosen to correspond to the minimum value such that $\boldsymbol{\pi}$ can be expressed as in (2), one obtains a rank decomposition. Rank decompositions are not necessarily unique, but uniqueness is not necessary for our purposes in developing a Bayesian nonparametric procedure.

Corollary 1—For any $\boldsymbol{\pi}\in\Pi_{d_1\cdots d_p}$ decomposed as in (2), we can obtain an equivalent decomposition

$$\pi=\sum_{h=1}^{k}\nu_h\Psi_h,\qquad \Psi_h=\boldsymbol{\psi}_h^{(1)}\otimes\boldsymbol{\psi}_h^{(2)}\otimes\cdots\otimes\boldsymbol{\psi}_h^{(p)},$$

where $\boldsymbol{\nu}=(\nu_1,\ldots,\nu_k)'$ is a probability vector, $\Psi_h\in\Pi_{d_1\cdots d_p}$, and $\boldsymbol{\psi}_h^{(j)}$ is a $d_j\times 1$ probability vector, for $h=1,\ldots,k$ and $j=1,\ldots,p$.

From Corollary 1 it is clear that any multivariate categorical data distribution can be expressed as a latent structure model,

$$\Pr(x_{i1}=c_1,\ldots,x_{ip}=c_p)=x_{c_1\cdots c_p}=\sum_{h=1}^{k}\nu_h\prod_{j=1}^{p}\psi_{hc_j}^{(j)},\quad (3)$$

where $\boldsymbol{\nu}$ is a vector of component probabilities, $z_i\in\{1,\ldots,k\}$ is a latent class index, $\mathbf{x}_i=(x_{i1},\ldots,x_{ip})'$ are conditionally independent given z_i and $\Pr(x_{ij}=c_j|z_i=h)=\psi_{hc_j}^{(j)}$ is the probability of $x_{ij}=c_j$ given allocation of individual i to class h.

2.2 Infinite Mixture of Product Multinomials

Although any multivariate categorical data distribution can be expressed as in (3) for a sufficiently large k, a number of practical issues arise in the implementation. Firstly, it is not straightforward to obtain a well-justified approach for estimation of k. Because the data are often very sparse with most of the cells in the $d_1 \times \cdots \times d_p$ contingency table being empty, a unique maximum likelihood estimate of the parameters in (3) often does not exist even when a modest k is chosen. Such problems may lead one to choose a very small k, which may be in-sufficiently large to provide an adequate approximation to the true multivariate distribution. Hence, inferences on the dependence structure in the observations may be severely biased.

These issues provide motivation for a Bayesian nonparametric approach, which avoids selection of a single finite k, allowing the number of components that are occupied by individuals in the sample to grow with sample size, while also allowing model averaging over the posterior distribution for the number of components. Such model averaging is preferable to approaches that conduct inferences conditional on a selected k, and hence ignore the (often substantial) uncertainty in estimation of k.

We propose to induce a prior, $\boldsymbol{\pi} \sim P$, through the following specification

$$\begin{gathered} \pi = \sum_{h=1}^{\infty} \nu_h \Psi_h, \quad \Psi_h = \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)}, \\ \boldsymbol{\psi}_h^{(j)} \sim P_{0j}, \quad \text{independently for } j = 1, \ldots, p, \; h = 1, \ldots, \infty, \\ \nu \sim Q, \end{gathered} \tag{4}$$

where P_{0j} is a probability measure on the d_j-dimensional probability simplex, $\mathcal{S}_{d_j}$, and Q is a probability measure on the countably infinite probability simplex, $\mathcal{S}_\infty$. For example, P_{0j} may correspond to a Dirichlet measure, while Q corresponds to a Dirichlet process or more broadly to a GEM or Poisson–Dirichlet measure. In the Dirichlet process case, the stick-breaking representation of Sethuraman (1994) implies that $\nu_h = V_h \prod_{l<h} (1 - V_l)$ with $V_h \sim$ beta(1, α) independently for $h = 1, \ldots, \infty$, where $\alpha > 0$ is a precision parameter characterizing Q. For small values of α, ν_h decreases towards zero rapidly with increases in the index h, so that the prior favors a sparse representation with most of the weight on few components. By choosing a hyperprior for α, one can allow the data to inform about an appropriate degree of sparsity, with the intrinsic Bayes penalty for model complexity protecting against overfitting.

From Section 2.1, it is clear that any $\boldsymbol{\pi}^0 \in \Pi_{d_1 \cdots d_p}$ can be characterized using the infinite mixture representation in the first line of (4). However, in placing prior distributions on the components characterizing (4), it is not immediately clear that this flexibility is maintained. Certainly, this is not true for any choice of $P_0 = P_{01} \otimes \cdots \otimes P_{0p}$ and Q. For this reason, it is necessary to place conditions on P_0 and Q so that P has full support on $\Pi_{d_1 \cdots d_p}$ with respect to some appropriate topology.

Theorem 2—Letting $\mathcal{N}_\varepsilon(\boldsymbol{\pi}^0) = \{\boldsymbol{\pi} : \|\boldsymbol{\pi} - \boldsymbol{\pi}^0\|_1 < \varepsilon\}$ denote an L_1 neighborhood around an arbitrary $\boldsymbol{\pi}^0 \in \Pi_{d_1 \cdots d_p}$, the probability $P\{\mathcal{N}_\varepsilon(\boldsymbol{\pi}^0)\} > 0$ for any $\varepsilon > 0$ under the following conditions:

- **i.** $P_{0j}\{\mathcal{N}_\varepsilon(\boldsymbol{\psi}_0^{(j)})\} > 0$ for any $\boldsymbol{\psi}_0^{(j)} \in \mathcal{S}_{d_j}$ and $\varepsilon > 0$, $j = 1, \ldots, p$;
- **ii.** $Q\{\mathcal{N}_\varepsilon(\boldsymbol{\nu}_0)\} > 0$ for any $\varepsilon > 0$ and $\boldsymbol{\nu}_0 \in \mathcal{S}_\infty$ such that $\nu_{0h} = 0$ for $h > r$ with r the (finite) maximum possible rank of $\boldsymbol{\pi}$.

For P_{0j} chosen as a finite Dirichlet and Q a Dirichlet process, it is straightforward to verify that the conditions of Theorem 2 are satisfied, so the resulting prior P has L_1 support on $\Pi_{d_1 \cdots d_p}$. Because there are finitely many parameters in $\boldsymbol{\pi}$, this condition is sufficient for posterior consistency, with

$$\lim_{n\to\infty} \Pr\{\pi \in \mathcal{N}_\varepsilon(\pi^0)|\mathbf{X}\} \to 1 \quad \text{for any } \varepsilon>0,$$

where $\boldsymbol{\pi}^0$ is the true value of $\boldsymbol{\pi}$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ is the data for subject i, and $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is the data for all n subjects.

