Journal of the American Statistical Association

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

Bayesian Factorizations of Big Sparse Tensors

Jing Zhou, Anirban Bhattacharya, Amy H. Herring & David B. Dunson

To cite this article: Jing Zhou, Anirban Bhattacharya, Amy H. Herring & David B. Dunson (2015) Bayesian Factorizations of Big Sparse Tensors, Journal of the American Statistical Association, 110:512, 1562-1576, DOI: 10.1080/01621459.2014.983233

To link to this article: http://dx.doi.org/10.1080/01621459.2014.983233

Accepted author version posted online: 14 Nov 2014.
Published online: 15 Jan 2016.

Bayesian Factorizations of Big Sparse Tensors

Jing ZHOU, Anirban BHATTACHARYA, Amy H. HERRING, and David B. DUNSON

It has become routine to collect data that are structured as multiway arrays (tensors). There is an enormous literature on low rank and sparse matrix factorizations, but limited consideration of extensions to the tensor case in statistics. The most common low rank tensor factorization relies on parallel factor analysis (PARAFAC), which expresses a rank k tensor as a sum of rank one tensors. In contingency table applications in which the sample size is massively less than the number of cells in the table, the low rank assumption is not sufficient and PARAFAC has poor performance. We induce an additional layer of dimension reduction by allowing the effective rank to vary across dimensions of the table. Taking a Bayesian approach, we place priors on terms in the factorization and develop an efficient Gibbs sampler for posterior computation. Theory is provided showing posterior concentration rates in high-dimensional settings, and the methods are shown to have excellent performance in simulations and several real data applications.

KEY WORDS: Bayesian; Categorical data; Contingency table; Log-linear model; Low rank; PARAFAC; Sparsity; Tensor factorization.

1. INTRODUCTION

In many application areas, it is standard to collect high-dimensional categorical data, which can be organized as a contingency table. Contingency tables correspond to a multiway array or tensor, with each cell containing a count of the number of individuals having a particular combination of values for the categorical variables being measured. In contingency table analyses, the focus is typically on inferring associations among the different variables, but challenges arise when there are many variables, so that the number of cells in the table is vastly bigger than the sample size. Usual log-linear modeling approaches have difficulty scaling to such settings; even when sparsity is imposed, the number of possible terms in the model is so massive that computation becomes intractable. This article proposes a solution to this problem using a novel class of Bayesian tensor factorizations.

For subjects $i = 1, \ldots, n$, data consist of multivariate categorical response vectors, $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$, with $y_{ij} \in \{1, \ldots, d_j\}$ for $j = 1, \ldots, p$. Letting $\Pr(y_{i1} = c_1, \ldots, y_{ip} = c_p) = \pi_{c_1 \ldots c_p}$ denote the probability mass function, the tensor of interest is $\pi = \{\pi_{c_1 \ldots c_p}\} \in \Pi_{d_1 \times \cdots \times d_p}$, with $\Pi_{d_1 \times \cdots \times d_p}$ the space of p-way probability tensors having d_j rows in the jth direction. Probability tensors have nonnegative elements that sum to one across all the cells, with the total number of cells being $\prod_{j=1}^{p} d_j$. When p is not small, we obtain $\prod_{j=1}^{p} d_j \gg n$, so that the vast majority of the cells of the table have zero counts.

To combat this data sparsity, it is necessary to substantially reduce dimensionality in estimating π. The usual way to accomplish this is through a low rank assumption. Unlike for matrices, there is no unique definition of rank but the most common convention is to define the rank k of a tensor π as the smallest value of k such that π can be expressed as

$$\pi = \sum_{h=1}^{k} \psi_h^{(1)} \otimes \cdots \otimes \psi_h^{(p)}, \tag{1}$$

which is the sum of k rank one tensors, each an outer product of vectors[1] for each dimension (Kolda and Bader 2009). Expression (1) is commonly referred to as parallel factor analysis (PARAFAC) (Harshman 1970; Bro 1997). For k small, the number of parameters is massively reduced from $\prod_{j=1}^{p} d_j$ to $k \sum_{j=1}^{p} d_j$; as the low rank assumption often holds approximately, this leads to an effective approach in many applications, and a rich variety of algorithms are available for estimation.

However, the decrease in degrees of freedom from exponential in p to linear in p is not sufficient when p is big. Large p small n problems arise routinely, and a usual solution outside of tensor settings is to incorporate sparsity. For example, in linear regression, many of the coefficients are set to zero (Tibshirani 1996; Scott and Berger 2010), while in estimation of large covariance matrices, sparse factor models are used that assume few factors and many zeros in the factor loadings matrices (West 2003; Carvalho et al. 2008). In contingency table analyses, sparsity can be imposed by setting many coefficients to zero in a saturated log-linear model, but as p grows it rapidly becomes computationally impossible to consider even all possible two-way interactions. Instead, a salient feature of the proposed approach is the ability to recover near sparse models in the log-linear parameterization by inducing shrinkage on the log-linear parameters. We remark here that in factorization (1), including zeros in the component vectors $\{\psi_h^{(j)}\}$ is not a viable solution, particularly as we do not want to enforce exact zeros in blocks of the tensor π.

Our notion is as follows. For component h ($h = 1, \ldots, k$), we partition the dimensions into two mutually exclusive subsets $S_h \cup S_h^c = \{1, \ldots, p\}$. The proposed sparse PARAFAC

Jing Zhou, School of Public Health, University of North Carolina, Chapel Hill, NC 27599 (E-mail: *amanistat@gmail.com*). Anirban Bhattacharya is Assistant Professor, Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843 (E-mail: *anirbanb@stat.tamu.edu*). Amy H. Herring, Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (E-mail: *aherring@bios.unc.edu*). David B. Dunson is Professor, Department of Statistical Science, Box 90251, 214 Old Chemistry Building, Duke University, Durham, NC 27708-0251 (E-mail: *dunson@stat.duke.edu*). The research was partially supported by grant number R01 ES017240-01 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH). The second author would like to acknowledge support from the Office of Naval Research (ONR BAA 14-0001).

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

[1]For $p = 2$, $\psi^{(1)} \otimes \psi^{(2)} = \psi^{(1)}\psi^{(2)\mathrm{T}}$. In general, $(\psi^{(1)} \otimes \cdots \otimes \psi^{(p)})_{c_1 \ldots c_p} = \psi_{c_1}^{(1)} \ldots \psi_{c_p}^{(p)}$.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2014.983233

(sp-PARAFAC) factorization is then

$$\pi = \sum_{h=1}^{k} \psi_h^{(1)} \otimes \cdots \otimes \psi_h^{(p)}, \quad \psi_h^{(j)} = \psi_0^{(j)} \text{ for } j \in S_h^c. \quad (2)$$

Hence, instead of having to introduce a separate vector $\psi_h^{(j)}$ for every h and j, we allow there to be more degrees of freedom used to characterize the tensor structure in certain directions than in others by setting a large fraction of the $\psi_h^{(j)}$s to a *baseline factor* $\psi_0^{(j)}$. If $j \in S_h^c$ for $h = 1, \ldots, k$, then the jth variable is independent of the other variables with $\Pr(y_{ij} = c_j) = \psi_{0c_j}^{(j)}$. By including $j \in S_h^c$ for some but not all $h \in \{1, \ldots, k\}$ one can use fewer degrees of freedom in characterizing the interaction between the jth factor and the other factors. In practice, we will learn $\{S_h\}$ using a Bayesian approach, as the appropriate lower dimensional structure is typically not known in advance.

We conjecture that many categorical datasets can be concisely represented via (2), with results substantially improved over usual PARAFAC factorizations due to the second layer of dimension reduction. Contingency table analysis is routine in practice; refer to Agresti (2002) and Fienberg and Rinaldo (2007). However, in stark contrast to the well developed literature on linear regression and covariance matrix estimation in big data settings, very few flexible methods are scalable beyond small tables. Our interest is in situations where the dimensionality p is comparable or even larger than the number of samples n.

2. SPARSE FACTOR MODELS FOR TABLES

2.1 Model and Prior

We focus on a Bayesian implementation of sp-PARAFAC in (2). Let $\mathcal{S}^{r-1} = \{x \in \Re^r : x_j \geq 0, \sum_{j=1}^r x_j = 1\}$ denote the $(r-1)$-dimensional probability simplex. Dunson and Xing (2009) proposed the following probabilistic PARAFAC factorization.

