Journal of the American Statistical Association

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

Bayesian Modeling of Temporal Dependence in Large Sparse Contingency Tables

Tsuyoshi Kunihama & David B. Dunson

To cite this article: Tsuyoshi Kunihama & David B. Dunson (2013) Bayesian Modeling of Temporal Dependence in Large Sparse Contingency Tables, Journal of the American Statistical Association, 108:504, 1324-1338, DOI: 10.1080/01621459.2013.823866

To link to this article: http://dx.doi.org/10.1080/01621459.2013.823866

View supplementary material

Accepted author version posted online: 25 Jul 2013.
Published online: 25 Jul 2013.

Submit your article to this journal

Article views: 452

View related articles

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=uasa20

Download by: [Board of Governors of the Federal Reserve System] **Date:** 23 October 2017, At: 11:22

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Bayesian Modeling of Temporal Dependence in Large Sparse Contingency Tables

Tsuyoshi KUNIHAMA and David B. DUNSON

It is of interest in many applications to study trends over time in relationships among categorical variables, such as age group, ethnicity, religious affiliation, political party, and preference for particular policies. At each time point, a sample of individuals provides responses to a set of questions, with different individuals sampled at each time. In such settings, there tend to be an abundance of missing data and the variables being measured may change over time. At each time point, we obtained a large sparse contingency table, with the number of cells often much larger than the number of individuals being surveyed. To borrow information across time in modeling large sparse contingency tables, we propose a Bayesian autoregressive tensor factorization approach. The proposed model relies on a probabilistic Parafac factorization of the joint pmf characterizing the categorical data distribution at each time point, with autocorrelation included across times. We develop efficient computational methods that rely on Markov chain Monte Carlo. The methods are evaluated through simulation examples and applied to social survey data. Supplementary materials for this article are available online.

KEY WORDS: Dynamic model; Multivariate categorical data; Nonparametric Bayes; Panel data; Parafac; Probabilistic tensor factorization; Stick-breaking.

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

1. INTRODUCTION

Time-indexed multivariate categorical data are collected in many areas, with partially overlapping categorical variables measured for different subjects at the different time points. As a motivating application, we consider social science surveys that are conducted at regular time intervals, containing many categorical questions such as gender, race, age group, ethnicity, religious affiliation, political party, and preference for particular policies. For such surveys and other types of time-indexed multivariate categorical data, it is common for the variables measured (questions asked) to vary somewhat over time while a subset of the variables will be measured at all times. In addition, the number of variables measured can be moderate to large leading to a contingency table with an *enormous* number of cells, the vast majority of which are empty. Given the fact that social science data often contain complex interactions, it becomes extremely challenging to build realistic and computationally tractable models that allow ultra-sparse data. We define ultra-sparse contingency tables as having exponentially or super-exponentially more cells than the sample size.

Let $\mathbf{x}_{ti} = (x_{ti1}, \ldots, x_{tip})'$ denote the multivariate response for the ith subject in the survey at time t, with the jth categorical question having d_j elements, $x_{tij} \in \{1, \ldots, d_j\}$, $j = 1, \ldots, p$. We accommodate the case in which the specific variables measured can vary across time by introducing missingness indicators, $m_{ti} = (m_{ti1}, \ldots, m_{tip})'$, with $m_{tij} = 1$ if variable j is missing for subject i at time t; we allow design-based missingness in which certain variables are not measured for any subjects at a particular time and for individual-specific missingness in which certain individuals fail to answer all the questions posed to them. In both cases we assume missing at random.

There is a rich literature on the analysis of contingency tables (Agresti 2002; Fienberg and Rinaldo 2007). Log-linear models are perhaps the most commonly used modeling framework. Routine implementations rely on maximum likelihood estimation, though there is also a rich Bayesian literature. For large, sparse contingency tables, maximum likelihood estimates do not exist in many cases except for overly simplistic log-linear models and richer classes of models become challenging to implement computationally. There is a rich literature on graphical modeling approaches to estimating conditional independence structures in categorical variables, with Dobra and Lenkoski (2011) proposing a recent Bayesian approach. Although their method is computationally efficient, except for very small tables, the number of possible graphical models is so enormous that it becomes infeasible to visit more than a vanishingly small fraction of the models making accurate model selection or averaging difficult. To facilitate scaling to large tables, Dunson and Xing (2009) and Bhattacharya and Dunson (2012) recently proposed Bayesian probabilistic tensor factorizations. These methods express the probability tensor corresponding to the joint probability mass function of the categorical variables as a convex combination of independent components. Such methods have not yet been developed for time-indexed contingency tables.

There is a rich literature on categorical time series and longitudinal data analysis in which the same categorical variable is repeatedly measured for each subject over time. For example, Markov models, state space models, and random effects models are routinely applied in such settings. However, these models are not relevant to the problem of incorporating dependence over time in modeling of large sparse contingency tables. As subjects can be different over time, we do not focus on the problem of incorporating within-subject dependence in repeated observations; instead our goal is to include dependence in the parameters characterizing the time-dependent joint pmfs for the categorical variables. To our knowledge, this problem has not yet

Tsuyoshi Kunihama is Professor (*tsuyoshi.kunihama@stat.duke.edu*) and David B. Dunson is Professor (*dunson@stat.duke.edu*), Department of Statistical Science, Duke University, Durham, NC 27708-0251. This research was supported by Nakajima Foundation and grants ES017436 and ES017240 from the National Institute of Environmental Health Sciences (NIEHS) of the US National Institutes of Health. The computational results are mainly generated using Ox (Doornik 2006).

© 2013 American Statistical Association
Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.823866

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

been addressed in the literature. Although one can potentially adapt log-linear or graphical models developed for contingency tables at one time in a somewhat straightforward manner, the hurdles mentioned above for the static case are compounded in the dynamic setting.

To facilitate routine implementations in ultra-sparse cases, we propose to adopt the Dunson and Xing (DX) (2009) probabilistic Parafac factorization to the dynamic setting. The DX model induces a tensor factorization through a Dirichlet process (DP) mixture of product multinomial distributions for the categorical observations. There is an increasingly rich literature proposing nonparametric Bayes' dynamic models, which allow time-indexed dependent random probability measures. Perhaps the most common approach relies on a dependent DP (MacEachern 1999, 2000), which incorporates time dependence in the weights and/or atoms in a stick-breaking representation (Griffin and Steel 2006; Rodriguez and Horst 2008; Chung and Dunson 2011). Most applications of dependent DPs fix the weights and allow the atoms to vary, as varying weights can lead to computational complexities. For dynamic modeling of contingency tables it is more parsimonious to allow varying weights, and varying atoms can lead to a substantial computational burden. An alternative approach, which allows varying weights in a computationally convenient and flexible manner, relies of dynamic mixtures of DPs (Dunson 2006; Ren et al. 2010). Recently, a class of probit stick-breaking processes was proposed (Chung and Dunson 2009), which has the appealing feature of allowing one to induce time dependence in random probability measures through Gaussian time series models (Rodriguez and Dunson 2011).

We propose a new nonparametric state space model for time-indexed ultra-sparse contingency tables. Relying on a DX-type probabilistic Parafac factorization, we place a dynamic model on the weights, which relies on transformed normal random variables in a similar manner to probit stick-breaking. The model is nonparametric in the sense that the induced prior for each time-indexed joint pmf assigns positive probability in arbitrarily small neighborhoods of any "true" data-generating pmf. Hence, our model can allow higher-order interactions and complex dependencies, while shrinking toward a low-dimensional structure and borrowing information across time to address the curse of dimensionality. In addition, and crucially for the approach to be useful in the motivating applications, posterior computation can be implemented via a highly efficient Markov chain Monte Carlo (MCMC) algorithm relying on a slice sampler related to Kalli, Griffin, and Walker (2011). Finally, the factorization produces a low-dimensional representation of the joint pmf, which is otherwise characterized by a daunting number of parameters in many cases, as the number of cells of the tables can be truly massive.

2. MODEL SPECIFICATION

2.1 Modeling of Multivariate Categorical Data

We review the nonparametric Bayes' approach of Dunson and Xing (2009) for a static large sparse contingency table. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ be multivariate categorical data for the ith subject, with $x_{ij} \in \{1, \ldots, d_j\}$, $j = 1, \ldots, p$. Let

$$\boldsymbol{\pi} = \left\{\pi_{c_1 \cdots c_p},\ c_j = 1, \ldots, d_j,\ j = 1, \ldots, p\right\} \in \Pi_{d_1 \cdots d_p}$$

be a probability tensor where $\pi_{c_1 \cdots c_p} = P(x_{i1} = c_1, \ldots, x_{ip} = c_p)$ is a cell probability and $\Pi_{d_1 \cdots d_p}$ is the set of all probability tensors of size $d_1 \times \cdots \times d_p$. Dunson and Xing (2009) showed that any $\boldsymbol{\pi} \in \Pi_{d_1 \cdots d_p}$ can be decomposed as

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

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)'$ is a probability vector, $\Psi_h \in \Pi_{d_1 \cdots d_p}$ and $\boldsymbol{\psi}_h^{(j)} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)})'$ is a $d_j \times 1$ probability vector for $h = 1, \ldots, k$ and $j = 1, \ldots, p$. This expression relies on a Parafac tensor factorization (Harshman 1970; Kolda 2001). It follows that any multivariate categorical data distribution can be expressed as a mixture of product multinomials,

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

By introducing a latent class index $s_i \in \{1, \ldots, k\}$ for the ith subject, the multivariate responses $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ are conditionally independent given s_i. Instead of conditioning on a fixed k, Dunson and Xing (2009) developed a nonparametric Bayes' approach that lets

$$\begin{aligned}
&\boldsymbol{\pi} = \sum_{h=1}^{\infty} \nu_h \Psi_h, \quad \Psi_h = \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)}, \qquad (2)\\
&\boldsymbol{\psi}_h^{(j)} \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}), \text{ independently for } j = 1, \ldots, p,\\
&\qquad\qquad h = 1, \ldots, \infty,\\
&\nu_h = V_h \prod_{l<h} (1 - V_l),\\
&V_h \sim \text{beta}(1, \alpha), \text{ independently for } h = 1, \ldots, \infty,
\end{aligned}$$

where $a_{jl} > 0$ for $l = 1, \ldots, d_j$ and $\alpha > 0$. Although (2) allows infinitely many components, the number k_n occupied by the n subjects in the sample will tend to be $k_n << n$, so few components will be occupied. The model corresponds to a Dirichlet process mixture of product multinomial distributions relying on a stick-breaking representation (Sethuraman 1994). A prior is induced on the joint pmf which has large support in the sense of assigning positive probability to L_1 neighborhoods of any true joint pmf.

2.2 Modeling of Time-Indexed Multivariate Categorical Data

Relying on the DX-type probabilistic Parafac factorization, we propose a new nonparametric Bayes' approach for time-indexed large sparse contingency tables. In a dynamic setting, we obtain the time-indexed multivariate response $\mathbf{x}_{ti} = (x_{ti1}, \ldots, x_{tip})'$, $x_{tij} \in \{1, \ldots, d_j\}$, for the ith subject at time t for $i = 1, \ldots, n_t$, $t = 1, \ldots, T$ and $j = 1, \ldots, p$. At time t, we have a probability tensor π_t for the multivariate categorical response given by

$$\boldsymbol{\pi}_t = \{\pi_{tc_1 \cdots c_p},\ c_j = 1, \ldots, d_j,\ j = 1, \ldots, p\} \in \Pi_{d_1 \cdots d_p},$$

where $\pi_{tc_1 \cdots c_p} = P(x_{ti1} = c_1, \ldots, x_{tip} = c_p)$ is a cell probability at time t. Relying on the probabilistic Parafac factorization, each probability tensor $\boldsymbol{\pi}_t$ can be expressed as a mixture of

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

product multinomials

$$\boldsymbol{\pi}_t = \sum_{h=1}^{k_t} \nu_{th} \Psi_{th}, \quad \Psi_{th} = \boldsymbol{\psi}_{th}^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_{th}^{(p)}, \tag{3}$$

where $k_t \in \mathbb{N}$, $\boldsymbol{\nu}_t = (\nu_{t1}, \ldots, \nu_{tk_t})'$ is a probability vector, $\Psi_{th} \in \Pi_{d_1 \cdots d_p}$ and $\boldsymbol{\psi}_{th}^{(j)} = (\psi_{th1}^{(j)}, \ldots, \psi_{thd_j}^{(j)})'$ is a $d_j \times 1$ probability vector for $h = 1, \ldots, k_t$. Letting $s_{ti} \in \{1, \ldots, k_t\}$ denote a latent class index for the ith subject at time t, the observations $\mathbf{x}_{ti}$ are conditionally independent given s_{ti}.