3. POSTERIOR COMPUTATION

3.1 Slice Sampling Algorithm

In the special case in which P_{0j} are finite Dirichlet and Q is a Dirichlet process, posterior computation under prior (4) can proceed via a simple-to-implement and efficient Gibbs sampling algorithm, which modifies the slice sampling approach of Walker (2007). To clarify, the model can be expressed in the following hierarchical form in this case:

$$\begin{aligned} x_{ij} &\sim \text{Multinomial}\left(\{1,\ldots,d_j\}, \psi^{(j)}_{z_i 1}, \ldots, \psi^{(j)}_{z_i d_j}\right), \\ z_i &\sim \sum_{h=1}^{\infty} V_h \prod_{l<h}(1-V_l)\delta_h, \qquad V_h \sim \text{beta}(1,\alpha), \\ \psi^{(j)}_h &\sim \text{Dorichlet}(a_{j1}, \ldots, a_{jc_j}), \end{aligned} \tag{5}$$

where the x_{ij}'s are sampled independently conditional on the latent class z_i and the stick-breaking random variables $\mathbf{V} = \{V_h\}$ and probability vectors $\boldsymbol{\psi} = \{\boldsymbol{\psi}^{(j)}_h\}$ are mutually independent. We introduce a vector of latent variables, $\mathbf{u} = (u_1, \ldots, u_n)'$, and define the joint likelihood of $\mathbf{u}$ and $\mathbf{X}$ given $\mathbf{V}$ and $\boldsymbol{\psi}$ as

$$\prod_{i=1}^{n}\left\{\sum_{h=1}^{\infty} 1(u_i<\nu_h)\prod_{j=1}^{p}\prod_{l=1}^{d_j}(\psi^{(j)}_{hl})^{1(x_{ij}=l)}\right\}, \tag{6}$$

where $\nu_h = V_h \Pi_{l<h}(1 - V_l)$, for $h = 1, \ldots, \infty$. In marginalizing out $\mathbf{u}$, it is clear that (6) is consistent with (5). The augmented joint posterior distribution is proportional to the prior on $\mathbf{V}$, $\boldsymbol{\psi}$ and α multiplied by the augmented data likelihood (6). Conditional posterior distributions for each of the unknowns can be derived using standard algebra, and a data augmentation Gibbs sampler is then used for posterior computation.

This Gibbs sampler iterates through the following sampling steps:

1. Update u_i, for $i = 1, \ldots, n$, by sampling from the conditional posterior, Uniform(0, ν_{z_i}).
2. For $h = 1, \ldots, k^*$, with $k^* = \max\{z_1, \ldots, z_n\}$, update $\boldsymbol{\psi}^{(j)}_h$ from the full conditional posterior distribution obtained in updating the Dirichlet prior with the likelihood of the jth response for subjects in component h,

$$\text{Dirichlet}\left(a_{j1}+\sum_{i:z_i=h}1(x_{ij}=1),\ldots,a_{jd_j}+\sum_{i:z_i=h}1(x_{ij}=d_j)\right).$$

3. For $h = 1, \ldots, k^*$, sample V_h from the full conditional posterior, which is beta(1, α) truncated to fall into the interval

$$\left[\max_{i:z_i=h}\left\{\frac{u_i}{\prod_{l<h}(1-V_l)}\right\},\ 1-\max_{i:z_i>h}\left\{\frac{u_i}{V_{z_i}\prod_{l<z_i,l\neq h}(1-V_l)}\right\}\right].$$

4. To update z_i, for $i = 1, \ldots, n$, sample from the multinomial full conditional with

$$\Pr(z_i=h|\cdots)=\frac{1(h\in A_i)\prod_{j=1}^{p}\psi_{hx_{ij}}^{(j)}}{\sum_{l\in A_i}\prod_{j=1}^{p}\psi_{lx_{ij}}^{(j)}},$$

where $A_i = \{h : \nu_h > u_i\}$. To identify the elements in $A_1, \ldots, A_n$, first update V_h, for $h = 1, \ldots, \tilde{k}$, where $\tilde{k}$ is the smallest value satisfying $\sum_{h=1}^{\tilde{k}}\nu_h > 1-\min\{u_1,\ldots,u_n\}$.

5. Finally, assuming a gamma(a_α, b_α) hyperprior for α, with the gamma parameterized to have mean a_α/b_α and variance a_α/b_α^2, the conditional posterior is

$$\text{gamma}\left(a_\alpha+k^*,\ b_\alpha-\sum_{h=1}^{k^*}\log(1-V_h)\right).$$

These steps are quite similar to those presented in Walker (2007), though he considered Dirichlet process mixtures of normals. Each step involves sampling from a standard distribution, so should be very simple and efficient to implement. Note that in using the slice sampler we avoid the need to approximate (4) through truncation, and only update those components that are needed. This is conceptually related to the retrospective sampling approach of Papaspiliopoulos and Roberts (2008), though the slice sampler is simpler to implement. One can potentially gain efficiency while maintaining simple sampling steps by combining the slice sampler with the retrospective sampler using the approach described by Papaspiliopoulos (2008). Such an algorithm is similar to our proposed algorithm, but updates the V_hs from a beta conditional posterior distribution obtained in marginalizing out the latent variables **u**.

3.2 Testing and Inferences

It is often of interest to test for independence of the elements of $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$. For example, in the transcription factor binding motif application, there has been considerable debate in the literature regarding the appropriateness of the independence assumption implicit in the product multinomial model. Under our proposed formulation, the null hypothesis of independence is nested within a nonparametric alternative that accommodates a sequence of models of increasing complexity including the saturated model. In particular, the independence model corresponds to $H_0 : \nu_1 = 1$. As noted in Berger and Sellke (1987), interval null hypotheses are often preferred to point null hypotheses. Motivated by this reasoning and by computational considerations, we focus instead on the interval null $H_0 : \nu^* > 1 - \varepsilon$, with $\nu^* = \max\{\nu_h, h = 1, \ldots, k^*\}$ and ε a prespecified small positive constant. In

particular, ε is chosen to permit deviations from independence that are so small as to be nonsignificant in the subject matter area. We find that $\varepsilon = 0.05$ provides a good default value based on simulations from the prior in a variety of cases. For very small ε, the interval null provides an approximation to the point null.

Under the prior proposed in Section 3.1, the probability allocated to H_0 is approximately

$$\Pr(\nu_1 > 1-\varepsilon) = \int_{1-\varepsilon}^{1} \int_0^{\infty} \alpha(1-V_1)^{\alpha-1} \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha - 1} \exp(-b_\alpha \alpha)\, d\alpha\, dV_1$$

which can easily be calculated numerically. From this expression, it is clear that the hyperparameters a_α and b_α control the prior probability allocated to H_0. In order to assign approximately equal probabilities to H_0 and H_1, one should choose values of a_α, b_α to produce $\Pr(H_0) \approx 0.5$. Noting that $a_\alpha + b_\alpha$ is often interpreted as a prior sample size, we recommend letting $a_\alpha + b_\alpha = 1/2$ as a default corresponding to a vague prior. As ν_1 is a parameter in each of the models under consideration, it is acceptable to choose a vague prior. We then choose $a_\alpha = 1/4$ to obtain $\Pr(H_0 : \nu^* > 1 - \varepsilon) \approx 0.5$. To complete a default prior specification, we let $a_{j1} = \cdots = a_{jc_j} = 1$, for $j = 1, \ldots, p$, which corresponds to choosing uniform priors for the category probabilities in each class for each outcome type.