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

where $\nu = \{\nu_h\} \in \mathcal{S}^{k-1}$ and $\lambda_h^{(j)} = (\lambda_{h1}^{(j)}, \ldots, \lambda_{hd_j}^{(j)}) \in \mathcal{S}^{d_j-1}$ is a vector of probabilities of $y_{ij} = 1, \ldots, d_j$ in component h. Introducing a latent subpopulation index $z_i \in \{1, \ldots, k\}$ for subject i, the elements of y_i are conditionally independent given z_i with $\Pr(y_{ij} = c_j \mid z_i = h) = \lambda_{hc_j}^{(j)}$, and marginalizing out the latent index z_i leads to a mixture of product multinomial distribution for y_i. Placing Dirichlet priors on the component vectors leads to a simple and efficient Gibbs sampler for posterior computation. We will refer to this model (3) as standard PARAFAC.

This approach has excellent performance in small to moderate p problems, but as p increases there is an inevitable breakdown point. The number of parameters increases linearly in p, as for other PARAFAC factorizations, so problems arise as p approaches the order of n or $p \gg n$. For example, we are particularly motivated by epidemiology studies collecting many categorical predictors, such as occupation type, demographic variables, and single nucleotide polymorphisms. For continuous response vectors $y_i \in \Re^p$, there is a well developed literature on Gaussian sparse factor models that are adept at accommodating $p \gg n$ data (West 2003; Lucas et al. 2006; Carvalho et al. 2008; Bhattacharya and Dunson 2011). These models include many zeros in the loadings matrices to induce additional dimension reduction on top of the low rank assumption. Pati et al. (2014) provided theoretical support through characterizing posterior concentration.

Our sp-PARAFAC factorization provides an analog of sparse factor models in the tensor setting. We let

$$\pi_{c_1 \ldots c_p} = \sum_{h=1}^{k} \nu_h \prod_{j \in S_h} \lambda_{hc_j}^{(j)} \prod_{j \in S_h^c} \lambda_{0c_j}^{(j)}, \quad (4)$$

where $|S_h| \ll p$ ($|S|$ denotes the cardinality of a set S) and the $\lambda_0^{(j)}$ vectors are *fixed in advance*; we consider two cases:

$$\text{(i) } \lambda_0^{(j)} = \left(\frac{1}{d_j}, \ldots, \frac{1}{d_j}\right)^{\mathrm{T}}$$

and

$$\text{(ii) } \lambda_0^{(j)} = \left(\frac{1}{n}\sum_{i=1}^{n} 1(y_{ij} = 1), \ldots, \frac{1}{n}\sum_{i=1}^{n} 1(y_{ij} = d_j)\right)^{\mathrm{T}},$$

corresponding to a discrete uniform and empirical estimates of the marginal category probabilities. By fixing the baseline dictionary vectors $\{\lambda_0^{(j)}\}$ in advance, and allocating a large subset of the variables within each cluster h to the baseline component, we dramatically reduce the size of the model space. In particular, the probability tensor π in (4) can be parameterized as $\theta_\pi = (\nu, \{S_h\}_{1 \leq h \leq k}, \{\lambda_h^{(j)}\}_{1 \leq h \leq k, j \in S_h})$, where $\nu \in \mathcal{S}^{k-1}$, $S_h \subset \{1, \ldots, p\}$, $\lambda_h^{(j)} \in \mathcal{S}^{d_j-1}$. Thus, the effective number of model parameters is now reduced to $(k-1) + \sum_{h=1}^{k} |S_h| + \sum_{h=1}^{k} \sum_{j \in S_h} (d_j - 1)$, which is substantially smaller than the $(k-1) + \sum_{j=1}^{p} k(d_j - 1)$ parameters in the original specification, provided $|S_h| \ll p$ for all $h = 1, \ldots k$. The size of S_h is penalized via a sparsity favoring prior on $|S_h|$ in (5) below. We will illustrate that this can lead to huge differences in practical performance.

Completing a Bayesian specification with priors for the unknown parameter vectors and expressing the model in hierarchical form, we have[2]

$$\begin{aligned}
y_{ij} &\sim \text{Mult}\big(\{1, \ldots, d_j\}; \lambda_{z_i 1}^{(j)}, \ldots, \lambda_{z_i d_j}^{(j)}\big), \\
\Pr(z_i = h) = \nu_h &= V_h \prod_{l<h} (1 - V_l), \\
\lambda_h^{(j)} &= (1 - \zeta_{jh})\lambda_0^{(j)} + \zeta_{jh}\tilde{\lambda}_h^{(j)}, \\
\zeta_{jh} &\sim \text{Bernoulli}(\tau_h), \quad \tilde{\lambda}_h^{(j)} \sim \text{Diri}(a_{j1}, \ldots, a_{jd_j}), \\
V_h &\sim \text{Beta}(1, \alpha), \quad \alpha \sim \text{Gamma}(a_\alpha, b_\alpha), \\
\tau_h &\sim \text{Beta}(1, \gamma). \qquad (5)
\end{aligned}$$

It is implicit that the probability statements in any row are made conditionally on parameters appearing in later rows. Clearly, the hierarchical prior in (5) is supported on the space of probability tensors with a sp-PARAFAC decomposition as in (4), since (5) is equivalent to letting the subset-size $|S_h| \sim \text{Binom}(p, \tau_h)$ and drawing a random subset S_h uniformly from all subsets of $\{1, \ldots, p\}$ of size $|S_h|$ in (4). A stick-breaking prior

[2] $\text{Mult}(\{1, \ldots, d\}; \lambda_1, \ldots, \lambda_d)$ denotes a discrete distribution on $\{1, \ldots, d\}$ with probabilities $\lambda_1, \ldots, \lambda_d$ associated to each atom.

(Sethuraman 1994) is chosen for the component weights $\{\nu_h\}$, taking a nonparametric Bayes approach that allows $k = \infty$, with a hyperprior placed on the concentration parameter α in the stick-breaking process to allow the data to inform more strongly about the component weights. The probability of allocation τ_h to the *active* (nonbaseline) category in component h is chosen as Beta$(1, \gamma)$, with $\gamma > 1$ favoring allocation of many of the $\lambda_h^{(j)}$'s to the baseline category $\lambda_0^{(j)}$. In the limiting case as $\gamma \to \infty$, the joint probability tensor π becomes an outer product of the baseline probabilities for the individual variables, $\pi = \lambda_0^{(1)} \otimes \cdots \otimes \lambda_0^{(p)}$. On the other hand, as $\gamma \to 0$, one reduces back to standard PARAFAC (3).

Line 2 of expression (5) is key in inducing the second level of dimensionality reduction in our Bayesian sparse PARAFAC factorization. The inclusion of the baseline component that does not vary with h massively reduces the number of parameters, and can additionally be argued to have minimal impact on the flexibility of the specification. The $\lambda_h^{(j)}$'s are incorporated within $\prod_{j=1}^p \lambda_{hc_j}^{(j)}$, which for large p is highly concentrated around its mean since the $\lambda_h^{(j)}$'s are independent across j. This is a manifestation of the concentration of measure phenomenon (Talagrand 1996), which roughly states that a random variable that depends in a smooth way on the influence of many independent variables, but not too much on any one of them, is essentially constant. For example, if $\theta_j \overset{\text{iid}}{\sim} U(0, 1)$ and $\Theta = \prod_{j=1}^p \theta_j$, then $\mathrm{E}(\Theta) = (1/2)^p$ and $\mathrm{var}(\Theta) = (1/3)^p$, which rapidly converges to zero. This implies that replacing a large randomly chosen subset of the $\lambda_h^{(j)}$'s by $\lambda_0^{(j)}$ should have minimal impact on modeling flexibility.

2.2 Induced Prior in Log-Linear Parameterization

An important challenge is accommodating higher order interactions, which play an important role in many applications (e.g., genetics), but are typically assumed to equal zero for tractability. As p grows, it is challenging to even accommodate two-way interactions in traditional categorical data models (log-linear, logistic regression) due to an explosion in the number of terms. In contrast, the tensor factorization does not explicitly parameterize interactions, but indirectly induces a shrinkage prior on the terms in a saturated log-linear model. One can then reparameterize in terms of the log-linear model in conducting inferences in a post model-fitting step. Focusing on binary variables for ease of exposition ($d_j = d = 2$ for all j) and working with the *corner* parameterization for log-linear models (see Massam, Liu, and Dobra 2009, Sec. 2), we illustrate the induced priors on the main effects and interactions below. Details of transforming to the log-linear parameterization are provided in Appendix A.