To borrow information across time, we place a dynamic structure on the probability tensor $\boldsymbol{\pi}_t$ in (3) assuming time varying weights ν_{th} and static atoms $\psi_{th}^{(j)} = \psi_h^{(j)}$. Time dependence is induced in the weights through a state space model, which assumes that stick-breaking increments on ν_{th} arise through transforming Gaussian autoregressive processes using a monotone differentiable link function $g : \Re \to (0, 1)$. This characterization is motivated by the probit stick-breaking process (Chung and Dunson 2009; Rodriguez and Dunson 2011), and leads to a parsimonious but flexible characterization of time-dependence in joint pmfs underlying large, sparse contingence tables.

Similar to expression (2), we develop a nonparametric Bayes' approach that sets the number of components to $k_t = \infty$, though the number of occupied components will tend to be much less than the sample size and can vary across time. The specific model is

$$\boldsymbol{\pi}_t = \sum_{h=1}^{\infty} \nu_{th} \Psi_h, \quad \Psi_h = \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)}, \tag{4}$$

$$\boldsymbol{\psi}_h^{(j)} \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}), \text{ independently for } j = 1, \ldots, p, \quad h = 1, \ldots, \infty, \tag{5}$$

$$\nu_{th} = g(W_{th}) \prod_{l<h} \{1 - g(W_{tl})\}, \tag{6}$$

$$W_{th} = \alpha_{th} + \varepsilon_{th}, \quad \varepsilon_{th} \sim N\left(0, \sigma_\varepsilon^2\right), \tag{7}$$

$$\alpha_{th} = \mu + \phi \alpha_{t-1h} + \eta_{th}, \quad \eta_{th} \sim N\left(0, \sigma_\eta^2\right), \tag{8}$$

where $|\phi| < 1$, $\{\varepsilon_{th}\}$ and $\{\eta_{th}\}$ are sequences of independently normally distributed random variables with mean 0 and variance σ_ε^2 and σ_η^2, respectively. The parameter ϕ controls the autocorrelation in the weights ν_{th} on the different components over time. For sake of parsimony and simplicity in modeling and computation, we include a single time-stationary correlation parameter ϕ instead of allowing dependence to be time- or element-specific. In the limiting case in which $\phi = 0$, the weights ν_{th} will be modeled as independent. This does not mean that independent priors are placed on the unknown joint pmfs at each time, as the incorporation of common atoms automatically induces some degree of a priori dependence. However, in applications one typically expects that the joint pmfs will be quite similar over time, and by using varying weights one does not rule out arbitrarily large changes in the pmfs over time. When ϕ is close to 1, there will be very high time dependence in the weights, leading to effective collapsing on a model that assumes a single time stationary joint pmf. For the initial state variables, we assume the stationary distributions, $\alpha_{1h} \sim N(\mu/(1-\phi), \sigma_\eta^2/(1-\phi^2))$ independently for $h = 1, \ldots, \infty$. Also, we choose priors $\mu \sim N(\mu_0, \sigma_0^2)$, $\phi \sim U(-1, 1)$, $\sigma_\varepsilon^2 \sim \text{IG}(m_\varepsilon/2, S_\varepsilon/2)$ and $\sigma_\eta^2 \sim \text{IG}(m_\eta/2, S_\eta/2)$, respectively.

Due to the Parafac factorization leading to a massive reduction in the number of parameters, the proposed method can efficiently estimate all the cell probabilities using cells with both positive and zero observed counts; the cells having zero counts can vary over time and are not assumed to be structural zeros. The marginal posterior distributions for the cell probabilities will not be concentrated at 0 even if the observed counts are 0.

Expressions (4)–(8) induce a prior on the time-dependent joint pmfs, but it is not immediately obvious how the chosen hyperpriors in the hierarchical specification impact the properties of the prior for $\{\boldsymbol{\pi}_t\}$. In particular, it is important to obtain characterizations of the moments of the induced prior for the cell probabilities, as well as the prior covariance between different elements and across time. Such expressions are provided in Lemma 1, with the proof in Appendix A. Lemma 2 shows that the prior is well defined in the sense that $\sum_{h=1}^{\infty} \nu_{th}$ converges to 1 almost surely.

Lemma 1. The expectation, variance, and covariance of the joint prior on the elements of $\{\boldsymbol{\pi}_t\}$ induced through (4)–(8) are

$$E\{\pi_{tc_1 \cdots c_p}\} = \prod_{j=1}^{p} \frac{a_{jc_j}}{\hat{a}_j},$$

$$V\{\pi_{tc_1 \cdots c_p}\} = \left(\prod_{j=1}^{p} \frac{a_{jc_j}(a_{jc_j}+1)}{\hat{a}_j(\hat{a}_j+1)} - \prod_{j=1}^{p} \frac{a_{jc_j}^2}{\hat{a}_j^2} \right) \left(\frac{\beta_2}{2\beta_1 - \beta_2} \right),$$

$$\text{cov}\{\pi_{tc_1 \cdots c_p}, \pi_{t+kc_1' \cdots c_p'}\} = \left(\prod_{j=1}^{p} \frac{a_{jc_j}\{a_{jc_j'} + 1(c_j = c_j')\}}{\hat{a}_j(\hat{a}_j+1)} - \prod_{j=1}^{p} \frac{a_{jc_j} a_{jc_j'}}{\hat{a}_j^2} \right) \left(\frac{\gamma_k}{2\beta_1 - \gamma_k} \right),$$

where $\beta_1 = E\{g(W_{th})\}$, $\beta_2 = E\{g^2(W_{th})\}$, $\gamma_k = E\{g(W_{th})\, g(W_{t+kh})\}$, $\hat{a}_j = \sum_{l=1}^{d_j} a_{jl}$ and $1(\cdot)$ is an indicator function.

The expectation of cell probabilities can be expressed as the product of expectations of Dirichlet priors for atoms. The variance and covariance are expressed as the product of two terms, the first one is related to atoms and the second one comes from time varying weights. As $\mu \to \infty$, then $\beta_2/(2\beta_1 - \beta_2) \to 1$ and $\gamma_k/(2\beta_1 - \gamma_k) \to 1$, and the variance and covariance will be influenced only by atoms. In such a case, the measure corresponding to the stick-breaking process will become a point mass at a random atom almost surely. In addition, β_1, β_2, and γ_k do not depend on time t, hence all expectation, variance, and covariance are independent of t though the covariance depends on the time difference k. Also, the covariance between cell probabilities with $c_j = c_j'$ for all j is always positive and, on the other hand, those with $c_j \neq c_j'$ for all j have negative covariance. In a special case in which the hyperparameters in the Dirichlet prior are $a_{j1} = \cdots = a_{jd_j} = a$ the variance and covariance is 0 in the limit as $a \to \infty$. The proof is in Appendix A.

Lemma 2. $\sum_{h=1}^{\infty} \nu_{th} = 1$ almost surely.

Lemma 2 is important in showing that the prior is well defined. The proof is in Appendix B.

Our proposed prior setting is parsimonious but highly flexible in the sense that the induced prior assigns positive probability in arbitrarily small neighborhoods of any true data-generating

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

pmf. Let Π denote the space having elements of the form $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t \in \Pi_{d_1 \cdots d_p},\ t \in \{1, \ldots, T\}\}$. We show in Theorem 1 that the proposed prior has large support on Π.

Theorem 1. Let $\mathcal{Q}$ denote the prior on Π through the proposed model and $\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)$ denote an L_1 neighborhood around an arbitrary $\boldsymbol{\pi}^0 \in \Pi$. Then for any $\boldsymbol{\pi}^0 \in \Pi$ and $\epsilon > 0$, the prior assigns positive probability in the ϵ-neighborhood, $\mathcal{Q}\{\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)\} > 0$.

Since the proposed prior is defined on a space with finitely many components, a straightforward extension of Theorem 4.3.1 in Ghosh and Ramamoorthi (2003) ensures that the posterior concentrates in arbitrary small neighborhoods of any true data-generating distribution as the sample size increases.

2.3 Interpretability, Prior Elicitation and Structural Zeros

This section discusses how to interpret interactions, induce prior information, and accommodate structural zero conditions, relying on different expressions of cell probabilities.

In categorical data analysis, detecting interactions of various order is one of the main interests. To interpret our model, we propose a novel approach where cell probabilities are expressed relying on generalized linear modeling. Let $x_t^A, x_t^B, x_t^C, x_t^D$ be categorical variables with $x_t^* \in \{1, \ldots, d_*\}$ where $*$ is one of $\{A, B, C, D\}$. We express the cell probability as

$$P\left(x_t^A = a, x_t^B = b, x_t^C = c, x_t^D = d\right) = \pi_{tabcd} = \frac{\mu_{tabcd}}{\sum_{\hat{a}=1}^{d_A} \sum_{\hat{b}=1}^{d_B} \sum_{\hat{c}=1}^{d_C} \sum_{\hat{d}=1}^{d_D} \mu_{t\hat{a}\hat{b}\hat{c}\hat{d}}}, \tag{9}$$

where μ_{tabcd} is defined through its logarithm form as

$$\log \mu_{tabcd} = \lambda_t + \lambda_{ta}^A + \lambda_{tb}^B + \lambda_{tc}^C + \lambda_{td}^D \tag{10}$$

$$+ \lambda_{tab}^{AB} + \lambda_{tac}^{AC} + \lambda_{tad}^{AD} + \lambda_{tbc}^{BC} + \lambda_{tbd}^{BD} + \lambda_{tcd}^{CD} \tag{11}$$

$$+ \lambda_{tabc}^{ABC} + \lambda_{tabd}^{ABD} + \lambda_{tacd}^{ACD} + \lambda_{tbcd}^{BCD} + \lambda_{tabcd}^{ABCD}. \tag{12}$$

This model corresponds to the multinomial model for contingency tables (Agresti 2002). The right term in (10) represents constant term and main effects, (11) shows two-way interactions, and (12) corresponds to three-way and four-way interactions. For identification, we assume $\mu_{td_Ad_Bd_Cd_D} = 1$ and main effects and interactions are 0 if at least one of $a = d_A$, $b = d_B$, $c = d_C$ and $d = d_D$ is satisfied. The constraints imply $\lambda_t = 0$. Given the cell probabilities, the interactions can be deterministically computed. This approach can be applied to a general case since the proposed prior has Kolmogorov consistency properties so that if one starts with a prior for P variables the same marginal is obtained for any subset of variables as if one started with a prior for the joint pmf of the subset.

In addition, we propose an approach for inducing an informative prior on time varying main effects and interactions. We start with eliciting a prior sample size n as well as an informative prior $\pi^*(\theta^*)$ for the parameters θ^* in a parametric model, such as (9)–(12). To update our initial default prior to include this information, we follow a data augmentation approach in which we add data y^n generated from the prior predictive under the parametric model to our observed data, with y^n structured to have the same variables and sample size n at each time. That is, $y^n = \{\boldsymbol{y}_{ti}, t = 1, \ldots, T, i = 1, \ldots, n\}$ where $\boldsymbol{y}_{ti}$ is a vector of categorical variables with the same structure as $\mathbf{x}_{ti}$ in Section 2.2. To marginalize over the prior predictive in inducing an informative prior for the parameters in our nonparametric model, we generate a new value of y^n at each MCMC iteration.

In many contingency table applications, certain combinations of categories are known to have probability 0 a priori. For example, males cannot get pregnant. There are two ways in which structural zeros can be easily accommodated by the proposed method. Manrique-Vallier and Reiter (2012) recently proposed a Bayesian approach for latent structure models with structural zeros. Under their approach, the observed cell counts equal latent cell counts multiplied by an indicator function, which gives value 0 for structural zeros. Our model could be used for the latent cell counts, with the missing data in the cells with structural zeros imputed within an MCMC sampler.