To conduct a Bayesian test of independence, we can rely on the Bayes factor in favor of the alternative hypothesis, which corresponds to

$$\mathrm{BF} = \frac{\Pr(H_1|\mathbf{X})/\Pr(H_1)}{\Pr(H_0|\mathbf{X})/\Pr(H_0)}, \quad (7)$$

which can easily be estimated based on the output of the Gibbs sampler proposed in Section 3.1, with $\widehat{\Pr}(H_1|\mathbf{X})$ equal to the proportion of samples for which $\nu^* \leq 1 - \varepsilon$ and $\widehat{\Pr}(H_0|\mathbf{X}) = 1 - \widehat{\Pr}(H_1|\mathbf{X})$. The performance of the Bayes factor-based test is assessed in a simulation study in Section 4.

In addition to testing for independence, it may be of interest to estimate the marginal distribution for each x_{ij} and to conduct inferences on the dependence structure in the different elements of $\mathbf{x}_i$. To obtain approximate draws from the posterior distribution for $\pi_{jl} = \Pr(x_{ij} = l)$, which can be used to obtain point and interval estimates, we suggest using the approximation

$$\pi_{jl} \approx \sum_{h=1}^{k^*} \nu_h \psi_{hl}^{(j)},$$

where k^* is the maximum class index. Clearly, unless the sample size is quite small, the error in this approximation is neglible, with $1 - \sum_{h=1}^{k^*} \nu_h$ providing an easily calculated upper bound on this error. In small samples, one can obtain an approximation that is as accurate as desired by drawing ν_h's and $\psi_h^{(j)}$ for $h = k^* + 1, \ldots, k_m$, with k_m chosen so that the error bound is below a prespecified constant. This idea is conceptually related to that proposed by Muliere and Tardella (1998).

To conduct inferences on the dependence between x_{ij} and $x_{ij'}$ for all $j, j' \in \{1, \ldots, p\}$, it is necessary to first choose a measure of association for two nominal variables. Most of the symmetric measures of association that are commonly used are based on the Pearson χ^2, with various adjustments proposed to attempt to remove the dependence on sample and table size. For example, Cramer's $V = \sqrt{\chi^2/n(k-1)}$, with $k = \min\{d_j, d_{j'}\}$, is a commonly used measure that ranges from 0 to 1. From a Bayesian perspective, the measure of association should not depend on the data directly, but should be a function of the parameters characterizing the multivariate distribution. Hence, we recommend

$$\rho^2_{jj'} = \frac{1}{\min\{d_j, d_{j'}\} - 1} \sum_{c_j=1}^{d_j} \sum_{c_{j'}=1}^{d_{j'}} \frac{(\pi_{c_j c_{j'}} - \overline{\psi}^{(j)}_{c_j} \overline{\psi}^{(j')}_{c_{j'}})^2}{\overline{\psi}^{(j)}_{c_j} \overline{\psi}^{(j')}_{c_{j'}}}, \text{ with } \overline{\psi}^{(j)}_{l} = \sum_{h=1}^{k^*} \nu_h \psi^{(j)}_{hl}, \quad (8)$$

where $\rho_{jj'}$ is a model-based version of Cramer's V and ranges from 0 to 1, with $\rho_{jj'} \approx 0$ when x_{ij} and $x_{ij'}$ are independent.

Our Bayesian approach has the nice feature that we can estimate the posterior distribution of $\rho_{jj'}$ for all j, j' pairs based on the output of the Gibbs sampler from Section 3.1. To conduct inference on the dependence structure between each of the pairs, we recommend reporting a $p \times p$ association matrix, with the elements corresponding to posterior means for each $\rho_{jj'}$. In addition, we can calculate posterior probabilities and Bayes factors for local null hypotheses, $H_{1jj'}: \rho_{jj'} > \varepsilon$ from the Gibbs output.

Using posterior probabilities of $H_{1jj'}$ for each j, j' pair as a basis for inferences on the dependence structure has substantial advantages over model selection-based approaches. In particular, in a standard Bayesian network analysis, one would estimate a directed acyclic graph (DAG) characterizing the dependence structure in the elements of $\mathbf{x}_i$. Unless p is small, this estimation involves a massive-dimensional model selection problem, as the number of possible DAGs is enormous. In very large model spaces, the available data are typically consistent with a large number of different models, and it is extremely unlikely that the estimated DAG will correspond to the true model.

Although we focus on testing of global dependence and dependence between pairs of variables, since these are the hypotheses most often of interest, our approach can be adapted for comparing competing dependence structure models. For example, one may have two or more alternative graphical models of dependence representing different biological hypotheses. In this case, it is likely that none of the graphical models in the list of those being compared is exactly true. In this setting, one can represent the true model using our proposed Dirichlet process mixture of product multinomial models. The graphical model is then selected that produces a predictive distribution closest to that for the nonparametric model in KL distance using the methods of Walker and Gutiérrez-Peña (2007).

4. SIMULATION STUDIES

To assess the performance of the proposed approach, we conducted a simulation study. Simulated data consisted of A, C, G, T nucleotides ($d_j = d = 4$) at $p = 20$ positions for $n = 100$ sequences. We considered two simulation scenarios, generating the nucleotides (1) independently, and (2) assuming dependence in locations 2, 4, 12, and 14. One mechanism by which positional dependence can be induced is the existence of multiple subpopulations, with each subpopulation having different A, C, G, T probabilities at certain locations. Within a subpopulation, nucleotides are positionally independent. However, marginalizing

out the latent subpopulation indicator, one obtains dependence in those locations that have different nucleotide frequencies across the subpopulations. The presence of such sub-populations is biologically motivated, with different subpopulations potentially having a different combination of dominant (i.e., high probability) nucleotides at certain locations. Presumably there is more than one combination that can lead to a functioning motif, so that different working combinations can arise through evolution in isolated groups of individuals. We assumed 80% of individuals fall into the first subpopulation, with the remaining individuals in a second subpopulation. Nucleotides were generated from a discrete uniform except at locations 2, 4, 12, and 14. For these locations, we used a product multinomial, with the probabilities of A, C, G, or T differing between the subpopulations in case 2 but not in case 1.

For each case, we generate 100 simulated datasets, analyzing each separately using the default prior of Section 3.2 and the Gibbs sampler of Section 3.1 run for 100,000 iterations with a 20,000 iteration burn-in. Mixing and convergence rates were good based on examination of trace plots. We also implemented the Bayes network search method of Xie and Geng (2008) for comparison, using a threshold of 0.05 on the p-values for inclusion of an edge.