We first focus on a case where $p = 3$ and $d_j = d = 2$ for $j = 1, 2, 3$. Given a $2 \times 2 \times 2$ tensor π, we can equivalently characterize π in terms of its log-linear parameterization

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{13}, \beta_{23}, \beta_{123})^{\mathrm{T}},$$

consisting of three main effect terms $\beta_1, \beta_2, \beta_3$, three second-order interaction terms $\beta_{12}, \beta_{13}, \beta_{23}$, and one third order interaction term β_{123}. We generate 10^4 random probability tensors $\pi^{(t)} = (\pi^{(t)}_{c_1c_2c_3})$, $t = 1, \ldots, 10^4$, distributed according to (5), where we fix the baseline $\lambda_0^{(j)} = (1/2, 1/2)^{\mathrm{T}}$ for all j. Given each prior sample $\pi^{(t)}$, we transform to the log-linear parameterization to obtain a sample $\boldsymbol{\beta}^{(t)}$ from the induced prior on $\boldsymbol{\beta}$, which allows us to estimate the marginal densities of the main effects and interactions and also their joint distributions. In particular, since γ plays an important role in placing weights on the baseline component, we would like to see how our induced priors differ with different γ values.

In our simulation exercise, we fix three values of γ, namely, $\gamma = 1, 5, 20$. Note that $\gamma = 1$ corresponds to a $U(0, 1)$ prior on τ_h. For different values of γ, we show the histograms of one main effect term β_1, one two-way interaction β_{12} and the three-way interaction β_{123} in Figure 1. Table 1 additionally reports summary statistics.

In high-dimensional regression, $y_i = x_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$, there has been substantial interest in shrinkage priors, which draw β_j a priori from a density concentrated at zero with heavy tails. Such priors strongly shrink the small coefficients to zero, while limiting shrinkage of the larger signals (Park and Casella 2008; Carvalho, Polson, and Scott 2010; Polson and Scott 2010; Hans 2011; Armagan, Dunson, and Lee 2013a). In Figure 1, the induced prior on any of the log-linear model parameters is symmetric about zero, with a large spike very close to zero, and heavy tails. Thus, we have indirectly induced a shrinkage prior on the main effects and interactions through our tensor decomposition approach. In addition, the prior automatically shrinks more aggressively as the interaction order increases. Such greater shrinkage of interactions is commonly recommended (Gelman et al. 2008). Importantly, we do not zero out small interactions but allow many small coefficients, which is an important distinction in applications, such as genomics, having many small signals. The hyperparameter γ serves as a penalty controlling the degree of shrinkage. We note that greedy methods like iterative hard thresholding or their convex relaxations like the lasso can only produce exactly sparse models.

The induced priors on the main effects and interactions are not analytically tractable and it seems difficult to obtain expressions for $\mathbb{P}(|\beta_j| < t)$ and $\mathbb{P}(|\beta_{jj'}| < t)$ for small t to theoretically compare the induced degree of shrinkage. However, if we truncate the stick-breaking prior in (5) to a finite number of components and set the baseline $\lambda_0^{(j)} = (1/2, 1/2)^{\mathrm{T}}$ for all j, then the induced shrinkage priors on the main effects and interactions have an explicit point mass at zero and we can compare the mass at zero to compare the degree of shrinkage. Under this setting, we provide expressions for $\mathbb{P}(\beta_j = 0)$ and $\mathbb{P}(\beta_{jj'} = 0)$ in Proposition 2.1 below. A proof can be found in Appendix A.

Proposition 2.1. Suppose $d_j = 2$ for all j. If the stick-breaking prior in (5) is truncated to K components and the baseline $\lambda_0^{(j)} = (1/2, 1/2)^{\mathrm{T}}$ for all j, then for any $1 \le j \ne j' \le p$,

$$\mathbb{P}(\beta_j = 0) = \left(\frac{\gamma}{1+\gamma}\right)^K,$$
$$\mathbb{P}(\beta_{jj'} = 0) = 2\left(\frac{\gamma}{1+\gamma}\right)^K - \left(\frac{\gamma}{2+\gamma}\right)^K. \qquad (6)$$

For fixed γ, both $\mathbb{P}(\beta_j = 0)$ and $\mathbb{P}(\beta_{jj'} = 0)$ become zero in the limit as $K \to \infty$, so that the point mass at zero vanishes. Second, for fixed K, $\mathbb{P}(\beta_j = 0)$ and $\mathbb{P}(\beta_{jj'} = 0)$ are both increasing functions of γ, implying that larger values of γ induce more shrinkage. Third, for fixed K and γ, $\mathbb{P}(\beta_j = 0) \le \mathbb{P}(\beta_{jj'} =$

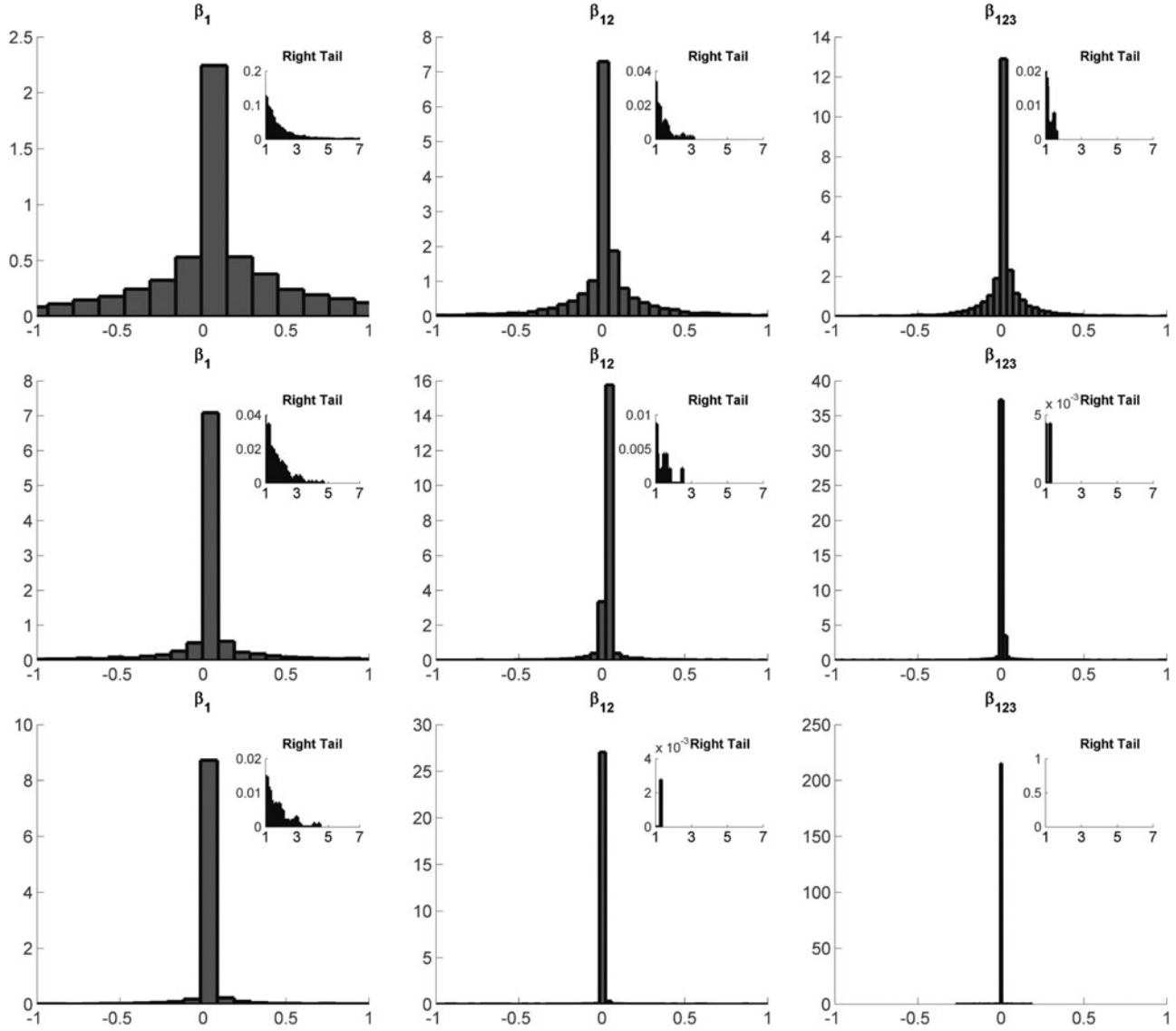

Figure 1. Histograms of induced priors for one main effect β_1, one two-way interaction β_{12}, and the three-way interaction β_{123}. Top row: $\gamma = 1$; middle row: $\gamma = 5$; bottom row: $\gamma = 20$.