Another approach is to combine variables that include combinations of structural zeros. For example, consider a survey that contains indicator variables of gender and pregnancy where $x_{ti1} = 1$ if subject i at time t is female and $x_{ti2} = 1$ if pregnant, leading to a 2×2 subtable with a structural zero. Then, we define a new variable $\tilde{x}_{ti}$ by vectorizing all cell components in the table except those that are impossible to occur. In this example, $\tilde{x}_{ti}$ has three categories such that $\tilde{x}_{ti} = 1$ corresponds to $(x_{ti1}, x_{ti2}) = (0, 0)$, $\tilde{x}_{ti} = 2$ to $(x_{ti1}, x_{ti2}) = (1, 0)$, and $\tilde{x}_{ti} = 3$ to $(x_{ti1}, x_{ti2}) = (1, 1)$. Replacing x_{ti1} and x_{ti2} by $\tilde{x}_{ti}$, the corresponding contingency tables have no structural zeros, and the proposed method can be applied. The cell probabilities of x_{ti1} and x_{ti2} can be computed from those of $\tilde{x}_{ti}$, so that inferences are based on relationships among the original variables.

3. MCMC ALGORITHM FOR POSTERIOR COMPUTATION

For posterior computation in DP mixtures, one common approach is marginalizing out the random probability measure with the Polya urn scheme (Bush and MacEachern 1996). Avoiding marginalization, Ishwaran and James (2001) developed the blocked Gibbs sampler relying on truncation approximation of the stick-breaking representation. Without truncation, Walker (2007) and Papaspiliopoulos and Roberts (2008) proposed the slice sampler and retrospective MCMC methods respectively. Though the slice sampler is simpler to implement, conditional constraints on sticks can cause slow mixing of the chain. Kalli, Griffin, and Walker (2011) proposed a more efficient slice sampler avoiding such a mixing problem.

Relying on a slice sampler related to Kalli, Griffin, and Walker (2011), we developed a simple and efficient MCMC algorithm for the proposed model. In the motivating application, we have two types of missing data, design-based missingness and individual-specific missingness. We assume missing at random for both cases and handle the missing data using missingness indicators, $m_{ti} = (m_{ti1}, \ldots, m_{tip})'$, with $m_{tij} = 1$ if variable j is missing for subject i at time t. In addition, we introduce latent variables $u_t = (u_{t1}, \ldots, u_{tn_t})'$ for the slice sampler. The likelihood of $\{u_t\}$ and $\{\mathbf{x}_t\}$ given $\{m_{ti}\}$, $\{\boldsymbol{v}_t\}$ and $\{\boldsymbol{\psi}_h^{(j)}\}$ can be expressed as

$$\prod_{t=1}^{T} \prod_{i=1}^{n_t} \left\{ \sum_{h=1}^{\infty} 1(u_{ti} < v_{th}) \prod_{j: m_{tij}=0} \prod_{l=1}^{d_j} \left(\psi_{hl}^{(j)}\right)^{1(x_{tij}=l)} \right\}.$$

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

This representation is consistent with the original model setting if latent variables $\{u_t\}$ are marginalized out. In a special case in which g is a probit link function, the data augmentation approach in Albert and Chib (2001) can improve efficiency of the posterior sampling by introducing independent normal latent variables $\{z_{tih}\}$ with mean W_{th} and variance 1 satisfying

$$P(z_{tih} > 0,\ z_{til} \le 0,\ l < h) = \Phi(W_{th}) \prod_{l<h} \{1 - \Phi(W_{th})\} = v_{th} = P(s_{ti} = h).$$

We propose the following MCMC sampling steps:

1. For $h = 1, \ldots, k^*$, with $k^* = \max\{s_{ti}\}$, update $\boldsymbol{\psi}_h^{(j)}$ from the following Dirichlet full conditional posterior distribution,
$$\text{Dirichlet}\Big(a_{j1} + \sum_{(t,i)\in A_{jh}} 1(x_{tij} = 1),\ \ldots,\ a_{jd_j} + \sum_{(t,i)\in A_{jh}} 1(x_{tij} = d_j)\Big),$$
where $A_{jh} = \{(t, i) : m_{tij} = 0,\ s_{ti} = h\}$.
2. Update z_{tih} from the marginal (w.r.t. u_{ti}) conditional posterior distribution,
$$z_{tih} \mid \cdots \sim \begin{cases} N_-(W_{th}, 1) & h < s_{ti}, \\ N_+(W_{th}, 1) & h = s_{ti}, \end{cases}$$
where $N_-(W_{th}, 1)$ and $N_+(W_{th}, 1)$ denote the normal distributions with mean W_{th} and variance 1 truncated on $(-\infty, 0]$ and $(0, \infty)$, respectively.
3. Update W_{th} from the normal marginal (w.r.t. u_{ti}) conditional posterior distribution, $N(\hat{W}_{th}, \sigma^2_{W_{th}})$, where
$$\hat{W}_{th} = \sigma^2_{W_{th}} \left(\sum_{i:s_{ti}\ge h}^{n_t} z_{tih} + \sigma_\varepsilon^{-2} \alpha_{th} \right),$$
$$\sigma^2_{W_{th}} = \frac{1}{\sum_{i=1}^{n_t} 1(s_{ti} \ge h) + \sigma_\varepsilon^{-2}}.$$
4. Update u_{ti} from the full conditional distribution, Uniform$(0, v_{ts_{ti}})$.
5. Update s_{ti} from the multinomial full conditional distribution,
$$\Pr(s_{ti} = h \mid \cdots) = \frac{1(h \in B_{ti}) \prod_{j:m_{tij}=0} \psi^{(j)}_{hx_{tij}}}{\sum_{l\in B_{ti}} \prod_{j:m_{tij}=0} \psi^{(j)}_{lx_{tij}}},$$
where $B_{ti} = \{h :\ v_{th} > u_{ti}\}$. To identify the elements in $\{B_{ti}\}$, we first update α_{th} and W_{th} for $t = 1, \ldots, T$ and $h = 1, \ldots, \tilde{k}$ where $\tilde{k}$ is the smallest number with $\sum_{h=1}^{\tilde{k}} v_{th} > 1 - \min\{s_{ti}\}$ for all t.
6. For $h = 1, \ldots, k^*$, update α_{th} using the forward filtering backward sampling algorithm by Früwirth-Schnatter (1994) and Carter and Kohn (1994), or Kalman filter and the simulation smoother by de Jong and Shephard (1995) and Durbin and Koopman (2002).
7. Update μ from the conditional posterior, $N(\mu_*, \sigma^2_\mu)$, where $\mu_* = \sigma^2_\mu(\hat{\sigma}^{-2}\hat{\mu} + \sigma_0^{-2}\mu_0)$, $\sigma^2_\mu = (\hat{\sigma}^{-2} + \sigma_0^{-2})^{-1}$ and
$$\hat{\mu} = \frac{\sum_{h=1}^{k^*} \sum_{t=2}^{T} (\alpha_{th} - \phi\alpha_{t-1h}) + (1+\phi) \sum_{h=1}^{k^*} \alpha_{1h}}{k^* \{T - 1 + (1+\phi)/(1-\phi)\}},$$
$$\hat{\sigma}^2 = \frac{\sigma^2_\eta}{k^* \{T - 1 + (1+\phi)/(1-\phi)\}}.$$
8. Update ϕ using the independence MH algorithm in which the proposal distribution is constructed relying on the mode and Hessian of the logarithm of the conditional posterior densities $\pi(\phi|\cdots)$. First, we compute $\hat{\phi}$ which maximizes (or approximately maximizes) the conditional posterior density. Then, we generated a candidate from a truncated normal distribution $TN_{(-1,1)}(\phi_*, \sigma^2_\phi)$, where
$$\phi_* = \hat{\phi} + \sigma^2_\phi \left. \frac{\partial \log \pi(\phi|\cdots)}{\partial \phi} \right|_{\phi=\hat{\phi}},$$
$$\sigma^2_\phi = \left\{ - \left. \frac{\partial \log \pi(\phi|\cdots)}{\partial^2 \phi} \right|_{\phi=\hat{\phi}} \right\}^{-1}.$$
9. Update σ^2_ε from the conditional distribution IG$(\hat{m}_\varepsilon/2, \hat{S}_\varepsilon/2)$, where $\hat{m}_\varepsilon = Tk^* + m_\varepsilon$ and $\hat{S}_\varepsilon = \sum_{t=1}^{T} \sum_{h=1}^{k^*} (W_{th} - \alpha_{th})^2 + S_\varepsilon$.
10. Update σ^2_η from the conditional distribution IG$(\hat{m}_\eta/2, \hat{S}_\eta/2)$, where $\hat{m}_\eta = Tk^* + m_\eta$ and $\hat{S}_\eta = \sum_{h=1}^{k^*} \sum_{t=2}^{T} (\alpha_{th} - \mu - \phi\alpha_{t-1h})^2 + (1-\phi^2) \sum_{h=1}^{k^*} \{\alpha_{1h} - \mu/(1-\phi)\}^2 + S_\eta$.

If g is an arbitrary link function, we update W_{th} using the independent MH algorithm, instead of steps 2 and 3 above. We generate a candidate from a normal distribution relying on the mode and Hessian of the logarithm of the conditional posterior densities of W_{th}. For logistic and t-links we can instead change the variance of z_{tih} from one to an observation-specific random variable having an appropriate mixture distribution; by treating the mixture distribution as unknown using a Dirichlet process, one can estimate the link nonparametrically.

4. SIMULATION STUDY

In this section, we assess the impact of borrowing of information over time by comparing our proposed method to static approaches, such as Dunson and Xing (DX) (2009). The static models are applied independently at each time point with no time-dependence included. First, we simulate time-indexed contingency tables from the model shown in expressions (4)–(8) with $T = 10$, $P = 20$, $d_j = 4$ for all j, $\mu = 0$, $\phi = 0.8$, $\sigma_\varepsilon = 0.1$, and $\sigma_\eta = 0.8$. At the respective time points we generated 120, 110, 150, 80, 100, 120, 100, 140, 110, and 150 observations, tiny sample sizes compared with the number of cells. For prior distributions, we assumed $\boldsymbol{\psi}_h^{(j)} \sim$ Dirichlet$(1, \ldots, 1)$, $\mu \sim N(0, 1)$, $\phi \sim U(-1, 1)$, $\sigma^2_\varepsilon \sim$ IG$(2.5, 0.025)$, $\sigma^2_\eta \sim$ IG$(2.5, 0.025)$. We draw 60,000 MCMC samples after the initial 20,000 samples are discarded as a burn-in period and every fifth sample is saved. We observed that the sample paths were stable and the sample autocorrelations dropped smoothly. Therefore, the chains apparently converged and mixed rapidly.