For simulation case 1 (no positional dependence, H_0 is true), we obtained good results, with the estimated values of $\rho_{jj'}$ close to zero for each j, j' pair and all of the simulation replicates. Figure 1(a) provides a histogram showing the estimated posterior probabilities of H_1 across the 100 simulated datasets using $\varepsilon = 0.10$. The method appropriately assigns a value close to zero to $\Pr(H_1|\text{data})$ when H_0 is true in most cases, with only 1/100 having an estimated $\Pr(H_1|\text{data})>0.5$. In contrast, the Xie and Geng (2008) approach badly over-fit the data, including dependence in 26.7 pairs of variables on average (range = [5, 171] out of 190 possible). These results improved somewhat using a threshold of 0.01 for inclusion of an edge, but the average number of false positives was still 3.5 (range = [0, 25]).

Figure 1(b) provides results in simulation case 2. The posterior probability assigned to H_1 was close to one in the majority of the simulations. The left panel of Figure 2 shows the proportions of simulations having $\Pr(H_{1jj'}|\mathbf{X}) > 0.95$. These proportions are 0/100 for all position pairs that are truly uncorrelated and ranged from 53/100 to 97/100 for position pairs that were dependent. The right panel of Figure 2 shows results for the Xie and Geng (2008) approach, plotting the proportion of simulations in which each position pair is flagged as correlated. The Xie and Geng (2008) approach had slightly higher power, but a substantially higher Type I error rate. Their method detected 5.7 of the 6 truly dependent pairs of variables on average (range = [3, 6]), while also reporting 36.7 false positives on average (range = [10, 105]). For a threshold of 0.01 for edge inclusion, the average number of true and false positives detected were 5.36 and 5.6, respectively.

The relatively poor performance of the Xie and Geng (2008) method does not reflect on their approach for BN search, but instead on the general drawbacks of relying on model selection in very large model spaces. For comparison, we conducted a simple frequentist alternative, which avoids BN search by focusing on pairwise testing, while adjusting for multiplicity. In particular, we implemented separate chi-square tests for each position pair, flagging those positions having p-values below the Benjamini and Hochberg (1995) threshold to maintain a false discovery rate ≤ 0.05. We find that dependence in (4, 12) and (12, 14) is detected in all of the simulations, but dependence in positions (2, 4), (2, 12), (2, 14), and (4, 14) was missed in all 100 simulations.

Hence, based on these simulations, the proposed Bayesian nonparametric approach has better performance than Bayes network-based methods and frequentist pairwise testing. In

addition to bypassing the difficulties involved in graphical model search, our approach has the conceptual advantage of providing a weight of evidence that two variables are correlated accounting for uncertainty in the dependence in other variables and automatically adjusting for dependence in the different hypothesis tests. Accounting for dependence in hypotheses induces an automatic Bayes adjustment for multiple testing, as discussed in the setting of simpler parametric models by Scott and Berger (2006).

5. APPLICATION TO MODELING POSITIONAL DEPENDENCE WITHIN MOTIFS

We applied the method to data for the p53 transcription factor binding motif. The data consisted of A, C, G, T nucleotides ($d_j = d = 4$) with $p = 20$ positions for $n = 574$ sequences obtained from 542 high-confidence p53 binding loci identified by ChIP-PET experiments (Wei et al. 2006). The p53 tumor suppressor is a sequence-specific DNA binding transcription factor. It regulates the expression of genes involved in a variety of cellular functions, including cell-cycle arrest, DNA repair, and apoptosis (Vogelstein, Lane, and Levine 2000). Hence, p53 provides a natural starting application for our methodology for flexibly characterizing and testing for positional dependence within gene sequences.

We applied the same approach implemented in the simulation study examples. The null hypothesis test for p53 data gave that $\widehat{\Pr}(H_0|\mathbf{X}) = 1 - \widehat{\Pr}(H_1|\mathbf{X}) = 0.00$, with $\widehat{\Pr}(H_1|\mathbf{X})$ equal to the proportion of samples from which $\max\{\nu_h | h = 1,2, \ldots, k^*\} \leq 1 - \varepsilon$, where ε is set to be 0.10 (identical results were obtained for 0.01 and 0.05). To estimate the pairwise positional dependence structure, we used the $\rho_{jj'}$ correlation measure defined in expression (8). Figure 3(a) shows the posterior means of $\rho_{jj'}$ for each pair $j, j' \in \{1, \ldots, 20\}$, while Figure 3(b) shows the frequentist Cramer's V estimates applied separately to each pair of locations. Our estimates are consistent with the Cramer's V estimates in that we assign relatively high correlations to similar pairs of locations. However, as expected in using our Bayes sparseness-favor approach, the smaller correlations are shrunk closer to zero.

Figure 3(c) shows the estimated posterior probabilities of $H_{1jj'}: \rho_{jj'} > 0.1$, while Figure 3(d) shows $-\log_{10} p$-values from pairwise χ^2 tests. Given that $-\log_{10} 0.05 = 1.3$, it is apparent in examining Figure 3(d) that there is a large number of very small p-values. In fact, choosing those position pairs that satisfy the Benjamini and Hochberg (1995) threshold to maintain a false discovery rate (FDR) ≤ 0.05, we obtain 126 pairs, with a large number flagged even for FDR ≤ 0.01. The Xie and Geng (2008) approach selected all 190 pairs as dependent when using a p-value threshold of 0.05 or 0.01 for edge inclusion. In contrast, our Bayesian procedure selects only 16 pairs of positions having $\Pr(H_{1jj'}|\mathbf{X}) > 0.95$. The pairs of positions selected by our method are shown in Figure 4. Given the much lower Type I error rate of our Bayes nonparametric approach in the simulation study, we expect our results are closer to the truth.

6. DISCUSSION

This article has proposed a Bayesian nonparametric approach for inference in sparse contingency tables constructed for multivariate nominal data. We focused in particular on Dirichlet process mixtures of product multinomial models, which can be shown to be highly flexible and computationally convenient. In fact, we find the proposed Gibbs sampler for posterior computation to be quite efficient, which is promising in terms of scaling up the approach to deal with several problems that are of substantial interest. The first is flexible modeling of dependence in large numbers of single nucleotide polymorphisms. As long as the dependence structure can be characterized using a sparse mixture, with most of the weight on few components in the latent structure decomposition, then scaling up to much

larger p should be straightforward. In addition to inferences on dependence across the genome, one important application is to accommodating missing genotype data, which is straightforward within the proposed framework. In fact, following the argument of Formann (2007), one can even allow violations of the missing at random assumption.

A common focus of statistical analyses of sequence data is searching for new motifs in long nucleotide sequences, with such searches attempting to identify words (sequences of nucleotides) that are conserved across sequences collected for different subjects. Most search algorithms make the assumption of independence across the positions within a motif. However, the Gibbs sampling approach of Liu, Neuwald, and Lawrence (1995) can potentially be modified to allow dependence through our proposed approach, so that one can allow for dependence in searching for motifs. Jensen and Liu (2008) recently used a Dirichlet process-based approach for clustering of different transcription factor binding motifs allowing for different lengths.