0), implying the two-way interaction terms are shrunk more aggressively compared to the main effects. A similar result is true for any higher-order interactions as evident from the proof. Finally, all of these conform to the findings from the simulation study which leads us to believe that the results are true in greater generality.

Table 1. Summary statistics of induced priors on coefficients in log-linear model parameterization

γ	Coefficient	Mean	Std.dev	Min	Max	Skewness	Kurtosis
1	β_1	0.014	0.831	−6.765	6.389	0.210	9.109
1	β_{12}	−0.002	0.340	−2.895	3.105	−0.025	16.583
1	β_{123}	0.002	0.196	−2.223	2.632	0.525	24.686
5	β_1	−0.002	0.485	−5.648	5.433	0.031	27.980
5	β_{12}	0.000	0.124	−2.085	2.244	0.495	93.438
5	β_{123}	0.000	0.051	−1.214	0.745	−3.701	159.360
20	β_1	0.002	0.246	−3.109	5.669	2.474	99.554
20	β_{12}	0.000	0.042	−1.126	1.819	9.488	632.790
20	β_{123}	0.000	0.009	−0.664	0.214	−44.051	3014.000

Our next set of simulations involve larger values of p, where the necessity of the regularization implied by γ becomes strikingly evident. Let $\boldsymbol{\beta}_m = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ denote the p-dimensional vector consisting of all the main effects. Our object of interest now is the induced joint prior distribution of $\boldsymbol{\beta}_m$. In particular, we focus on two univariate functionals of $\boldsymbol{\beta}_m$: (i) the l_1 norm $\|\boldsymbol{\beta}_m\|_1 = \sum_{j=1}^{p} |\beta_j|$, and (ii) the *numerical sparsity* $s(\boldsymbol{\beta}_m) = (\|\boldsymbol{\beta}_m\|_1/\|\boldsymbol{\beta}_m\|_2)^2$. For $x \in \mathbb{R}^p$, $\|x\|_1$ and $s(x)$ are continuous functions of x which are commonly used as surrogate measures of sparsity (Lopes 2013). In particular, it follows from the Cauchy–Schwatrz inequality that $s(x)$ is a sharp lower bound to $\|x\|_0$, the number of nonzero entries in x.

We consider two values of p, namely $p = 50$ and $p = 200$, and sample 5×10^4 draws from the prior (5) in each case. Transforming the prior draws of π to the log-linear parameterization $\boldsymbol{\beta}$, we plot histograms of the induced density of $\|\boldsymbol{\beta}_m\|_1$ in Fig-

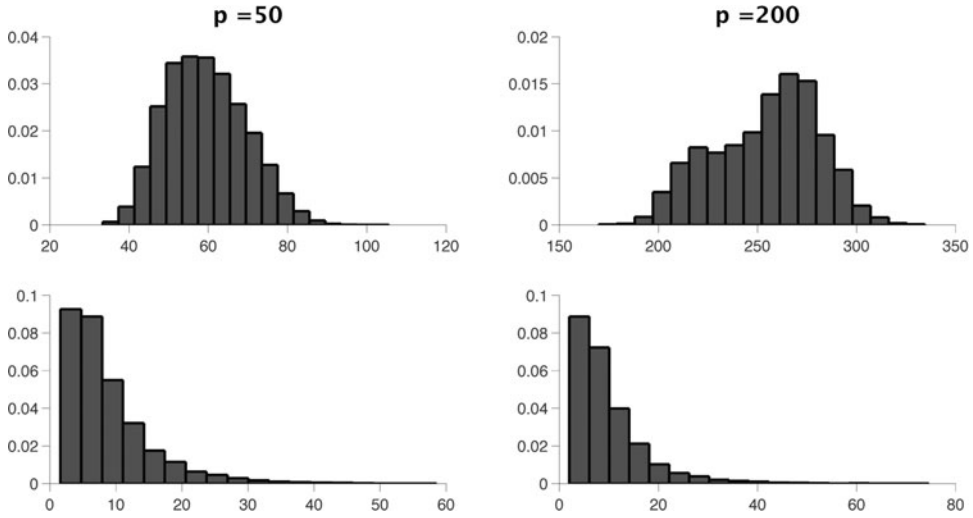

Figure 2. Histograms of $\|\boldsymbol{\beta}_m\|_1$ for $p = 50$ (left panel) and $p = 200$ (right panel). The top row corresponds to the standard PARAFAC model and the bottom row to the sp-PARAFAC model with $\gamma = 0.1p$.

ure 2 and that of $s(\boldsymbol{\beta}_m)$ in Figure 3. In each of the two figures, the top row corresponds to $\gamma = 0$ so that the sp-PARAFAC formulation reduces back to the standard PARAFAC (3), while the bottom row corresponds to the sp-PARAFAC with γ/p set to a constant $\kappa \in (0, 1)$. Figures 2 and 3 reveal a highly undesirable property of the standard PARAFAC in high dimensions, where the entire distributions of $\|\boldsymbol{\beta}_m\|_1$ and $s(\boldsymbol{\beta}_m)$ shift to the right with increasing p, with $\mathbb{E}\|\boldsymbol{\beta}_m\|_1, \mathbb{E}\, s(\boldsymbol{\beta}_m) \asymp p$. The induced prior on $\boldsymbol{\beta}_m$ for the standard PARAFAC clearly lacks any automatic multiplicity adjustment property (Scott and Berger 2010), and would bias inferences for moderate to large values of p. On the other hand, under the sp-PARAFAC model with $\gamma = \kappa p$, the induced priors on $\|\boldsymbol{\beta}_m\|_1$ and $s(\boldsymbol{\beta}_m)$ are robust to increasing p, as evident from the bottom rows of Figures 2 and 3. The choice $\gamma = \kappa p$ acts as a penalty on the size of the non-null group, forcing the prior to concentrate on smaller subsets; see Castillo and van der Vaart (2012) for a similar choice of the hyperparameter in a regression setting and also Remark 3.2 in Section 3.3.

3. POSTERIOR CONCENTRATION

3.1 Preliminaries

In this section, we provide theoretical justification to the proposed sp-PARAFAC procedure in high dimensional settings by studying the concentration properties of the posterior with growing sample size. When the parameter space is finite dimensional, it is well known that the posterior contracts at the parametric rate of $n^{-1/2}$ under mild regularity conditions (Ghosal, Ghosh, and van der Vaart 2000). However, we are interested in the asymptotic framework of the dimension $p = p_n$ growing with the sample size n, potentially at a faster rate, reflecting the applications we are interested in. There is a small but increasing literature on asymptotic properties of Bayesian procedures in models with growing dimensionality, with most of the focus being on linear models or generalized linear models belonging to the exponential family; refer to Ghosal (1999, 2000), Belitser and Ghosal (2003), Jiang (2007), Armagan et al. (2013b), Bontemps (2011), and Castillo and van der Vaart