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

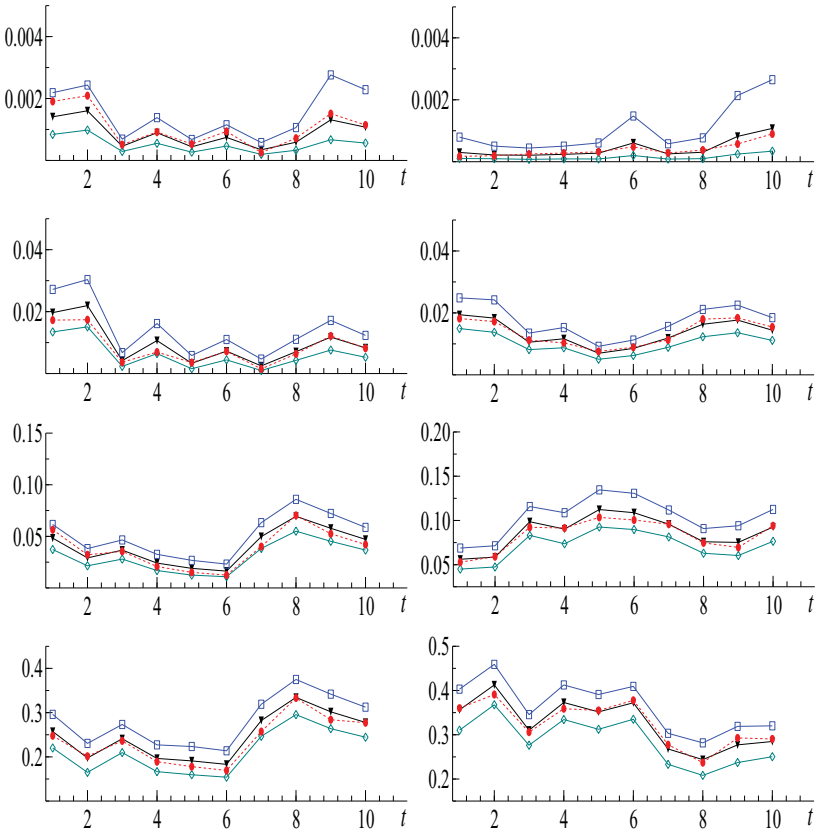

Figure 1. Estimation results of cell probabilities by the proposed method. Red lines with circle symbols represent true values, black lines with triangles posterior means, blue lines with squares upper bounds of 95% intervals and blue-green lines with diamonds lower bounds of 95% intervals. The first row: $P(x_{ti4} = 1, x_{ti6} = 2, x_{ti10} = 3, x_{ti15} = 4)$ and $P(x_{ti7} = 3, x_{ti9} = 1, x_{ti13} = 4, x_{ti19} = 2)$. The second row: $P(x_{ti1} = 3, x_{ti7} = 2, x_{ti20} = 4)$ and $P(x_{ti3} = 4, x_{ti12} = 2, x_{ti18} = 1)$. The third row: $P(x_{ti11} = 2, x_{ti17} = 2)$ and $P(x_{ti5} = 3, x_{ti19} = 2)$. The fourth row: $P(x_{ti8} = 1)$ and $P(x_{ti20} = 4)$. The online version of this figure is in color.

We first assess performance in estimation of cell probabilities. We randomly picked several cell probabilities and tracked their movements over time. We report true values, posterior means, and 95% credible intervals in Figure 1 (the proposed method) and Figure 2 (DX method). The proposed approach covers all true values in 95% intervals and interval widths are much narrower than for the DX approach consistently across time.

We additionally investigate performance in estimating associations among the categorical variables using the following measure of dependence from Dunson and Xing (2009)

$$\rho_{tjj'}^2 = \frac{1}{\min\{d_j, d_{j'}\} - 1} \sum_{c_j=1}^{d_j} \sum_{c_{j'}=1}^{d_{j'}} \frac{\left(\pi_{tc_jc_{j'}} - \bar{\psi}_{tc_j}^{(j)} \bar{\psi}_{tc_{j'}}^{(j')}\right)^2}{\bar{\psi}_{tc_j}^{(j)} \bar{\psi}_{tc_{j'}}^{(j')}}, \tag{13}$$

where $\bar{\psi}_{tl}^{(j)} \equiv P(x_{tij} = l) \approx \sum_{h=1}^{k^*} \nu_{th} \psi_{hl}^{(j)}$. The first row of Figure 3 reports plots of all pairs of true values (y-axis) and posterior means (x-axis) of $\rho_{tjj'}$ at time $t = 2$ and 7. Since all cell probabilities are given in the simulation study, the true $\rho_{tjj'}$ can be computed through (13). At each time point, coordinate points by our approach locate closely to the $y = x$ line, compared to widely scattered points by the DX method. In addition, Table 1 shows correlations between true values and posterior means of $\rho_{tjj'}$. Although correlations by the DX method are high, the proposed method consistently produces higher correlations.

Log-linear models provide a standard choice for the analysis of contingency tables. However, one issue is that flexible log-linear models that accommodate arbitrary interactions among the variables and allow time dependence cannot be applied directly to large, sparse tables. Certainly, maximum likelihood estimates typically do not exist and Bayesian methods, which allow an unknown dependence structure, do not scale beyond small tables. Dahinden, Kalisch, and Buehlmann (2010) proposed an approach for high-dimensional log-linear models with interactions, which relies on solving several low-dimensional subproblems that are then combined. An earlier approach by Dahinden et al. (2007) instead relied on L1 penalized log-linear

Table 1. Correlations between true values and posterior means of $\rho_{tjj'}$ using the first simulation data

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$	$t = 10$	Total
Proposed	0.948	0.977	0.990	0.977	0.983	0.986	0.985	0.965	0.969	0.968	0.974
DX	0.837	0.794	0.880	0.761	0.766	0.921	0.846	0.817	0.831	0.793	0.841

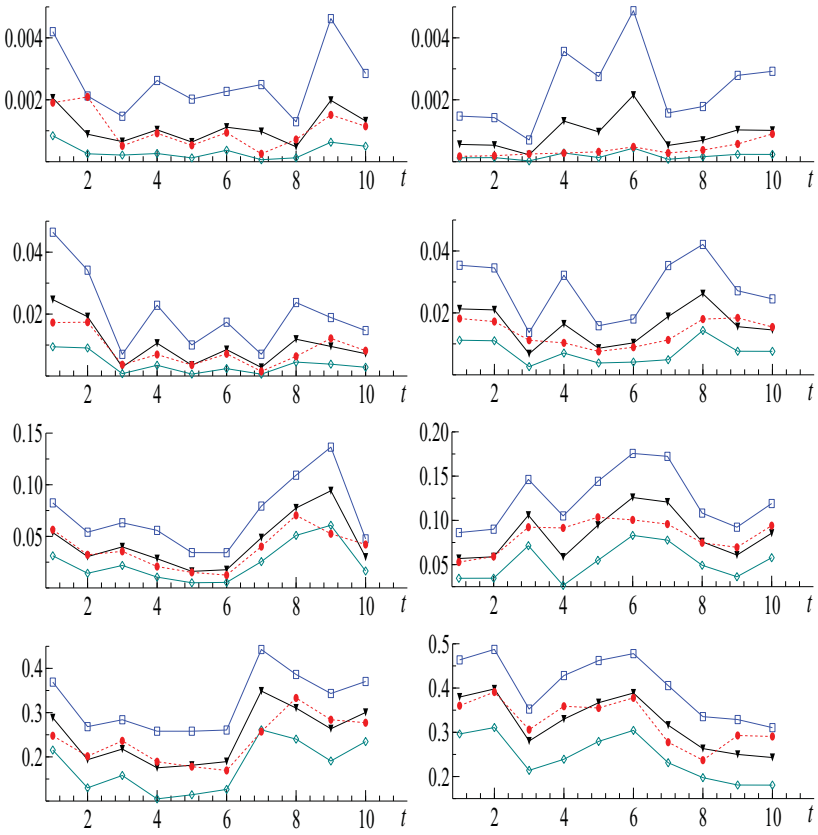

Figure 2. Estimation results of cell probabilities by DX method. Red lines with circle symbols represent true values, black lines with triangles posterior means, blue lines with squares upper bounds of 95% intervals and bule-green lines with diamonds lower bounds of 95% intervals. The first row: $P(x_{ti4} = 1, x_{ti6} = 2, x_{ti10} = 3, x_{ti15} = 4)$ and $P(x_{ti7} = 3, x_{ti9} = 1, x_{ti13} = 4, x_{ti19} = 2)$. The second row: $P(x_{ti1} = 3, x_{ti7} = 2, x_{ti20} = 4)$ and $P(x_{ti3} = 4, x_{ti12} = 2, x_{ti18} = 1)$. The third row: $P(x_{ti11} = 2, x_{ti17} = 2)$ and $P(x_{ti5} = 3, x_{ti19} = 2)$. The fourth row: $P(x_{ti8} = 1)$ and $P(x_{ti20} = 4)$. The online version of this figure is in color.

models allowing sparsity of tables. Also, Dahinden et al. (2007) proposed an efficient estimation algorithm for model selection for two level categorical variables.

As a second alternative to our proposed approach, we implemented the method of Dahinden et al. (DH) (2007) in a second simulation example with $T = 8$, $P = 13$, and $d_j = 2$ for all j. Other settings are the same as in the first simulation case. As DH did not consider time-indexed contingency tables, we applied their approach separately at each time point using the logilasso R package, with five-way cross-validation used to choose penalty parameters. The second rows of Figure 3 and Table 2, respectively, summarize the resulting dependence measures $\rho_{tjj'}$ at time $t = 2$ and 7 for each method. For the proposed method, the posterior means are close to true values and correlations between estimates and true values are uniformly high. The DH method has a tendency to underestimate dependence, particularly when true values are low, and has the lowest correlation between the estimates and truth.

To gauge robustness we also simulated data from a time-dependent log-linear model in which all main effects and two-way interactions independently follow random walk processes, $\xi_t \sim N(\xi_{t-1}, 1)$ with $\xi_0 = 0$ where ξ_t is a main effect or two-way interaction at time t, and other higher interactions are zero. The third rows of Figure 3 and Table 3, respectively, report the estimation results. Although we find less difference among them in this case, the proposed method still shows the best performance.

In addition, for the interactions discussed in Section 2.3, we generated contingency tables from the model (9)–(12) with $N = 1,200$, $T = 30$, $P = 4$, and $d_j = 2$ for all j. For the first 10 time points, we assumed two-way interactions (x_t^A, x_t^B), (x_t^A, x_t^C), (x_t^B, x_t^C), (x_t^B, x_t^D), (x_t^C, x_t^D), where we express the structure of these interactions as

$$M_1 = \{AB, AC, BC, BD, CD\}.$$

Table 2. Correlations between true values and posterior means of $\rho_{tjj'}$ using the second simulation data

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	Total
Proposed	0.951	0.978	0.979	0.984	0.986	0.969	0.981	0.944	0.965
DX	0.872	0.803	0.838	0.599	0.807	0.884	0.932	0.827	0.696
DH	0.705	0.557	0.733	0.466	0.725	0.506	0.763	0.487	0.562

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

Table 3. Correlations between true values and posterior means of $\rho_{tjj'}$ using the third simulation data

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	Total
Proposed	0.725	0.827	0.768	0.798	0.818	0.916	0.791	0.807	0.817
DX	0.642	0.640	0.726	0.664	0.611	0.864	0.769	0.713	0.724
DH	0.371	0.716	0.821	0.491	0.611	0.877	0.764	0.715	0.624

For the next 10 time points, a three-way interaction involving variables x_t^A, x_t^B, and x_t^C also occurs:

$$M_2 = M_1 \cap \{ABC\}.$$

For the last 10 time points, another three-way interaction involving variables x_t^B, x_t^C, and x_t^D occurs:

$$M_3 = M_2 \cap \{BCD\}.$$

We assumed no four-way interaction for simplicity. Using the same choice of prior as in our other analyses, we generated 20,000 MCMC samples after a 10,000 burn-in and every fifth sample was saved. Our prior induces a flexible shrinkage prior on the interactions, centered at zero with heavier than Gaussian tails and concentration increasing with interaction order. Detailed descriptions of the induced priors are provided in the supplementary materials. Figure 4 reports the estimation results of interactions. The proposed method clearly has good performance in estimating the interactions and detecting the time changes, with 95% credible intervals covering the true parameters 97.5% of the time for the constant interactions and 96.6% for the time-varying interactions.