Another interesting extension of our proposed approach would be to include predictors, $\mathbf{w}_i = (w_{i1}, \ldots, w_{iq})'$, that potentially impact the joint distribution of $\mathbf{x}_i$. This can be accomplished in a sparse manner by allowing the weights to be predictor dependent, so that ν_h is replaced by $\nu_{ih} = \nu_h(\mathbf{w}_i)$. A predictor-dependent stick-breaking process, which generalizes the Dirichlet process, can then be defined for these weight functions. For example, the kernel stick-breaking process of Dunson and Park (2008) can be used directly.

References

Ansari A, Iyengar R. Semiparametric Thurstonian Models for Recurrent Choices: A Bayesian Analysis. Psychometrika. 2006; 71:631–657.

Ashford JR, Sowden RR. Multivariate Probit Analysis. Biometrics. 1970; 26:535–546. [PubMed: 5480663]

Barash, Y.; Elidan, G.; Friedman, M.; Kaplan, T. Modeling Dependence in Protein-DNA Binding Sites. RECOMB 2003; Berlin, Germany. April 2003; 2003.

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Ser B. 1995; 57:289–300.

Berger JO, Sellke T. Testing a Point Null Hypothesis—The Irreconcilability of p-Values and Evidence. Journal of the American Statistical Association. 1987; 82:112–122.

Bulyk ML, Johnson PLF, Church GM. Nucleotides of Transcription Factor Binding Sites Exert Interdependent Effects on the Binding Affinities of Transcription Factors. Nucleic Acids Research. 2002; 30:1255–1261. [PubMed: 11861919]

Chib S, Greenberg E. Analysis of Multivariate Probit Models. Biometrika. 1998; 85:347–361.

Dunson DB, Park J-H. Kernel Stick-Breaking Processes. Biometrika. 2008; 95:307–323. [PubMed: 18800173]

Ferguson TS. A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics. 1973; 1:209–230.

Ferguson TS. Prior Distributions on Spaces of Probability Measures. The Annals of Statistics. 1974; 2:615–629.

Formann AK. Mixture Analysis of Multivariate Categorical Data With Covariates and Missing Entries. Computational Statistics & Data Analysis. 2007; 51:5236–5246.

Gilula Z. Singular Value Decomposition of Probability Matrices: Probabilistic Aspects of Latent Dichotomous Variables. Biometrika. 1979; 66:339–344.

Good IJ. Some Applications of the Singular Decomposition of a Matrix. Technometrics. 1969; 11:823–831.

Goodman LA. Explanatory Latent Structure Assigning Both Identifiable and Unidentifiable Models. Biometrika. 1974; 61:215–231.

Hoff PD. Subset Clustering of Binary Sequences, With an Application to Genomic Abnormalities. Biometrics. 2005; 61:1027–1036. [PubMed: 16401276]

Jara A, Garcia-Zattera MJ, Lesaffre E. A Dirichlet Process Mixture Model for the Analysis of Correlated Binary Responses. Computational Statistics & Data Analysis. 2007; 51:5402–5415.

Jensen ST, Liu JS. Bayesian Clustering of Transcription Factor Binding Motifs. Journal of the American Statistical Association. 2008; 103:188–200.

Kolda TG. Orthogonal Tensor Decompositions. SIAM Journal on Matrix Analysis and Applications. 2001; 23:243–255.

Kottas A, Müller P, Quintana F. Nonparametric Bayesian Modeling for Multivariate Ordinal Data. Journal of Computational and Graphical Statistics. 2005; 14:610–625.

Lawrence CE, Reilly AA. An Expectation Maximization (EM) Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences. Proteins. 1990; 7:41–51. [PubMed: 2184437]

Lazarsfeld, PF.; Henry, NW. Latent Structure Analysis. Boston, MA: Houghton Mifflin; 1968.

Liu JS. The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem. Journal of the American Statistical Association. 1994; 89:958–966.

Liu JS, Neuwald AN, Lawrence CE. Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. Journal of the American Statistical Association. 1995; 90:1156–1170.

Muliere P, Tardella L. Approximating Distributions of Random Functionals of Ferguson–Dirichlet Priors. Canadian Journal of Statistics. 1998; 26:283–297.

Nikolajewa S, Pudimat R, Hiller M, Platzer M, Backofen R. BioBayesNet: A Web Server for Feature Extraction and Bayesian Network Modeling of Biological Sequence Data. Nucleic Acids Research. 2007; 35:W688–W693. [PubMed: 17537825]

Ochi Y, Prentice RL. Likelihood Inference in a Correlated Probit Regression Model. Biometrika. 1984; 71:531–544.

Papaspiliopoulos, O. Working Paper 08-20. Centre for Research in Statistical Methodology, University Warwick; Coventry, U.K: 2008. A Note on Posterior Sampling From Dirichlet Process Mixture Models.

Papaspiliopoulos O, Roberts GO. Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models. Biometrika. 2008; 95:169–186.

Quintana FA. Nonparametric Bayesian Analysis for Assessing Homogeneity in $k \times I$ Contingency Tables With Fixed Right Margin Totals. Journal of the American Statistical Association. 1998; 93:1140–1149.

Quintana FA, Newton MA. Computational Aspects of Non-parametric Bayesian Analysis With Applications to the Modeling of Multiple Binary Sequences. Journal of Computational and Graphical Statistics. 2000; 9:711–737.

Scott JG, Berger JO. An Exploration of Aspects of Bayesian Multiple Testing. Journal of Statistical Planning and Inference. 2006; 136:2144–2162.

Sethuraman J. A Constructive Definition of Dirichlet Priors. Statistica Sinica. 1994; 4:639–650.

Tomovic A, Oakeley EJ. Position Dependencies in Transcription Factor Binding Sites. Bioinformatics. 2007; 23:933–941. [PubMed: 17308339]

Vogelstein B, Lane D, Levine AJ. Surfing the p53 Network. Nature. 2000; 408:307–310. [PubMed: 11099028]

Walker SG. Sampling the Dirichlet Mixture Model With Slices. Communications in Statistics—Simulation and Computation. 2007; 36:45–54.

Walker SG, Gutiérrez-Peña E. Bayesian Parametric Inference in a Nonparametric Framework. TEST. 2007; 16:188–197.

Wei CL, et al. A Global Map of p53 Transcription-Factor Binding Sites in the Human Genome. Cell. 2006; 124:207–219. [PubMed: 16413492]

Xie X, Geng Z. A Recursive Method for Structural Learning of Directed Acyclic Graphs. Journal of Machine Learning Research. 2008; 9:459–483.

Zhang X, Boscardin WJ, Belin TR. Bayesian Analysis of Multivariate Nominal Measures Using Multivariate Multinomial Probit Models. Computational Statistics & Data Analysis. 2008; 52:3297–3708.