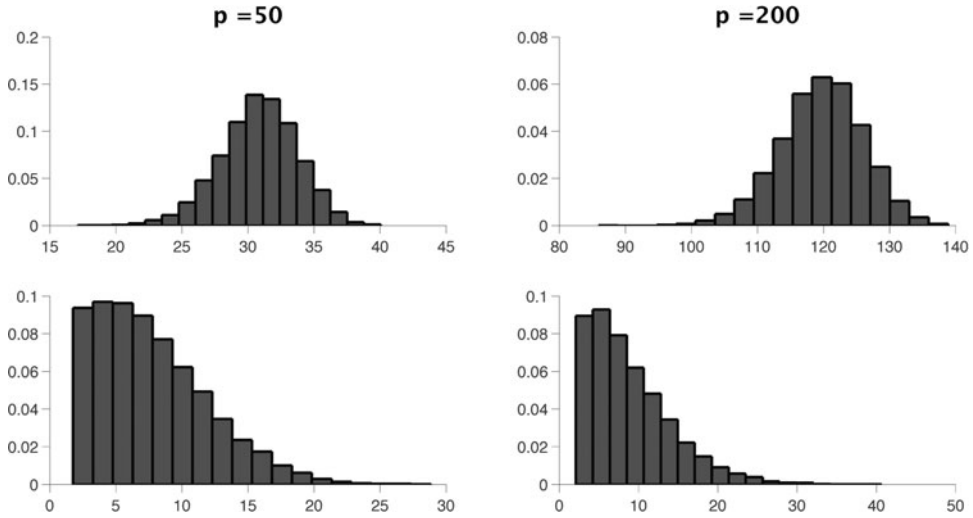

Figure 3. Histograms of the numerical sparsity $s(\boldsymbol{\beta}_m)$ for $p = 50$ (left panel) and $p = 200$ (right panel). The top row corresponds to the standard PARAFAC model and the bottom row to the sp-PARAFAC model with $\gamma = 0.1p$.

(2012) among others. In all these cases, the object of interest is a vector of high-dimensional regression coefficients or more generally, the conditional distribution $f(y \mid x)$ of a univariate response y given high-dimensional predictors x. However, we are concerned here with estimation of the high-dimensional joint probability tensor π.

Let $\mathcal{F}_n$ denote the class of all $d_1 \times \cdots \times d_{p_n}$ probability tensors; we shall assume $d_1 = \cdots = d_{p_n} = d$ in the sequel for notational convenience. Let $\pi^{(0n)} \in \mathcal{F}_n$ be a sequence of true tensors. We observe $y_1, \ldots, y_n \sim \pi^{(0n)}$ and set $\boldsymbol{y}^{(n)} = (y_1, \ldots, y_n)$. We denote the prior distribution on $\mathcal{F}_n$ induced by the sp-PARAFAC formulation by $\mathbb{P}_n$ and the corresponding posterior distribution by $\mathbb{P}_n(\cdot \mid \boldsymbol{y}^{(n)})$.

For two probability tensors $\pi^{(1)}$ and $\pi^{(2)} \in \mathcal{F}_n$, the L_1 distance is defined as

$$\left\|\pi^{(1)} - \pi^{(2)}\right\|_1 = \sum_{c_1=1}^{d} \cdots \sum_{c_{p_n}=1}^{d} \left|\pi^{(1)}_{c_1 \ldots c_{p_n}} - \pi^{(2)}_{c_1 \ldots c_{p_n}}\right|.$$

For a sequence of numbers $\epsilon_n \to 0$ and a constant $M > 0$ independent of ϵ_n, let

$$U_n = \left\{\pi : \left\|\pi - \pi^{(0n)}\right\|_1 \leq M\epsilon_n\right\} \tag{7}$$

denote a ball of radius $M\epsilon_n$ around $\pi^{(0n)}$ in the L_1 norm. We seek to find a minimum possible sequence ϵ_n such that

$$\lim_{n\to\infty} \mathbb{P}_n\left(U_n^c \mid \boldsymbol{y}^{(n)}\right) \to 0, \quad \text{a.s. } \pi^{(0n)}. \tag{8}$$

3.2 Assumptions

In this section, we state our assumptions on the true data generating model and briefly discuss their implications.

Assumption 3.1. The true sequence of probability tensors $\pi^{(0n)}$ are of the form

$$\pi^{(0n)}_{c_1 \ldots c_{p_n}} = \sum_{h=1}^{k_n} \nu_{0h} \prod_{j \in S_{0h}} \lambda^{(0j)}_{hc_j} \prod_{j \in S_{0h}^c} \lambda^{(j)}_{0c_j}, \; 1 \leq c_j \leq d, 1 \leq j \leq p_n, \tag{A0}$$

where $\lambda^{(j)}_0 \in \mathcal{S}^{d-1}$ are assumed to be fixed and known.

We now provide some intuition for assumption (A0). Letting $S_0 = \cup_{h=1}^{k_n} S_{0h}$, we can rewrite the expansion of $\pi^{(0n)}$ in (A0) as

$$\pi^{(0n)}_{c_1 \ldots c_{p_n}} = \sum_{h=1}^{k_n} \nu_{0h} \prod_{j \in S_0} \bar{\lambda}^{(0j)}_{hc_j} \prod_{j \in S_0^c} \lambda^{(j)}_{0c_j}, \tag{9}$$

where

$$\bar{\lambda}^{(0j)}_h = \begin{cases} \lambda^{(0j)}_h & \text{if } j \in S_{0h}, \\ \lambda^{(j)}_0 & \text{if } j \in S_0 \backslash S_{0h}. \end{cases}$$

In (9), the term $\prod_{j \in S_0^c} \lambda^{(j)}_{0c_j}$ doesn't involve h and can be factored out completely. Assumption (A0) thus posits that the variables in S_0^c are marginally independent and the entire dependence structure is driven by the variables in S_0. We shall refer to S_0 and S_0^c as the nonnull and null group of variables respectively.

Let $q_n = |S_0|$ and define a mapping $j \to e_j$ from $\{1, \ldots, q_n\}$ to the ordered elements of S_0, so that $e_1 \leq \cdots \leq e_{q_n}$. As j varies between 1 to q_n, e_j ranges over the elements of S_0. Denote by $\psi^{(0n)}$ the d^{q_n} joint probability tensor for the variables $\{y_{ij} : j \in S_0\}$, so that

$$\psi^{(0n)}_{c_1 \ldots c_{q_n}} = \Pr(y_{ie_1} = c_1, \ldots, y_{ie_{q_n}} = c_n) = \sum_{h=1}^{k_n} \nu_{0h} \prod_{j=1}^{q_n} \bar{\lambda}^{(0e_j)}_{hc_j}. \tag{10}$$

Thus, after factoring out the marginally independent variables in S_0^c, (A0) implies a standard PARAFAC expansion (10) for $\psi^{(0n)}$ with k_n many components. Since any nonnegative tensor admits a standard PARAFAC distribution (Lim and Comon 2009), we can always write an expansion of $\psi^{(0n)}$ as in (10).

The next set of assumptions are provided below.[3]

Assumption 3.2. In addition to (A0), $\pi^{(0n)}$ satisfies

(A1) The number of components $k_n = O(1)$.
(A2) Letting $s_n = \max_{1 \leq h \leq k_n} |S_{0h}|$, one has $s_n = O(\log p_n)$.
(A3) There exists a constant $\varepsilon_0 \in (0, 1)$ such that $\lambda^{(0j)}_{hc} \geq \varepsilon_0$ for all $1 \leq h \leq k_n, 1 \leq c \leq d, j \in S_{0h}$.

(A1) and (A2) imply that the size of the nonnull group is much smaller than p_n, since $q_n = |S_0| \leq \sum_{h=1}^{k_n} |S_h| \leq k_n s_n \ll p_n$.

Some discussion is in order for condition (A3). First, note that we can choose ε_0 in a way so that $\bar{\lambda}^{(0j)}_{hc} \geq \varepsilon_0$ for all h, c, and $j \in S_0$. Hence, (A3) implies a lower bound on the joint probability $\psi^{(0n)}$ in (10). Such a lower bound on a compactly supported target density is a standard assumption in Bayesian nonparametric theory; see for example, van der Vaart and van Zanten (2008). However, unlike univariate or multivariate density estimation in fixed dimensions where the density can be assumed to be bounded below by a constant, we need to precisely characterize the decay rate of the lower bound of the joint probability. Since $\psi^{(0n)}$ is a d^{q_n} probability tensor, $\min_{c_1, \ldots c_{s_n}} \psi^{(0n)}_{c_1 \ldots c_{s_n}} \leq (1/d)^{q_n} = \exp(-s_n k_n \log d)$. Assumption (A3) implies that

$$\min_{c_1, \ldots c_{s_n}} \psi^{(0n)}_{c_1 \ldots c_{s_n}} \geq \exp(-q_n \log(1/\varepsilon_0)) = \exp(-c_0 s_n) \tag{11}$$

for some constant $c_0 > 0$.