5. ANALYSIS OF SOCIAL SURVEY DATA

In this section, we apply the proposed method to data from the General Social Survey (GSS, *http://www3.norc.org/GSS+Website*). Our focus is on studying associations

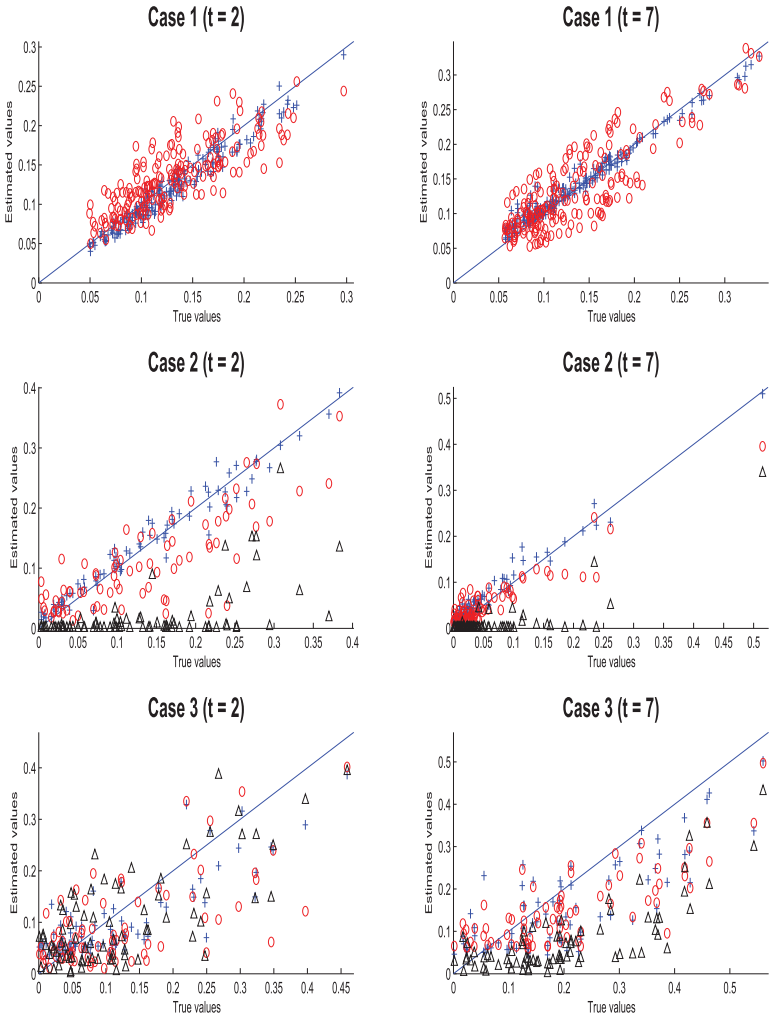

Figure 3. Plots of true and estimated values of $\rho_{tjj'}$ using the simulation data. y axis represents estimated values and x axis true values. Cross-shaped dots represent the proposed method, circles DX method, and triangles DH method. The first, second, and third rows show the results at time $t = 2$ and 7 using the first (case 1), second (case 2), third (case 3) simulation datasets. The online version of this figure is in color.

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

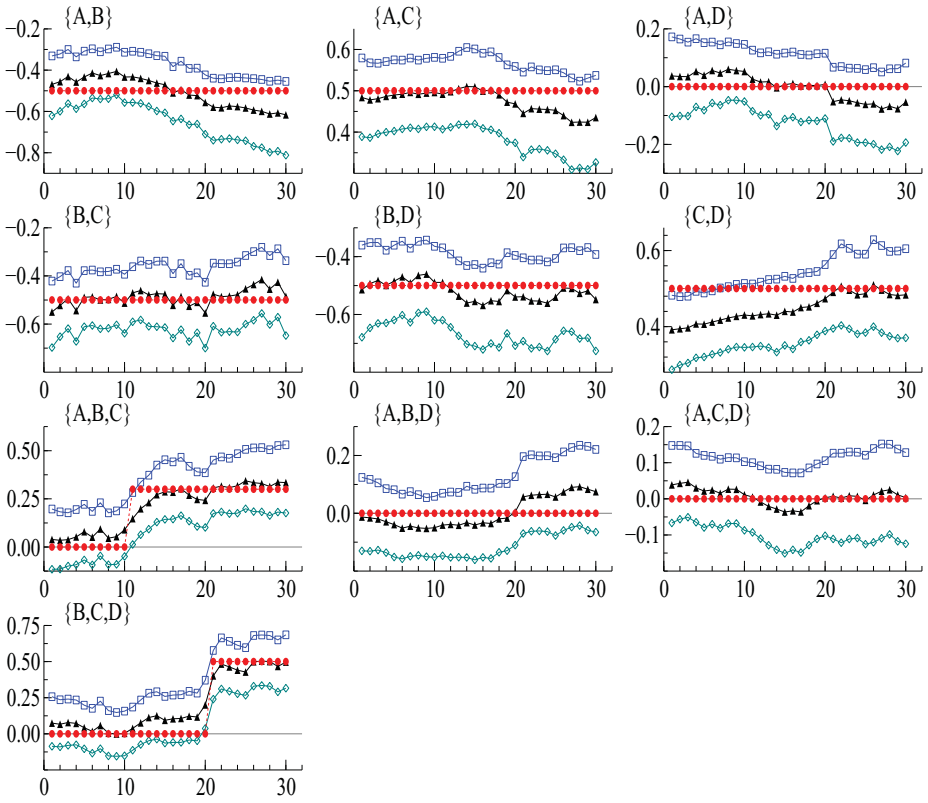

Figure 4. Estimation results of interactions in the parametric modeling. y axis represents estimated values and x axis time. Red lines with circle symbols represent true values, black lines with triangles posterior means, blue lines with squares upper bounds of 95% intervals, and blue-green lines with diamonds lower bounds of 95% intervals. The online version of this figure is in color.

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

among demographic and preference variables over time. We select $p = 29$ categorical variables from 1994 to 2010, including gender, ethnicity, preference for particular policies, and many more listed in the supplementary materials. The GSS was conducted every 2 years across this time period. The numbers of observations are 2,992 (1994), 2,904 (1996), 2,832 (1998), 2,817 (2000), 2,765 (2002), 2,812 (2004), 4,510 (2006), 2,023 (2008), and 2,044 (2010), respectively. There are abundant missing data in which only a subset of the variables were recorded for an individual, and compared to the number of cells, the sample size is quite small at each time point.

We first compared our proposed approach to log-linear models. Unfortunately, current methodology for fitting log-linear models that allow flexible dependence structures cannot accommodate these data due to the large sparse structure, time variation, and abundant missing data. Hence, to provide a comparison, we initially focused on a bivariate subset of the data consisting of religious preference ($i = 1, \ldots, 5$) and attitude toward abortion ($j = 1, 2$) from 1994 to 2010. We consider the following Log-Linear Poisson (LLP) models.

$$\text{LLP-Model 1:} \qquad \begin{aligned} N_{tij} &\sim \text{Poisson}(N_t\, \mu_{ij}), \\ \log \mu_{ij} &= \lambda + \lambda_i^R + \lambda_j^A + \lambda_{ij}^{RA}, \end{aligned}$$

where N_{tij} is count of the cell ij at time t, $N_t = \sum_i \sum_j N_{tij}$, λ_i^R is an effect of the first variable (religious preference), λ_j^A is an effect of the second variable (view of abortion), and λ_{ij}^{RA} is an association term. For identifiability, we assume constraints $\lambda_5^R = \lambda_2^A = \lambda_{5j}^{RA} = \lambda_{i2}^{RA} = 0$. LLP-Model 1 is a standard choice in the contingency table literature (Agresti 2002), and the cell probabilities $\mu_{ij} / \sum_{i'} \sum_{j'} \mu_{i'j'}$ do not depend on time. Next, we extended LLP-Model 1 to incorporate time-varying effects on cell probabilities.

LLP-Model 2:

$$\begin{aligned} &N_{tij} \sim \text{Poisson}(N_t\, \mu_{tij}), \qquad \log \mu_{tij} = \lambda_t + \lambda_{ti}^R + \lambda_{tj}^A + \lambda_{tij}^{RA}, \\ &\boldsymbol{\beta}_t = (\lambda_t, \lambda_{t1}^R, \ldots, \lambda_{t4}^R, \lambda_{t1}^A, \lambda_{t11}^{RA}, \ldots, \lambda_{t41}^{RA})', \\ &\beta_{tl} = \mu_l + \phi_l \beta_{t-1l} + \varepsilon_{tl}, \quad \varepsilon_{tl} \sim N(0, \sigma_l^2), \quad l = 1, \ldots, 10, \end{aligned}$$

where λ_{ti}^R, λ_{tj}^A, and λ_{tij}^{RA} are effects of the first variable, the second variable, and interactions at time t, respectively. We assume $\lambda_{t5}^R = \lambda_{t2}^A = \lambda_{t5j}^{RA} = \lambda_{ti2}^{RA} = 0$ at each time point and $\boldsymbol{\beta}_0 = \mathbf{0}$ for the initial values. LLP-Model 2 is a time-dependent hierarchical model in which all parameters in the log-linear model follow first order autoregressive processes independently.

We first estimate all models using the data from 1994 to 2008. Then, relying on the estimated parameters, we predict the contingency table in 2010 (Table 4). For the proposed model, we used the same MCMC settings as in the simulation study. For log-linear models, we estimated parameters using an MCMC algorithm where missing values are imputed from conditional probabilities given observed data at each iteration. For example, we generate the religious preference i given the view of abortion j with probability $\mu_{tij} / \sum_{i'} \mu_{ti'j}$. For priors, we assumed $\boldsymbol{\beta} = (\lambda, \lambda_1^R, \ldots, \lambda_4^R, \lambda_1^A, \lambda_{11}^{RA}, \ldots, \lambda_{41}^{RA})' \sim N(\mathbf{0}, I)$ for LLP-Model 1, $\mu_l \sim N(0, 1)$, $\phi_l \sim U(-1, 1)$, and $\sigma_l^2 \sim \text{IG}(2.5, 0.025)$ for all

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

Table 4. Contingency table of the religious preference and view of abortion in 2010

	Protestant	Catholic	Jewish	None	Other	Total
Agree	216	103	21	137	60	537
Disagree	372	182	7	81	47	689
Total	588	285	28	218	107	1226

l for LLP-Model 2. Using Gibbs' sampling, we generated posterior samples of μ_l and σ_l^2 from normal and Inverse-Gamma distributions, respectively. For $\boldsymbol{\beta}$, ϕ_l, $\boldsymbol{\beta}_t$, we used an MH algorithm in which candidates were generated from normal distributions relying on the mode and Hessian of the logarithm of the conditional posterior densities. We generated 10,000 MCMC samples after a 1000 burn-in for LLP-Model 1 and 20,000 MCMC samples after the 2000 burn-in for LLP-Model 2 and, for both cases, every fifth sample was saved.

We generated replications at every fifth MCMC iteration and computed averages of the following predictive criteria:

Absolute deviation (AD):

$$\sum_{i=1}^{5}\sum_{j=1}^{2}\left|N_{ij}^{\text{rep}}-N_{ij}^{\text{obs}}\right|,$$

Mean absolute percentage error (MAPE):

$$\frac{1}{10}\sum_{i=1}^{5}\sum_{j=1}^{2}\left|\frac{N_{ij}^{\text{rep}}-N_{ij}^{\text{obs}}}{N_{ij}^{\text{obs}}}\right|,$$

where N_{ij}^{rep} and N_{ij}^{obs} are the replication and observation of count of the cell ij, respectively. To keep the same total number of replications among all methods, predictions are generated from cell probabilities $\mu_{ij}/\sum_{i'}\sum_{j'}\mu_{i'j'}$ for LLP-Model 1 and $\mu_{2010ij}/\sum_{i'}\sum_{j'}\mu_{2010i'j'}$ for LLP-Model 2. Table 5 reports the prediction results. Although LLP-Model 2 produces better performance than LLP-Model 1 by incorporating time-dependence, our proposed method clearly outperforms log-linear models in terms of both predictive criteria. In addition, we compared 95% predictive intervals by the proposed method and LLP-Model 2 and found that our method could capture all actual observations, while LLP-Model 2 had poor coverage. The plot is included in the supplementary materials.

Next, we apply the proposed method to all 29 categorical variables. We generated 30,000 MCMC samples after the initial 10,000 samples are discarded as the burn-in and every fifth sample are saved. We observed the sample paths are stable and the sample autocorrelations are small. Table 6 shows the estimation result of parameters in the time-dependent stick-breaking processes. Concerning the measure of time-dependence ϕ, the posterior mean is close to 1 and the 95% credible interval locates

Table 5. Prediction results

	Proposed	Model 1	Model 2
AD	194.4	208.6	204.5
MAPE	0.216	0.232	0.227

Table 6. Estimation result of parameters in the proposed stick-breaking process

Parameter	Mean	Stdev.	95% interval
μ	−0.012	0.004	[−0.023, −0.005]
ϕ	0.988	0.004	[0.978, 0.994]
σ_ε	0.062	0.009	[0.046, 0.082]
σ_η	0.126	0.011	[0.104, 0.149]

near 1, which means the weights of the stick-breaking processes have strong time dependence over time.