Zhou Q, Liu JS. Modeling Within-Motif Dependence for Transcription Factor Binding Site Predictions. Bioinformatics. 2004; 20:909–916. [PubMed: 14751969]

APPENDIX A: PROOF OF THEOREM 1

To prove Theorem 1, first note that the singular value decomposition (SVD) expresses $\boldsymbol{\pi}$ as

$$\pi=\sum_{h=1}^{k}\lambda_h\boldsymbol{\xi}_h\boldsymbol{\kappa}_h', \quad \text{(A.1)}$$

where $\lambda_h \geq 0$ are singular values, $\lambda_1^2 \geq \cdots \geq \lambda_k^2 > 0$ are the nonzero eigenvalues of the matrix $\boldsymbol{\pi}\boldsymbol{\pi}'$, $\boldsymbol{\xi}_h$ is a $d_1 \times 1$ vector, $\boldsymbol{\kappa}_h$ is a $d_2 \times 1$ vector, and $\boldsymbol{\xi}=\{\boldsymbol{\xi}_h\}_{h=1}^k$ and $\boldsymbol{\kappa}=\{\boldsymbol{\kappa}_h\}_{h=1}^k$ are orthonormal bases for the spaces spanned by the rows of $\boldsymbol{\pi}\boldsymbol{\pi}'$ and $\boldsymbol{\pi}'\boldsymbol{\pi}$, respectively. The SVD decomposition in (A.1) is unique up to multiplication of $\boldsymbol{\xi}_h$ and $\boldsymbol{\kappa}_h$ by −1. As $\mathbf{A} = \boldsymbol{\pi}\boldsymbol{\pi}'$ is nonnegative, the rows of $\mathbf{A}$ ($\mathbf{A}'$) can be expressed as linear combinations of nonnegative orthonormal bases $\boldsymbol{\xi}$($\boldsymbol{\kappa}$). We remove sign ambiguity in (A.1) by restricting $\boldsymbol{\xi}$ and $\boldsymbol{\kappa}$ to be nonnegative. The spectral radius of the positive definite matrix $\mathbf{A}$ is defined as $\rho(\mathbf{A})=\max_h\{\lambda_h^2\}$. It is known that $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$, where $\|\cdot\|$ denotes an induced norm, such as $\|\mathbf{A}\|_1=\max_{1\leq j\leq d_1}\sum_{i=1}^{d_1}|a_{ij}|$. Because the elements of $\boldsymbol{\pi}$ are nonnegative probabilities summing to one, it is clear that $\sum_{i=1}^{d_1}|a_{ij}| \leq 1$ for all j, so that $\rho(\mathbf{A}) \leq 1$. From the Perron–Frobenius theorem, $\lambda_1^2=\rho(\mathbf{A})$, and hence $0 < \lambda_h \leq 1$ for $h = 1, \ldots, k$.

Using the known restrictions on $\{\lambda_h\}_{h=1}^k$, $\boldsymbol{\xi}$ and $\boldsymbol{\kappa}$, we can equivalently express (A.1) as

$$\pi=\sum_{h=1}^{k}\lambda_h\boldsymbol{\xi}_h\boldsymbol{\kappa}_h'=\sum_{h=1}^{k}\tilde{\lambda}_h\tilde{\boldsymbol{\xi}}_h\tilde{\boldsymbol{\kappa}}_h', \quad \text{(A.2)}$$

where $\boldsymbol{\xi}_h = c_{\xi,h}\tilde{\boldsymbol{\xi}}_h$, $c_{\xi,h}=\sum_{j=1}^{d_1}\xi_{hj}$, $\boldsymbol{\kappa}_h = c_{\kappa,h}\tilde{\boldsymbol{\kappa}}_h$, $c_{\kappa,h}=\sum_{j=1}^{d_2}\kappa_{hj}$, and $\tilde{\lambda}_h = c_{\xi,h}c_{\kappa,h}\lambda_h$. In the reparameterization, $\tilde{\boldsymbol{\xi}}_h$ and $\tilde{\boldsymbol{\kappa}}_h$ are probability vectors, with the elements in $\tilde{\boldsymbol{\xi}}_h\tilde{\boldsymbol{\kappa}}_h'$ summing to one. Because the elements of $\boldsymbol{\pi}$ sum to one, it is straightforward to show that $\sum_{h=1}^{k}c_{\xi,h}c_{\kappa,h}\lambda_h=1$, and hence $\tilde{\boldsymbol{\lambda}}=(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k)'$ is a probability, vector. Noting that (1) can be expressed as $\pi=\sum_{h=1}^{k}\nu_h\boldsymbol{\psi}_h^{(1)}\boldsymbol{\psi}_h^{(2)'}$, Theorem 1 follows directly.

APPENDIX B: PROOF OF COROLLARY 1

The approach used in proving Theorem 1 can be applied directly after showing that expression (2) holds under the constraints (i) $0 < \lambda_h \leq 1$ for $h = 1, \ldots, k$; and (ii) the elements of $\mathbf{u}_h^{(j)}$ are nonnegative for $h = 1, \ldots, k$ and $j = 1, \ldots, p$. Letting $\boldsymbol{\pi}_{[jj']}$ denote the $d_j \times d_{j'}$ matrix containing $\Pr(x_{ij} = c_j, x_{ij'} = c_{j'})$ in element $c_j, c_{j'}$, it follows from (2) that

$$\pi_{[jj']}=\sum_{h=1}^{k}\lambda_h \mathbf{u}_h^{(j)}\otimes \mathbf{u}_h^{(j')},\quad j,j'\in\{1,\ldots,p\},$$

where the rows of $\pi_{[jj']}\pi'_{[jj']}$ can be expressed as linear combinations of $\{\mathbf{u}_h^{(j)}\}_{h=1}^k$. Hence, $\{\mathbf{u}_h^{(j)}\}_{h=1}^k$ forms a basis for the union of the row spaces of $\pi_{[j1]}\pi'_{[j1]},\ldots,\pi_{[jp]}\pi'_{[jp]}$. Because $\boldsymbol{\pi}$ is a nonnegative tensor, the rows of $\pi_{[jj']}\pi'_{[jj']}$ are nonnegative and can be expressed as linear combinations of nonnegative vectors having positive weights. Hence one can choose each of the $\mathbf{u}_h^{(j)}$ vectors to be nonnegative, which satisfies property (ii). Because (2) is not uniquely defined in that $\lambda_h\mathbf{U}_h$ can be replaced by $\lambda_h CC^{-1}\mathbf{U}_h$ for any $C\in\Re^+$, the restriction $0<\lambda_h\leq 1$ can be included trivially.