3.3 Main Result

We are now in a position to state a theorem on posterior convergence rates.

Theorem 3.1. Assume the true sequence of tensors $\pi^{(0n)} \in \mathcal{F}_n$ satisfy assumptions (A0) – (A3) and $s_n \log p_n / n \to 0$. Also, assume the sp-PARAFAC model is fitted with the stick-breaking prior truncated to k_n many components and $\gamma = \kappa p_n^2$ for some constant $\kappa \in (0, 1)$ in (5). Then, (8) is satisfied with $\epsilon_n = \sqrt{s_n \log p_n / n}$ in (7).

A proof of Theorem 3.1 can be found in Appendix B. As an implication of Theorem 3.1, if $p_n = n^d$ for some constant d, then the posterior contracts at the near parametric rate $\sqrt{(\log n)^c / n}$ for some constant $c > 0$. Moreover, consistent estimation is possible even if p_n is exponentially large as long as $p_n \leq \exp(\sqrt{n})$.

[3]For sequences a_n, b_n, we write $a_n = o(b_n)$ if $a_n / b_n \to 0$ as $n \to \infty$ and $a_n = O(b_n)$ if $a_n \leq C b_n$ for all large n.

In particular, with $p_n = \exp(n^{\delta/2})$ for $\delta < 1$, the posterior contracts at least at the rate $n^{-(1-\delta)/2}$.

Remark 3.1. We assume the number of components k_n known in Theorem 3.1 for ease of exposition, with our main focus on dimensionality reduction. Adapting to an unknown number of components in mixture models is a well-studied problem; see, for example, Ge and Jiang (2006), Pati, Dunson, and Tokdar (2013), and Shen, Tokdar, and Ghosal (2013). For the infinite stick-breaking prior on the mixture components, one can use the sieving technique developed in Pati, Dunson, and Tokdar (2013) to estimate deviation bounds for the tail sum of a stick-breaking process.

Remark 3.2. We need $\gamma = \kappa p_n^2$ in Theorem 3.1 to assure that the probability mass function of the induced beta-Bernoulli prior on $|S_h|$ satisfies $\mathbb{P}(|S_h| = s) \asymp e^{-Cs \log p_n}$ for small s; refer to the proof of Theorem 3.1 for further details. Classes of priors on the model size proportional to $e^{-Cs \log p_n}$ are referred to as complexity priors in Castillo and van der Vaart (2012) and commonly used in high-dimensional Bayesian asymptotics, where the prior probability of a particular model size s is inversely proportional to the $\binom{p_n}{s}$ many models of size s. See also Section 2.1 of Arias-Castro and Lounici (2012) for an usage of a similar prior in a high-dimensional regression context.

The choice $\gamma = \kappa p_n$ leads to $\mathbb{P}(|S_h| = s)$ behaving like e^{-Cs} for small s, which is not sufficient given the current proof technique. However, for numerical stability, we recommend the choice $\gamma = \kappa p_n$ in all practical applications involving large p_n, with $\kappa = 0.2$ used as a default choice in all our examples.

4. POSTERIOR COMPUTATION

Under model (5), we can easily proceed to draw posterior samples from a Gibbs sampler since all the full conditionals have recognizable forms. We integrate out the ζ_{jh}s to obtain $\lambda_h^{(j)} \sim (1-\tau_h)\delta_{\lambda_0^{(j)}} + \tau_h \text{Diri}(a_{j1}, \ldots, a_{jd_j})$ and therefore do not update the ζ_{jh}s. The algorithm iterates through the following steps:

1. For variable $j = 1, \ldots, p$ and latent class $h = 1, \ldots, k^*$, where $k^* = \max\{z_1, \ldots, z_n\}$, update $\lambda_h^{(j)} \equiv (\lambda_{h1}^{(j)}, \ldots, \lambda_{hd_j}^{(j)})$ from a two component mixture distribution, having a point mass at the baseline probability:
$$\begin{aligned}\left(\lambda_h^{(j)} \middle| -\right) = w_{0h}^{(j)} \delta_{\lambda_0^{(j)}} &+ w_{1h}^{(j)} \text{Diri}\left(a_{j1} + \sum_{i=1}^n 1(y_{ij} = 1, z_i = h),\right. \\ &\left. \ldots, a_{jd_j} + \sum_{i=1}^n 1(y_{ij} = d_j, z_i = h)\right), \quad (12)\end{aligned}$$
where $w_{0h}^{(j)}$ and $w_{1h}^{(j)}$ are the mixture weights:
$$\begin{aligned} w_{0h}^{(j)} &= \frac{(1-\tau_h)\prod_{c=1}^{d_j} \lambda_{0c}^{(j)\sum_{i=1}^n 1(z_i=h, y_{ij}=c)}}{(1-\tau_h)\prod_{c=1}^{d_j} \lambda_{0c}^{(j)\sum_{i=1}^n 1(z_i=h, y_{ij}=c)} + \tau_h \frac{\Gamma(\sum_{c=1}^{d_j} a_{jc})}{\prod_{c=1}^{d_j}\Gamma(a_{jc})} \cdot \frac{\prod_{c=1}^{d_j}\Gamma\left(a_{jc} + \sum_{i=1}^n 1(z_i=h, y_{ij}=c)\right)}{\Gamma\left(\sum_{c=1}^{d_j} a_{jc} + \sum_{i=1}^n 1(z_i=h)\right)}}, \\ w_{1h}^{(j)} &= 1 - w_{0h}^{(j)}. \end{aligned}$$
2. Let $\eta_{hj} \in \{0, 1\}$ be a binary allocation variable indicating the component $\lambda_h^{(j)}$ is drawn from in (12), with $\eta_{hj} = 0$ if $\lambda_h^{(j)}$ is updated from the baseline component. Update τ_h, $h = 1, \ldots, k^*$ from a Beta full conditional:
$$\tau_h | - \sim \text{Beta}\left(1 + \sum_{j=1}^p 1(\eta_{hj} = 1), \gamma + \sum_{j=1}^p 1(\eta_{hj} = 0)\right). \quad (13)$$
3. The full conditional of V_h, $h = 1, \ldots, k^*$ only requires the updated information on latent class allocation for all subjects:
$$V_h | - \sim \text{Beta}\left(1 + \sum_{i=1}^n 1(z_i = h), \alpha + \sum_{i=1}^n 1(z_i > h)\right). \quad (14)$$
4. Sample z_i, $i = 1, \ldots, n$ from the multinomial full conditional with
$$\Pr(z_i = h | -) = \frac{\nu_h \prod_{j=1}^p \lambda_{hy_{ij}}^{(j)}}{\sum_{l=1}^{k^*} \nu_l \prod_{j=1}^p \lambda_{ly_{ij}}^{(j)}}, \quad (15)$$
where $\nu_h = V_h \prod_{l<h}(1 - V_l)$.
5. Update α from the Gamma full conditional:
$$\alpha | - \sim \text{Gamma}\left(a_\alpha + k^*, b_\alpha - \sum_{h=1}^{k^*} \log(1 - V_h)\right). \quad (16)$$

These steps are simple to implement and we gain efficiency by updating the parameters in blocks. For example, instead of updating $\lambda_h^{(j)}$ one at a time, we sample $\boldsymbol{\lambda} \equiv \{\lambda_h^{(j)}, h = 1, \ldots, k^*, j = 1, \ldots, p\}$ jointly with corresponding parameters in matrix form. In all our examples, we ran the chain for 25,000 iterations, discarding the first 10,000 iterations as burn-in and collecting every fifth sample post burn-in to thin the chain. Mixing and convergence were satisfactory based on the examination of trace plots and the run time scaled linearly with n and p. We also carried out sensitivity analysis by multiplying and dividing the hyperparamaters a_α, b_α , and γ in (5) by a factor of 2, with the conclusions remained unchanged from the default setting $a_\alpha = b_\alpha = 1$ and $\gamma = 0.2p$.