Then, we investigate cross-interactions among the variables over time. Figure 5 show the posterior means of $\rho_{tjj'}$ for all pairs in 2002 and 2010. Additional results for other years are included in the supplementary materials. We find the structure of interactions is complex at each time point. Also, though each interaction gradually changes over time, all tables look similar to one another, implying they have close dependence. This is consistent with the result of the strong dependent weights in the stick-breaking processes. Some categorical variables such as Race [$j = 3$], Attitude toward abortion [6], Political party affiliation [9], and Think of self as liberal or conservative [14] intricately correlate with many other variables. On the other hand, zodiac [11] shows little interactions with all other variables. Among all pairs of variables, {Age [1], Marital status [10]}, {Attitude toward abortion [6], Attitude toward homosexual [16]}, and {Attitude toward homosexual [16], Attitude toward Marijuana [19]} show strong interactions in the whole period. Also, we observed several pairs of variables showing relatively close interactions over time, such as {Attitude toward abortion [6], Think of self as liberal or conservative [14]}, {Race [3], Political party affiliation [9]}, and {Marital status [10], Having gun [17]}. In addition, the views of government expense show moderate interactions, especially to the environment [23], nation's health [24], halting the rising crime [25], dealing with drug addiction [26], and education system [27].

Next, we study trends of dependence between categorical variables. Figure 6 reports the posterior means and 95% credible intervals of $\rho_{tjj'}$ for pairs with close interactions. We observed various patterns of time paths. For {Age, Marital status}, the interaction increased around 2000 then declined sharply to a lower level. {Race, Political party affiliation} and {Race, Having gun} have peaks in 2006 and the interactions have steeply decreased after that. In addition, we can see similar trends in {Attitude toward abortion, Think of self as lib or con}, {Attitude toward abortion, Attitude toward homosexual}, {Attitude toward homosexual, Attitude toward Marijuana}, {Religion, Attitude toward abortion}, and {Religion, Attitude toward Marijuana}. The interactions have roughly increased over time, especially in the 2000s. On the other hand, the dependence in {Race, Death penalty for murder} decreased at first and kept stable in the middle of the period then declined again. {Having gun, Family income} gradually increased over the period but the difference is small. For {Marital status, Having gun}, the interaction dropped in the middle of the period but recovered recently at the same level as the beginning.

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

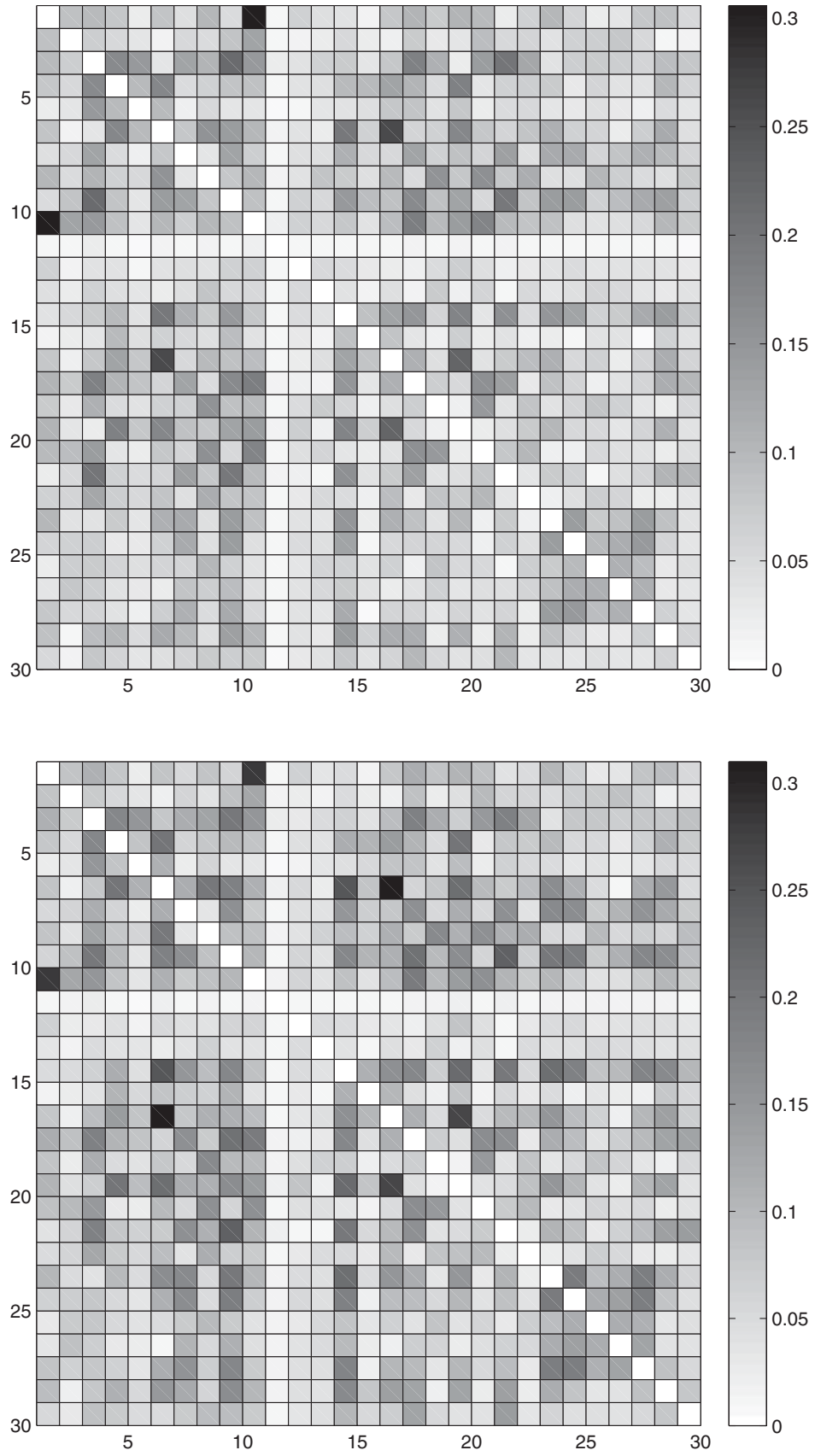

Figure 5. Posterior means of $\rho_{tjj'}$ in 2002 (above) and 2010 (below).

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

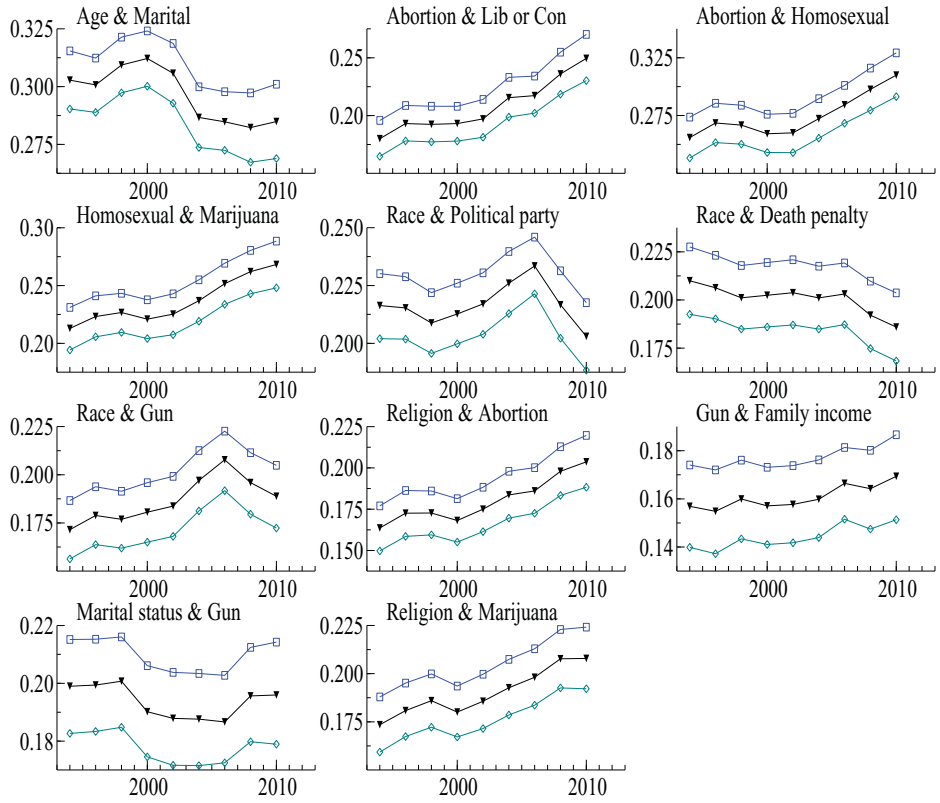

Figure 6. Estimation results of $\rho_{tjj'}$ for several pairs. Black lines with triangles posterior means, blue lines with squares upper bounds of 95% intervals and bule-green lines with diamonds lower bounds of 95% intervals. The first row: (Age group, Current marital status), (Attitude toward abortion, Think of self as liberal or conservative), and (Attitude toward abortion, Attitude toward homosexual sex relations). The second row: (Attitude toward homosexual sex relations, Should Marijuana be made legal), (Race, Political party affiliation), and (Race, Favor or oppose death penalty for murder). The third row: (Race, Have gun in home), (Religious preference, Attitude toward abortion), and (Have gun in home, Total family income). The fourth row: (Current marital status, Have gun in home) and (Religious preference, Should Marijuana be made legal). The online version of this figure is in color.

6. DISCUSSION

We have demonstrated that the proposed approach is useful in analyzing time-indexed large sparse contingency tables. One interesting extension is to accommodate joint modeling of mixed scale variables consisting of not only categorical data but also continuous and count variables. In such a case, one can potentially model the observed data vector for the ith subject at time t, $y_{ti} = (y_{ti1}, \ldots, y_{tip})'$, as conditionally independent given latent class variables $x_{ti} = (x_{ti1}, \ldots, x_{tip})'$, with x_{ti} modeled exactly as proposed in this article. For example, consider the simple case in which $p = 2$ with $y_{ti1} \in \Re$ continuous and $y_{ti2} \in \{1, \ldots, d_2\}$ categorical. Then, one can let $y_{ti1} \sim N(\mu_{x_{ti1}}, \sigma^2_{x_{ti1}})$ and $y_{ti2} = x_{ti2}$, with the proposed probabilistic tensor factorization approach flexibly accommodating dependence in y_{ti1} and y_{ti2} through dependence in x_{ti1} and x_{ti2}. The induced marginal distribution for the continuous variable y_{ti1} will be a mixture of normals, with the probability weight on each component potentially varying with the categorical variable y_{ti2}. This same strategy can be generalized to more complex settings involving many categorical, count, continuous, and even functional observations.

Another interesting direction in terms of generalizations is to accommodate dependence in the observations; for example, one may collect multivariate categorical longitudinal data in which the same variables are measured repeatedly on the sample study subjects or the data may have a nested structure. Log-linear and logistic regression-type models can be easily generalized to such settings, but clearly encounter computational challenges in large sparse settings. Potentially the simplex factor model of Bhattacharya and Dunson (2011) can be generalized to accommodate such dependence structures through the latent factors, with some challenges arising in terms of developing computationally efficient implementations and models that are both flexible and interpretable.