APPENDIX C: PROOF OF THEOREM 2

Using Corollary 1, $\boldsymbol{\pi}^0$ can be expressed as

$$\pi^0=\sum_{h=1}^{\infty}\nu_{0h}\Psi_{0h},\quad \Psi_{0h}=\boldsymbol{\psi}_{0h}^{(1)}\otimes\cdots\otimes\boldsymbol{\psi}_{0h}^{(p)},$$

where $\boldsymbol{\nu}_0=(\nu_{01},\ldots,\nu_{0k_0},0,\ldots,0)'$, k_0 is the rank of the tensor, $\boldsymbol{\pi}^0$, $\nu_{0h}=0$ for $h=k_0+1,\ldots,\infty$, and $\boldsymbol{\psi}_{0h}^{(1)},\ldots,\boldsymbol{\psi}_{0h}^{(p)}$ are probability vectors with respective dimensions $d_1,\ldots,d_p$. The L_1 distance between $\boldsymbol{\pi}^0$ and an arbitrary $\boldsymbol{\pi}\in\Pi_{d_1\cdots d_p}$ can be defined as

$$\left\|\pi-\pi^0\right\|_1=\sum_{c_1=1}^{d_1}\cdots\sum_{c_p=1}^{d_p}\sum_{h=1}^{\infty}\left|\nu_h\prod_{j=1}^{p}\psi_{hc_j}^{(j)}-\nu_{0h}\prod_{j=1}^{p}\psi_{0hc_j}^{(j)}\right|. \quad \text{(C.1)}$$

which is a function of the probability vectors characterizing the tensor SVD decompositions. Holding $\boldsymbol{\pi}^0$ fixed, the probability allocated to $\mathcal{N}_\varepsilon(\boldsymbol{\pi}^0)$ can then be defined as

$$\int 1(\left\|\pi-\pi^0\right\|_1<\varepsilon)dQ(\nu)\prod_{h=1}^{\infty}\prod_{j=1}^{p}dP_{0j}(\boldsymbol{\psi}_h^{(j)}). \quad \text{(C.2)}$$

For any $\varepsilon>0$, it is straightforward to show that there exist infinitely many values $\tilde{\varepsilon}>0$ and $\varepsilon^*>0$ such that

$$\left\|\nu-\nu_0\right\|_1<\tilde{\varepsilon}\quad\text{and}\quad \|\boldsymbol{\psi}_h^{(j)}-\boldsymbol{\psi}_{0h}^{(j)}\|_1<\varepsilon^*\quad \forall h,j$$

implies that $\|\boldsymbol{\pi}-\boldsymbol{\pi}^0\|_1<\varepsilon$. Hence, to show that (C.2) is strictly positive, it suffices to show

$$\Pr(\left\|\nu-\nu_0\right\|_1<\tilde{\varepsilon},\|\boldsymbol{\psi}_h^{(j)}-\boldsymbol{\psi}_{0h}^{(j)}\|_1<\varepsilon^*,\ h=1,\ldots,\infty,\ j=1,\ldots,p)>0,$$

which follows directly from conditions (i) and (ii) of the Theorem 2 due to independence.

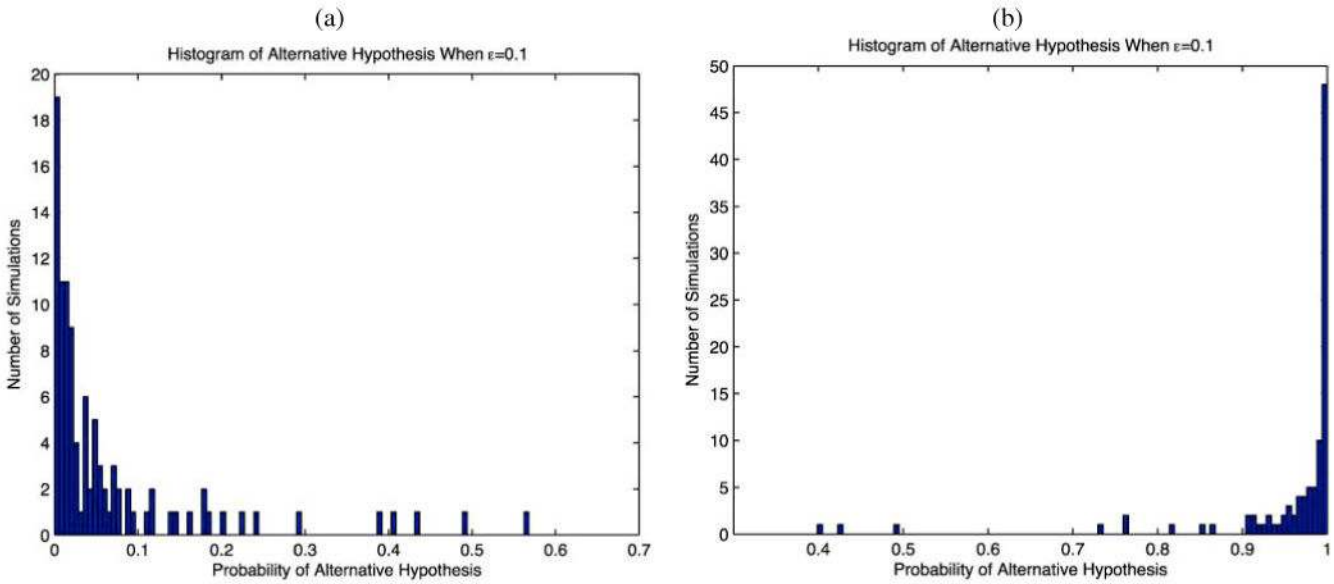

Figure 1.
Histograms of estimated posterior probabilities of H_1 in each of the 100 simulations under (a) case 1 (no positional dependence—H_0 is true) and (b) case 2 (positional dependence—H_1 is true).

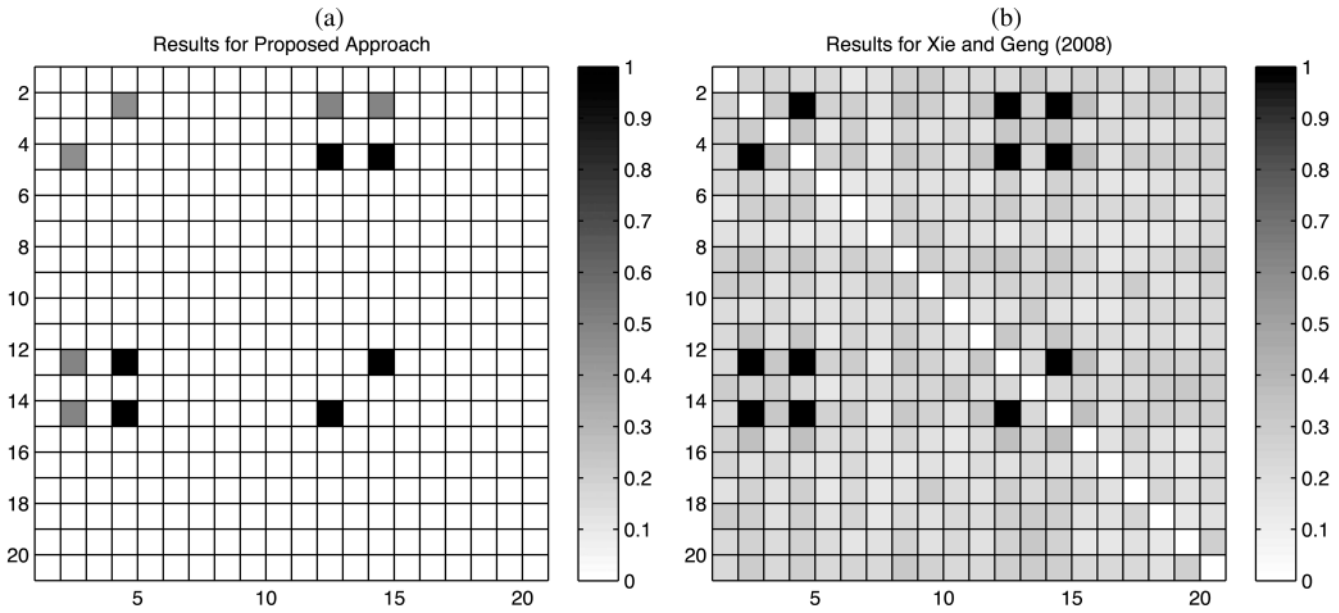

Figure 2.
Results of simulation case 2—percentages of simulations for which (a) $\Pr(H_{1jj'}|\mathbf{X})>0.95$, and (b) the Xie and Geng (2008) method estimated an association between positions j, j'. The true model has dependence in positions 2, 4, 12, and 14.