5. SIMULATION STUDIES

5.1 Estimating Sparse Interactions

We first conduct a replicated simulation study to assess the estimation of sparse interactions using the proposed sp-PARAFAC model. We simulated 100 dependent binary variables $y_{ij} \in \{0, 1\}$, $j = 1, \ldots, p = 100$ $(d_j = d = 2)$ for $i = 1, \ldots, n = 100$ subjects from a log-linear model having up to three-way interactions:
$$\log\left(\frac{\pi_{c_1 \ldots c_p}}{\pi_{0\ldots0}}\right) = \sum_{s=1}^3 \sum_{S \subset \{1,\ldots,p\}:|S|=s} \beta_S 1_{(c_S=1)}. \quad (17)$$
For example, if $S = \{1, 2, 4\}$, then $\beta_S = \beta_{1,2,4}$ and $1_{(c_S=1)} = 1_{(c_1=1, c_2=1, c_4=1)}$ with $1_{(\cdot)}$ denoting the indicator function. To mimic the situation where only a few interactions are present,

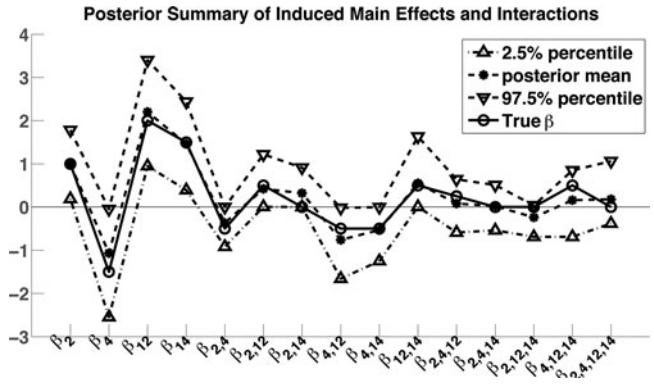

Figure 4. Posterior means and 95% credible intervals for all main effects and interactions in S^* compared with the true coefficients.

we restrict to $S \subset S^* = \{2, 4, 12, 14\}$ and set all interactions except

$$\boldsymbol{\beta} = (\beta_2, \beta_4, \beta_{12}, \beta_{14}, \beta_{2,4}, \beta_{2,12}, \beta_{4,12}, \beta_{4,14}, \beta_{12,14}, \times \beta_{2,4,12}, \beta_{4,12,14})^{\mathrm{T}}$$

to zero. This data-generating mechanism induces dependence among the variables in S^*, while rendering the other variables to be marginally independent. Figure 4 reports the posterior means and 95% credible intervals for all main effects and interactions for the variables in S^* averaged across 100 simulation replicates along with the true coefficients. As illustrated in Figure 4, averaging across the simulation replicates and different parameters, the 95% credible intervals cover the true parameter values 80% of the time.

Next, we study performance in estimating the dependence structure. Cramer's V is a popular statistic measuring the strength of association or dependence between two (nominal) categorical variables in a contingency table, ranging from 0 (no association) to 1 (perfect association). Let $\rho_{jj'}$ denote the Cramer's V statistics for variables j and j', so that

$$\rho^2_{jj'} = \frac{1}{\min\{d_j, d_{j'}\} - 1} \sum_{c_j=1}^{d_j} \sum_{c_{j'}=1}^{d_{j'}} \frac{\left(\pi^{(jj')}_{c_j c_{j'}} - \pi^{(j)}_{c_j} \pi^{(j')}_{c_{j'}}\right)^2}{\pi^{(j)}_{c_j} \pi^{(j')}_{c_{(j')}}}, \quad (18)$$

where $\pi^{(jj')}_{ll'} = \Pr(y_{ij} = l, y_{ij'} = l')$ and $\pi^{(j)}_{l} = \Pr(y_{ij} = l)$. Under the log-linear model (17), $\rho = (\rho_{jj'})$ is a sparse matrix with the Cramer's V for all pairs except those in $S^* \times S^*$ being zero. This is an immediate consequence of the fact that if $(j, j') \notin S^* \times S^*$, then y_{ij} and $y_{ij'}$ are independent.

We compare estimation of the off-diagonal entries of ρ under the sp-PARAFAC model with the empirical Cramer's V matrix $\hat{\rho}$. We can clearly convert posterior samples for the model parameters to posterior samples for $\rho_{jj'}$ through (18). The empirical estimator is obtained by replacing $\pi^{(jj')}_{c_j c_{j'}}$ and $\pi^{(j)}_{c_j}$ by their empirical estimators. The left panel in Figure 5 shows the posterior summaries (averaged across simulation replicates) of the Cramer's V values for all possible dependent pairs along with the true Cramer's V values (which can be calculated from (17)). In the right panel of Figure 5, we overlay kernel density estimators of posterior samples (in gray) and the empirical estimators (in red) of the Cramer's V values for all null pairs across all simulation replicates. Note the axes are also marked in gray and red for the respective cases. The sp-PARAFAC method clearly outperforms the empirical estimator convincingly, with the posterior density for the null pairs highly concentrated near zero while the empirical estimator has a mean Cramer's V value of 0.08 across the null pairs.

Furthermore, we can obtain power and Type I error rates for the nonnull and null variables respectively by computing the percentage of detected significance over the simulation replicates, with a coefficient declared significant if the 95% credible interval does not contain zero. Focusing on the power and Type I error of the main effects and interactions in S^*, most of the error rates are appealing barring a few cases (see Tables 2 and 3). It is not surprising that the approach may face difficulty assessing the exact interaction structure among a set of associated variables based on limited data. Further, given the Cramer's V results in the right panel of Figure 5, the Type I error for any variable not in S^* should be very small or zero. As an example, we tested the main effects and all possible interactions for positions 20, 30, 40, and 50. The Type I error rates are zero for all of them.

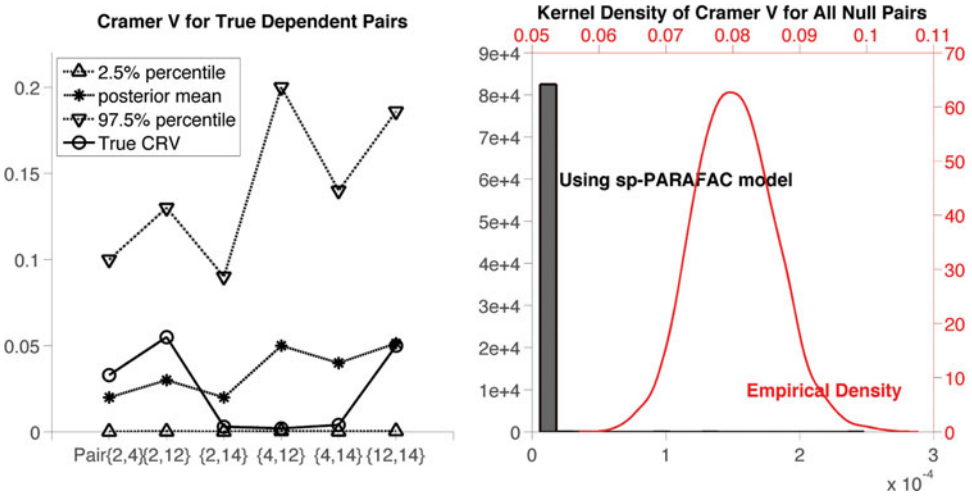

Figure 5. Left: Posterior summaries of the Cramer's V values for all dependent pairs versus the true Cramer's V values; Right: Estimated density of Cramer's V combining all null pairs under sp-PARAFAC vs. empirical estimation.

Table 2. Power for non-null variables based on 100 simulations

	β_2	β_4	β_{12}	β_{14}	$\beta_{2,4}$	$\beta_{2,12}$	$\beta_{4,12}$	$\beta_{4,14}$	$\beta_{12,14}$	$\beta_{2,4,12}$	$\beta_{4,12,14}$
Power	0.97	0.9	1	1	0.95	0.99	0.98	0.97	0.99	0	0
True coefficient	1	−1.5	2	1.5	−0.5	0.5	−0.5	−0.5	0.5	0.25	0.5

5.2 Comparison with Standard PARAFAC

We now conduct a simulation study to compare estimation of the Cramer's V matrix ρ under the proposed approach to the standard PARAFAC model in (3). We considered 100 simulation replicates, with data in each replicate consisting of $p = 100$ categorical variables for $n = 100$ subjects, with each variable having four possible levels ($d_j = d = 4$). Two simulation settings were considered to induce dependence between the variables in $S^* = \{2, 4, 12, 14\}$: (i) via multiple subpopulations as in the simulation study in Dunson and Xing (2009), and (ii) via a nominal GLM model $\Pr(y_{ij} = c) = \frac{\exp(\boldsymbol{y}_{i(j)}\boldsymbol{\beta}_c)}{1+\sum_{c=2}^{4}\exp(\boldsymbol{y}_{i(j)}\boldsymbol{\beta}_c)}$ for $j \in S^*$, where $\boldsymbol{y}_{i(j)}\boldsymbol{\beta}_c$ is a linear combination of all variables that are associated with the jth variable excluding the jth variable. The remaining variables were independently generated from a discrete uniform distribution.