APPENDIX A: PROOF OF LEMMA 1

The expectation of cell probability is

$$E\{\pi_{tc_1\cdots c_p}\} = E\left\{\sum_{h=1}^{\infty} \nu_{th} \prod_{j=1}^{p} \psi^{(j)}_{hc_j}\right\} = \sum_{h=1}^{\infty}\left[E\{\nu_{th}\}\prod_{j=1}^{p} E\left\{\psi^{(j)}_{hc_j}\right\}\right],$$
$$= \prod_{j=1}^{p} E\left\{\psi^{(j)}_{hc_j}\right\}\sum_{h=1}^{\infty} E\{\nu_{th}\} = \prod_{j=1}^{p} E\left\{\psi^{(j)}_{hc_j}\right\} = \prod_{j=1}^{p}\frac{a_{jc_j}}{\hat{a}_j}.$$

The marginal distribution of W_{th} can be expressed as $N(\mu/(1-\phi), \sigma^2_\eta/(1-\phi^2)+\sigma^2_\varepsilon)$, independent of t and h. Hence, we set $\beta_1 = E\{g(W_{th})\}$ and $\beta_2 = E\{g^2(W_{th})\}$. The second moment of cell

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

probability is

$$\begin{aligned}
E\{\pi^2_{tc_1\cdots c_p}\} &= E\left[\left\{\sum_{h=1}^{\infty}\nu_{th}\prod_{j=1}^{p}\psi^{(j)}_{hc_j}\right\}\left\{\sum_{l=1}^{\infty}\nu_{tl}\prod_{j=1}^{p}\psi^{(j)}_{lc_j}\right\}\right],\\
&= \sum_{h=1}^{\infty}\sum_{l=1}^{\infty}E\{\nu_{th}\nu_{tl}\}E\left\{\prod_{j=1}^{p}\psi^{(j)}_{hc_j}\psi^{(j)}_{lc_j}\right\},\\
&= \left[\prod_{j=1}^{p}E\left\{\left(\psi^{(j)}_{hc_j}\right)^2\right\}-\prod_{j=1}^{p}E^2\left\{\psi^{(j)}_{hc_j}\right\}\right]\sum_{h=1}^{\infty}E\{\nu^2_{th}\}\\
&\quad+\prod_{j=1}^{p}E^2\left\{\psi^{(j)}_{hc_j}\right\}\sum_{h=1}^{\infty}\sum_{l=1}^{\infty}E\{\nu_{th}\nu_{tl}\},\\
&= \left(\prod_{j=1}^{p}\frac{a_{jc_j}(a_{jc_j}+1)}{\hat{a}_j(\hat{a}_j+1)}-\prod_{j=1}^{p}\frac{a^2_{jc_j}}{\hat{a}^2_j}\right)\sum_{h=1}^{\infty}E\{\nu^2_{th}\}+\prod_{j=1}^{p}\frac{a^2_{jc_j}}{\hat{a}^2_j},
\end{aligned}$$

where

$$\begin{aligned}
\sum_{h=1}^{\infty}E\{\nu^2_{th}\} &= \sum_{h=1}^{\infty}E\left[g^2(W_{th})\prod_{l<h}\{1-g(W_{tl})\}^2\right],\\
&= \sum_{h=1}^{\infty}\beta_2\{1-2\beta_1+\beta_2\}^{h-1},\\
&= \frac{\beta_2}{2\beta_1-\beta_2}.
\end{aligned}$$

Hence,

$$V\{\pi_{tc_1\cdots c_p}\} = \left(\prod_{j=1}^{p}\frac{a_{jc_j}(a_{jc_j}+1)}{\hat{a}_j(\hat{a}_j+1)}-\prod_{j=1}^{p}\frac{a^2_{jc_j}}{\hat{a}^2_j}\right)\left(\frac{\beta_2}{2\beta_1-\beta_2}\right). \quad \text{(A.1)}$$

Similarly,

$$\begin{aligned}
&E\{\pi_{tc_1\cdots c_p}\pi_{t+kc'_1\cdots c'_p}\}\\
&= E\left[\left\{\sum_{h=1}^{\infty}\nu_{th}\prod_{j=1}^{p}\psi^{(j)}_{hc_j}\right\}\left\{\sum_{l=1}^{\infty}\nu_{t+kl}\prod_{i=1}^{p}\psi^{(i)}_{lc'_i}\right\}\right],\\
&= \left[\prod_{j=1}^{p}E\left\{\psi^{(j)}_{hc_j}\psi^{(j)}_{hc'_j}\right\}-\prod_{j=1}^{p}E\left\{\psi^{(j)}_{hc_j}\right\}E\left\{\psi^{(j)}_{lc'_j}\right\}\right]\sum_{h=1}^{\infty}E\{\nu_{th}\nu_{t+kh}\}\\
&\quad+\prod_{j=1}^{p}E\left\{\psi^{(j)}_{hc_j}\right\}E\left\{\psi^{(j)}_{lc'_j}\right\},\\
&= \left(\prod_{j=1}^{p}\frac{a_{jc_j}\{a_{jc'_j}+1(c_j=c'_j)\}}{\hat{a}_j(\hat{a}_j+1)}-\prod_{j=1}^{p}\frac{a_{jc_j}a_{jc'_j}}{\hat{a}^2_j}\right)\sum_{h=1}^{\infty}E\{\nu_{th}\nu_{t+kh}\}\\
&\quad+\prod_{j=1}^{p}\frac{a_{jc_j}a_{jc'_j}}{\hat{a}^2_j},
\end{aligned}$$

where

$$\begin{aligned}
&E\{\nu_{th}\nu_{t+kh}\}\\
&= E\left\{\left[g(W_{th})\prod_{l<h}\{1-g(W_{tl})\}\right]\left[g(W_{t+kh})\prod_{l<h}\{1-g(W_{t+kl})\}\right]\right\},\\
&= E\{g(W_{th})g(W_{t+kh})\}\prod_{l<h}E[\{1-g(W_{tl})\}\{1-g(W_{t+kl})\}],\\
&= E\{g(W_{th})g(W_{t+kh})\}\prod_{l<h}[1-2\beta_1+E\{g(W_{tl})g(W_{t+kl})\}].
\end{aligned}$$

From (7) and (8), $E\{g(W_{th})g(W_{t+kh})\}$ can be expressed as

$$\begin{aligned}
&E\{g(W_{th})g(W_{t+kh})\}\\
&= E\{g(\alpha_{th}+\varepsilon_{th})g(\alpha_{t+kh}+\varepsilon_{t+kh})\},\\
&= E\left\{g(\alpha_{th}+\varepsilon_{th})\,g\left(\frac{1-\phi^k}{1-\phi}\mu+\phi^k\alpha_{th}+\sum_{i=0}^{k-1}\phi^i w_{t+k-ih}+\varepsilon_{t+kh}\right)\right\}.
\end{aligned}$$

Since α_{th}, ε_{th}, w_{t+k-ih} $(i=0,\ldots,k-1)$, and ε_{t+kh} are independent of one another and their distributions do not depend on t or h, hence $\gamma_k \equiv E\{g(W_{th})g(W_{t+kh})\}$ is dependent on time difference k but independent of time t.

In addition,

$$\begin{aligned}
\sum_{h=1}^{\infty}E\{\nu_{th}\nu_{t+kh}\} &= \sum_{h=1}^{\infty}\gamma_k\prod_{l<h}\{1-2\beta_1+\gamma_k\},\\
&= \frac{\gamma_k}{2\beta_1-\gamma_k}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\operatorname{cov}\{\pi_{tc_1\cdots c_p},\pi_{t+kc'_1\cdots c'_p}\}\\
&= \left(\prod_{j=1}^{p}\frac{a_{jc_j}\{a_{jc'_j}+1(c_j=c'_j)\}}{\hat{a}_j(\hat{a}_j+1)}-\prod_{j=1}^{p}\frac{a_{jc_j}a_{jc'_j}}{\hat{a}^2_j}\right)\left(\frac{\gamma_k}{2\beta_1-\gamma_k}\right).
\end{aligned}$$

Since $\beta_2/(2\beta_1-\beta_2)>0$, $\gamma_k/(2\beta_1-\gamma_k)>0$ and (A.1), cell probabilities with $c_j=c'_j$ for all j have positive covariance and, on the other hand, those with $c_j\neq c'_j$ for all j have negative covariance.

In a case where $a_{j1}=\cdots=a_{jc_j}=a$, the variance and covariance are expressed as

$$\begin{aligned}
&V\{\pi_{tc_1\cdots c_p}\}\\
&= \left(\prod_{j=1}^{p}\frac{1+1/a}{d^2_j+d_j/a}-\prod_{j=1}^{p}\frac{1}{d^2_j}\right)\left(\frac{\beta_2}{2\beta_1-\beta_2}\right),\\
&\operatorname{cov}\{\pi_{tc_1\cdots c_p},\pi_{t+kc'_1\cdots c'_p}\}\\
&= \left(\prod_{j=1}^{p}\frac{1+1(c_j=c'_j)/a}{d^2_j+d_j/a}-\prod_{j=1}^{p}\frac{1}{d^2_j}\right)\left(\frac{\gamma_k}{2\beta_1-\gamma_k}\right).
\end{aligned}$$

Hence, $V\{\pi_{tc_1\cdots c_p}\}\to 0$ and $\operatorname{cov}\{\pi_{tc_1\cdots c_p},\pi_{t+kc'_1\cdots c'_p}\}\to 0$ as $a\to\infty$.

APPENDIX B: PROOF OF LEMMA 2

To prove $\sum_{h=1}^{\infty}\nu_{th}=1$ a.s., it is enough to show $\sum_{h=1}^{\infty}E\{\log(1-g(W_{th})\}=-\infty$ (Ishwaran and James 2001). g is a nonnegative monotone increasing link function: $\Re\to(0,1)$, therefore $0<\beta_1=E\{g(W_{th})\}<1$. Then, using Jensen's inequality,

$$E[\log\{1-g(W_{th})\}]\le\log[1-E\{g(W_{th})\}]=\log(1-\beta_1)<0.$$

Therefore, $\sum_{h=1}^{\infty}E\{\log(1-g(W_{th})\}=-\infty$ at each time point.

APPENDIX C: PROOF OF THEOREM

The proposed prior probability assigned to $\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)$ can be expressed as

$$\mathcal{Q}\left\{\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)\right\} = \int 1(\|\boldsymbol{\pi}-\boldsymbol{\pi}^0\|<\epsilon)d\mathcal{Q}(\boldsymbol{\nu}_t,\boldsymbol{\psi}^{(j)}_h,t\in\{1,\ldots,T\},\\ h=1,\ldots,\infty,\,j=1,\ldots,p).$$

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017

where $\boldsymbol{\nu}_t$ is a probability vector induced by the proposed stick-breaking process and we use the L_1 distance

$$\|\boldsymbol{\pi} - \boldsymbol{\pi}^0\| = \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} |\pi_{tc_1\cdots c_p} - \pi^0_{tc_1\cdots c_p}|,$$

where p_t is a probability mass function for time $t \in \{1, \ldots, T\}$.

For any $\boldsymbol{\pi}^0 \in \Pi$, each component in $\boldsymbol{\pi}^0$ can be expressed as

$$\boldsymbol{\pi}^0_t = \sum_{h=1}^{k_t} \nu^0_{th} \Psi_{th}, \quad \Psi_{th} = \boldsymbol{\psi}^{(1)}_{th} \otimes \cdots \otimes \boldsymbol{\psi}^{(p)}_{th},$$

where $k_t \in \mathbb{N}$, $\boldsymbol{\nu}^0_t = (\nu^0_{t1}, \ldots, \nu^0_{tk_t})'$ is a probability vector, $\Psi_{th} \in \Pi_{d_1\cdots d_p}$ and $\boldsymbol{\psi}^{(j)}_{th} = (\psi^{(j)}_{th1}, \ldots, \psi^{(j)}_{thd_j})'$ is a $d_j \times 1$ probability vector. We define $k^+_0 = 0$ and $k^+_t = \sum_{i=1}^{t} k_i$ for $t = 1, \ldots, T$. Then, we construct $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t, t \in \{1, \ldots, T\}\} \in \Pi$ induced by the proposed prior such that the component with the index h in $\boldsymbol{\pi}^0_t$ is approximated by the component with the index $k^+_{t-1} + h$ in $\boldsymbol{\pi}_t$. For any ϵ, we define a set $D(\boldsymbol{\pi}^0, \epsilon) \subset \Pi$ such that for any $\boldsymbol{\pi} \in D(\boldsymbol{\pi}^0, \epsilon)$, each $\boldsymbol{\pi}_t$ can be expressed as (4) satisfying $\boldsymbol{\nu} \in \mathcal{N}_{\epsilon'}(\tilde{\boldsymbol{\nu}})$, where $\boldsymbol{\nu} = \{\boldsymbol{\nu}_t, t \in \{1, \ldots, T\}\}$, $\tilde{\boldsymbol{\nu}} = \{\tilde{\boldsymbol{\nu}}_t, t \in \{1, \ldots, T\}\}$ and $\tilde{\boldsymbol{\nu}}_t = (\tilde{\nu}_{t1}, \tilde{\nu}_{t2}, \ldots)'$ is a probability vector, where

$$\tilde{\nu}_{tm} = \begin{cases} \nu^0_{tm-k^+_{t-1}}, & (k^+_{t-1} < m \le k^+_t), \\ 0, & (\text{otherwise}), \end{cases}$$

therefore $\tilde{\nu}_{tf(t,h)} = \nu^0_{th}$ where $f(t,h) = k^+_{t-1} + h$ for $1 \le h \le k_t$. Also, $\epsilon' = \epsilon/2\prod_{j=1}^{p} d_j$, and $\boldsymbol{\psi}^{(j)}_{k^+_{t-1}+h} \in \mathcal{N}_{\epsilon''}(\boldsymbol{\psi}^{(j)}_{th})$ for $h = 1, \ldots, k_t$ and $t = 1, \ldots, T$, where $\epsilon'' = \epsilon/2\sum_t p_t k_t p \prod_j d_j$.