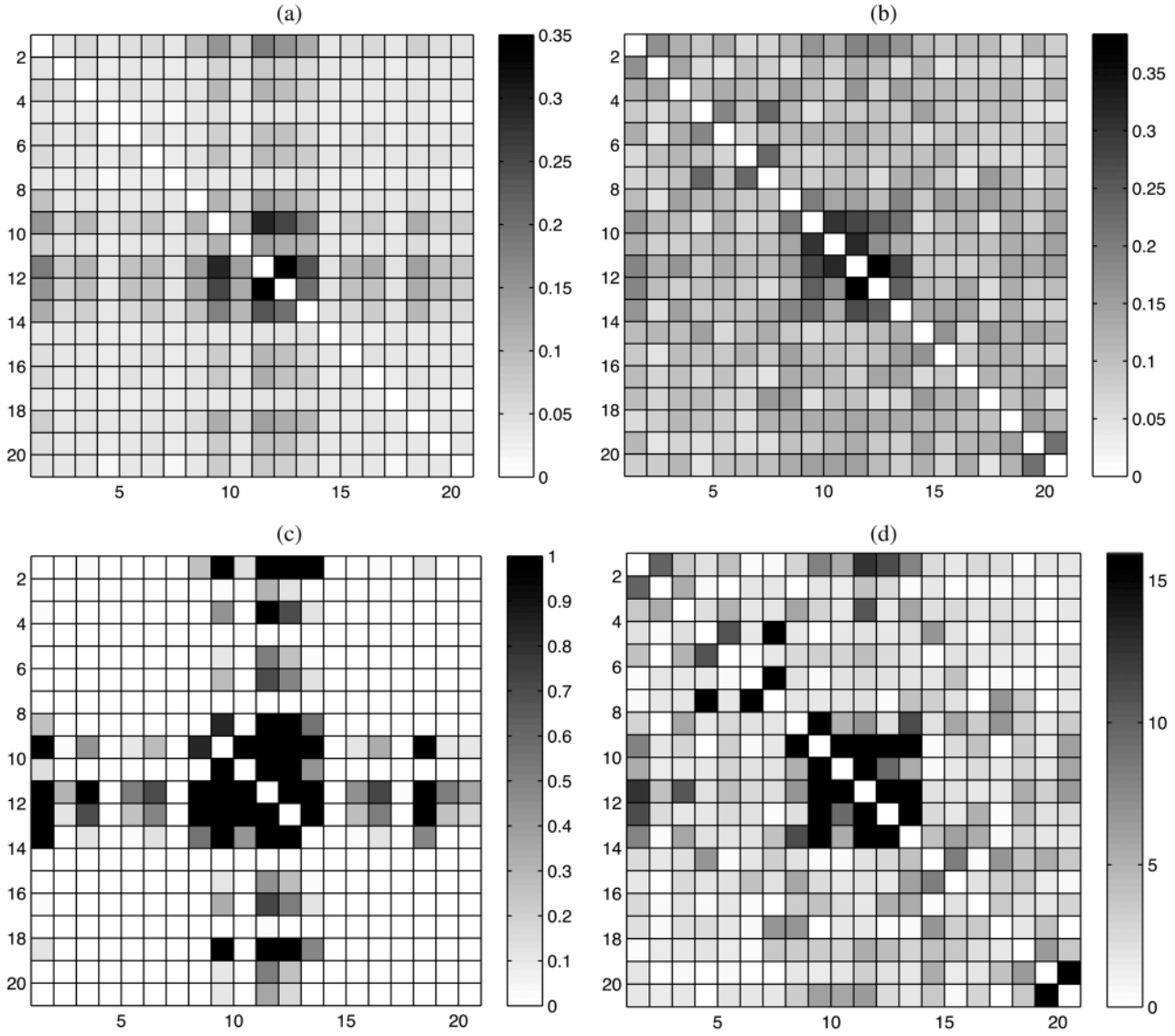

Figure 3.
Results for p53 data. (a) Posterior mean of $\rho_{jj'}$. (b) Pairwise Cramer's V values. (c) Pairwise posterior probabilities, $\Pr(H_{1jj'}|\mathbf{X})$. (d) $-\log_{10}$ p-values from pairwise χ^2 tests.

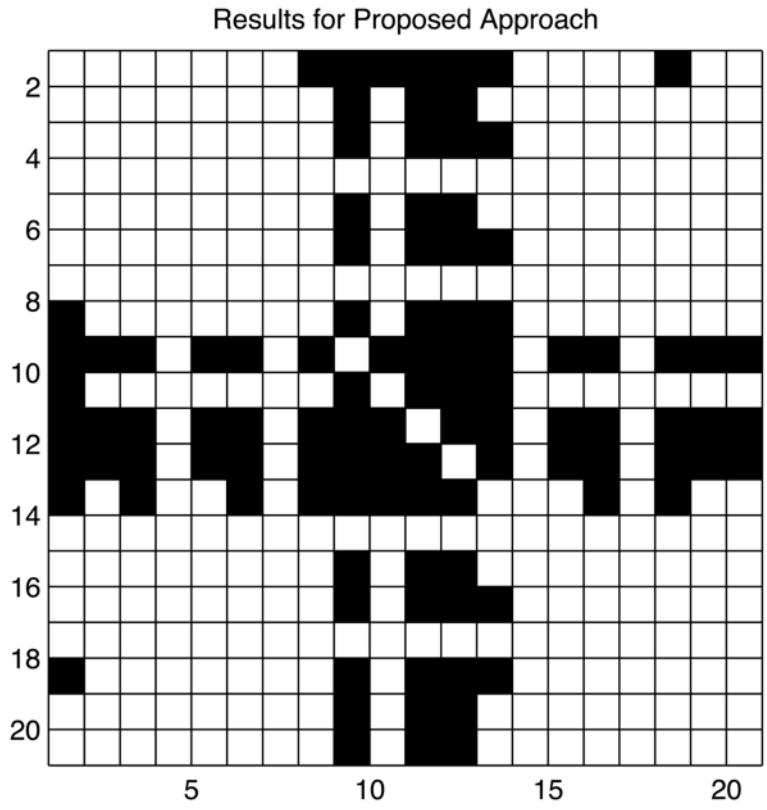

Figure 4.
Pairs of positions flagged as dependent for the p53 data using the proposed Bayesian nonparametric approach with $\Pr(H_{1jj'}|\mathbf{X}) > 0.95$.