We consider the intervals (a_{th}, b_{th}) in the real line for W_{th} in the proposed prior for $h = 1, \ldots, k^+_t$, and $t = 1, \ldots, T$, where

$$a_{th} = \begin{cases} g^{-1}\{\max(\tilde{\nu}_{th} - \tilde{\epsilon}, 0)\}, & (h = 1), \\ g^{-1}\left\{\dfrac{\max(\tilde{\nu}_{th} - \tilde{\epsilon}, 0)}{\prod_{l<h}\{1 - g(W_{tl})\}}\right\}, & (h = 2, \ldots, k^+_t), \end{cases}$$

$$b_{th} = \begin{cases} g^{-1}\{\tilde{\nu}_{th} + \tilde{\epsilon}\}, & (h = 1), \\ g^{-1}\left\{\dfrac{\tilde{\nu}_{th} + \tilde{\epsilon}}{\prod_{l<h}\{1 - g(W_{tl})\}}\right\}, & (h = 2, \ldots, k^+_t), \end{cases}$$

where $\tilde{\epsilon} = \epsilon'/2\sum_t p_t k^+_t$. In this case, it is straightforward to check $|\nu_{th} - \tilde{\nu}_{th}| < \tilde{\epsilon}$ for $h = 1, \ldots, k^+_t$ and the proposed prior assigns positive probability to these intervals. Then, the distance between $\boldsymbol{\nu}$ and $\tilde{\boldsymbol{\nu}}$ is

$$\begin{aligned} \|\boldsymbol{\nu} - \tilde{\boldsymbol{\nu}}\| &= \sum_{t=1}^{T} p_t \sum_{h=1}^{\infty} |\nu_{th} - \tilde{\nu}_{th}|, \\ &= \sum_{t=1}^{T} p_t \sum_{h=1}^{k^+_t} |\nu_{th} - \tilde{\nu}_{th}| + \sum_{t=1}^{T} p_t \sum_{h>k^+_t} \nu_{th}, \qquad \text{(C.1)} \\ &< 2\tilde{\epsilon} \sum_{t=1}^{T} p_t k^+_t = \epsilon'. \end{aligned}$$

For the second component in (C.1), $\sum_{h>k^+_t} \nu_{th} < k^+_t\tilde{\epsilon}$ because $\nu_{th} > \tilde{\nu}_{th} - \tilde{\epsilon}$ for $h = 1, \ldots, k^+_t$ and $\sum_{h=1}^{k^+_t} \nu_{th} > 1 - k^+_t\tilde{\epsilon}$. In addition, it is straightforward to show that the proposed prior assigns positive probability to $\mathcal{N}_{\epsilon''}(\boldsymbol{\psi}^{(j)}_{th})$. Therefore, since $D(\boldsymbol{\pi}^0, \epsilon)$ contains such case, $\mathcal{Q}\{D(\boldsymbol{\pi}^0, \epsilon)\} > 0$.

For any $\boldsymbol{\pi} \in D(\boldsymbol{\pi}^0, \epsilon)$,

$$\begin{aligned} &\|\boldsymbol{\pi} - \boldsymbol{\pi}^0\| \\ &= \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} |\pi_{tc_1\cdots c_p} - \pi^0_{tc_1\cdots c_p}|, \\ &= \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left| \sum_{h=1}^{\infty} \nu_{th} \prod_{j=1}^{p} \psi^{(j)}_{hc_j} - \sum_{l=1}^{k_t} \nu^0_{tl} \prod_{j=1}^{p} \psi^{(j)}_{tlc_j} \right|, \\ &= \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left| \sum_{h=1}^{k_t} \left(\nu_{tk^+_{t-1}+h} \prod_{j=1}^{p} \psi^{(j)}_{k^+_{t-1}+hc_j} - \nu^0_{th} \prod_{j=1}^{p} \psi^{(j)}_{thc_j} \right) \right. \\ &\quad \left. + \sum_{l \le k^+_{t-1}, k^+_t < l} \nu_{tl} \prod_{j=1}^{p} \psi^{(j)}_{lc_j} \right|, \\ &\le \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left(\sum_{h=1}^{k_t} \left| \nu_{tk^+_{t-1}+h} \prod_{j=1}^{p} \psi^{(j)}_{k^+_{t-1}+hc_j} - \nu^0_{th} \prod_{j=1}^{p} \psi^{(j)}_{thc_j} \right| \right. \\ &\quad \left. + \sum_{l \le k^+_{t-1}, k^+_t < l} \nu_{tl} \right), \\ &\le \sum_{t=1}^{T} p_t \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left(\sum_{h=1}^{k_t} \left| \nu_{tk^+_{t-1}+h} - \nu^0_{th} \right| + \sum_{l=1}^{k_t} \sum_{j=1}^{p} \left| \psi^{(j)}_{k^+_{t-1}+lc_j} - \psi^{(j)}_{tlc_j} \right| \right. \\ &\quad \left. + \sum_{l \le k^+_{t-1}, k^+_t < l} \nu_{tl} \right), \\ &= \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \sum_{t=1}^{T} p_t \sum_{h=1}^{\infty} |\nu_{th} - \tilde{\nu}_{th}| \\ &\quad + \sum_{t=1}^{T} p_t \sum_{l=1}^{k_t} \sum_{j=1}^{p} \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \left| \psi^{(j)}_{k^+_{t-1}+lc_j} - \psi^{(j)}_{tlc_j} \right|, \\ &< \prod_{j=1}^{p} d_j \epsilon' + \sum_{t=1}^{T} p_t k_t p \prod_{j=1}^{p} d_j \epsilon'', \\ &= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

Therefore, $\boldsymbol{\pi} \in \mathcal{N}_\epsilon(\boldsymbol{\pi}^0)$ and $D(\boldsymbol{\pi}^0, \epsilon) \subset \mathcal{N}_\epsilon(\boldsymbol{\pi}^0)$. Hence, $\mathcal{Q}\{\mathcal{N}_\epsilon(\boldsymbol{\pi}^0)\} > 0$.

SUPPLEMENTARY MATERIALS

Detailed descriptions of the induced priors.

[Received May 2012. Revised May 2013.]

REFERENCES

Agresti, A. (2002), *Categorical Data Analysis* (2nd ed.), New York: Wiley. [1324,1327,1332]

Albert, J. H., and Chib, S. (2001), "Sequential Ordinal Modeling With Applications to Survival Data," *Biometrics*, 57, 829–836. [1328]

Bhattacharya, A., and Dunson, D. B. (2012), "Simplex Factor Models for Multivariate Unordered Categorical Data," *Journal of the American Statistical Association*, 107, 362–377. [1324]

Bush, C., and MacEachern, S. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275–285. [1327]

Carter, C. K., and Kohn, R. (1994), "On Gibbs Sampling for State Space Models," *Biometrika*, 81, 541–553. [1328]

Chung, Y., and Dunson, D. B. (2009), "Nonparametric Bayes Conditional Distribution Modeling With Variable Selection," *Journal of the American Statistical Association*, 104, 1646–1660. [1325,1326]

——— (2011), "The Local Dirichlet Process," *Annals of the Institute for Statistical Mathematics*, 63, 59–80. [1325]
Dahinden, C., Kalisch, M., and Buehlmann, P. (2010), "Decomposition and Model Selection for Large Contingency Tables," *Biometrical Journal*, 52, 233–252. [1329]
Dahinden, C., Parmigiani, G., Emerick, M., and Buehlmann, P. (2007), "Penalized Likelihood for Sparse Contingency Tables With an Application to Full-Length cDNA Libraries," *BMC Bioinformatics*, 8, 476. [1329,1330]
de Jong, P., and Shephard, N. (1995), "The Simulation Smoother for Time Series Models," *Biometrika*, 82, 339–350. [1328]
Dobra, A., and Lenkoski, A. (2011), "Copula Gaussian Graphical Models and Their Application to Modeling Functional Disability Data," *Annals of Applied Statistics*, 5, 969–993. [1324]
Doornik, J. (2006), *Ox: Object Oriented Matrix Programming*, London: Timberlake Consultants Press. [1]
Dunson, D. B. (2006), "Bayesian Dynamic Modeling of Latent Trait Distributions," *Biostatistics*, 7, 551–568. [1325]
Dunson, D. B., and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, 104, 1042–1051. [1324,1325,1329]
Durbin, J., and Koopman, S. J. (2002), "Simple and Efficient Simulation Smoother for State Space Time Series Analysis," *Biometrika*, 89, 603–616. [1328]
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, 3430–3445. [1324]
Früwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models," *Journal of Time Series Analysis*, 15, 183–202. [1328]
Ghosh, J., and Ramamoorthi, R. (2003), *Bayesian Nonparametrics*, Berlin: Springer Verlag. [1327]
Griffin, J. E., and Steel, M. F. J. (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194. [1325]
Harshman, R. A. (1970), "Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multi-Modal Factor Analysis," *UCLA Working Papers in Phonetics*, 16, 1–84. [1325]
Ishwaran, H., and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173. [1327,1336]
Kalli, M., Griffin, J. E., and Walker, S. G. (2011), "Slice Sampling Mixture Models," *Statistics and Computing*, 65, 93–105. [1325,1327]
Kolda, T. G. (2001), "Orthogonal Tensor Decompositions," *SIAM Journal on Matrix Analysis and Applications*, 23, 243–255. [1325]
MacEachern, S. N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, pp. 50–55. [1325]
——— (2000), "Dependent Dirichlet Processes," Technical Report, Department of Statistics, Ohio State University. [1325]
Manrique-Vallier, D., and Reiter, J. (2013), "Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeroes," *Journal of Computational and Graphical Statistics*, to appear, DOI: 10.1080/10618600.2013.844700. [1327]
Papaspiliopoulos, O., and Roberts, G. O. (2008), "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models," *Biometrika*, 95, 169–186. [1327]
Ren, L., Dunson, D., Lindroth, S., and Carin, L. (2010), "Dynamic Nonparametric Bayesian Models for Analysis of Music," *Journal of the American Statistical Association*, 105, 458–472. [1325]
Rodriguez, A., and Dunson, D. B. (2011), "Nonparametric Bayesian Models Through Probit Stick-Breaking Processes," *Bayesian Analysis*, 6, 145–177. [1325,1326]
Rodriguez, A., and Horst, E. T. (2008), "Bayesian Dynamic Density Estimation," *Bayesian Analysis*, 3, 339–366. [1325]
Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650. [1325]
Walker, S. G. (2007), "Sampling the Dirichlet Mixture Model With Slices," *Communications in Statistics-Simulation and Computation*, 36, 45–54. [1327]

Downloaded by [Board of Governors of the Federal Reserve System] at 11:22 23 October 2017