The color plot on the left in Figure 6 shows the true pairwise Cramer's V values under simulation setting (i) (only the top-left 20×20 sub matrix of ρ is shown for clarity). Figure 6 (right) and Figure 7 represent one of the replicates, in which the right plot in Figure 6 shows the Cramer's V under the standard non-sparse PARAFAC method, while Figure 7 shows the Cramer's V using our method with the two different choices (i) and (ii) of the baseline components. It is obvious that our approach has much better estimates for not only the true dependent pairs but also the true nulls. Results for simulation (ii) shown in Figure 8 again show superiority of our sparse improvement to PARAFAC.

6. APPLICATION

6.1 Splice-Junction Gene Sequences

We applied the method to the splice-junction gene sequences, abbreviated as splice data below. The dataset is publicly available at the UCI machine learning repository. Splice junctions are points on a DNA sequence at which "superfluous" DNA is removed during the process of protein creation in higher organisms. These data consist of A, C, G, T nucleotides at $p = 60$ positions for $N = 3175$ sequences. Since the sample size is much larger than the number of variables, we compared our approach with the standard PARAFAC in two scenarios, first a small randomly selected subset (of size $n = 2p = 120$) of the full dataset, and second, the full dataset itself. Using two different sample sizes in this manner allows for a study of the new and existing methods and a comparison to a gold standard (a sufficiently large dataset). We ran the analysis to estimate the pairwise positional dependence structure under the standard PARAFAC method and the proposed approach with discrete uniform baseline component. As is apparent in Figure 10, both methods have similar performance when $n \gg p$. However, when the sample size is modest compared to the dimensionality, Figure 9 clearly demonstrates the advantage of our proposed method in identifying the dependence structure and pushing the independent pairs to zero, thereby obtaining a closer approximation to the gold standard (Figure 10).

6.2 The Public Use Microdata Sample (PUMS)

The PUMS data contains a sample of actual responses to the American Community Survey (available at *http://www2.census.gov/acs2010_1yr/pums/csv_pnc.zip*). The dataset includes behavioral, sociodemographic, and sociological variables in which 44 categorical variables are derived from the original survey data. There are 38,549 valid subjects without missing values. We used a similar strategy to that used for

Table 3. Type I error for null variables based on 100 simulations

	$\beta_{2,14}$	$\beta_{2,4,14}$	$\beta_{2,12,14}$	$\beta_{2,4,12,14}$
Type I error	0.97	0	0.68	0
True coefficient	0	0	0	0

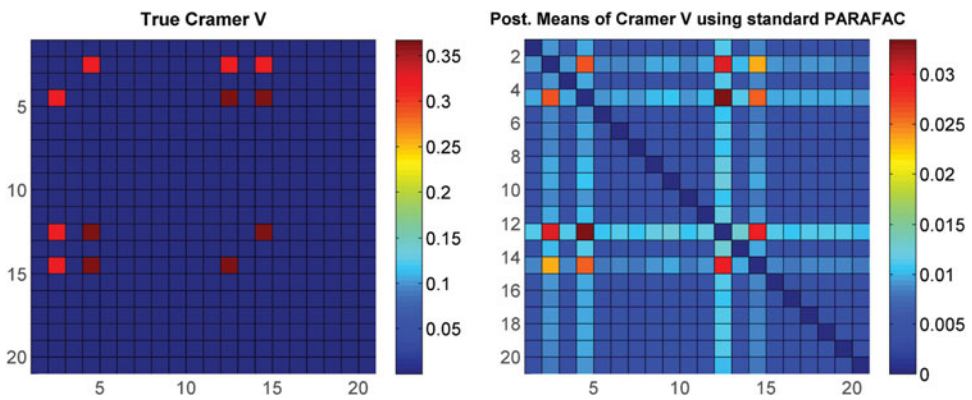

Figure 6. Simulation setting (i)—Left: True Cramer's V matrix; Right: Posterior means of Cramer's V using standard PARAFAC. Top 20×20 submatrix shown.

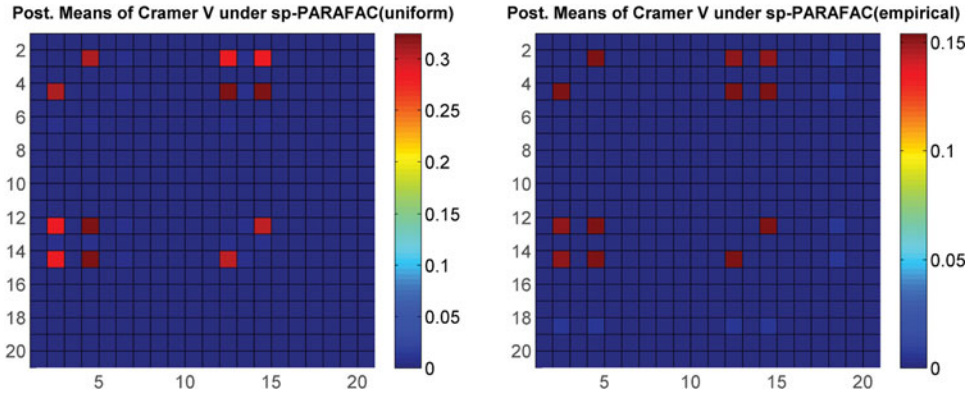

Figure 7. Posterior means of Cramer's V under simulation setting (i) using proposed method—Left: with $\lambda_0^{(j)}$ being discrete uniform; Right: with $\lambda_0^{(j)}$ being empirical estimates of the marginal category probabilities. Top 20×20 submatrix shown.

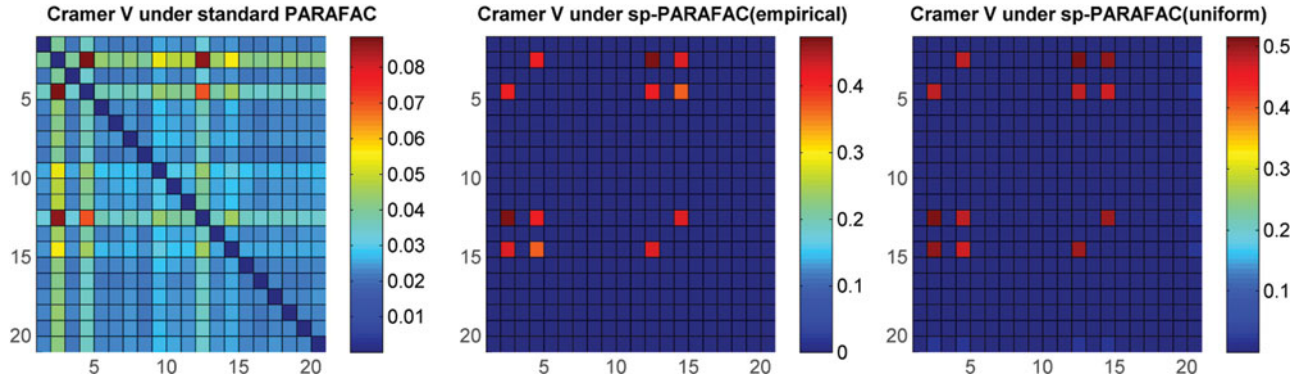

Figure 8. Posterior means of Cramer's V under simulation setting (ii)—Left: using standard PARAFAC; Middle: under proposed method using empirical marginal with Diri(1, . . . ,1) prior for λ_0; Right: using proposed method with discrete uniform λ_0. Top 20×20 submatrix shown.

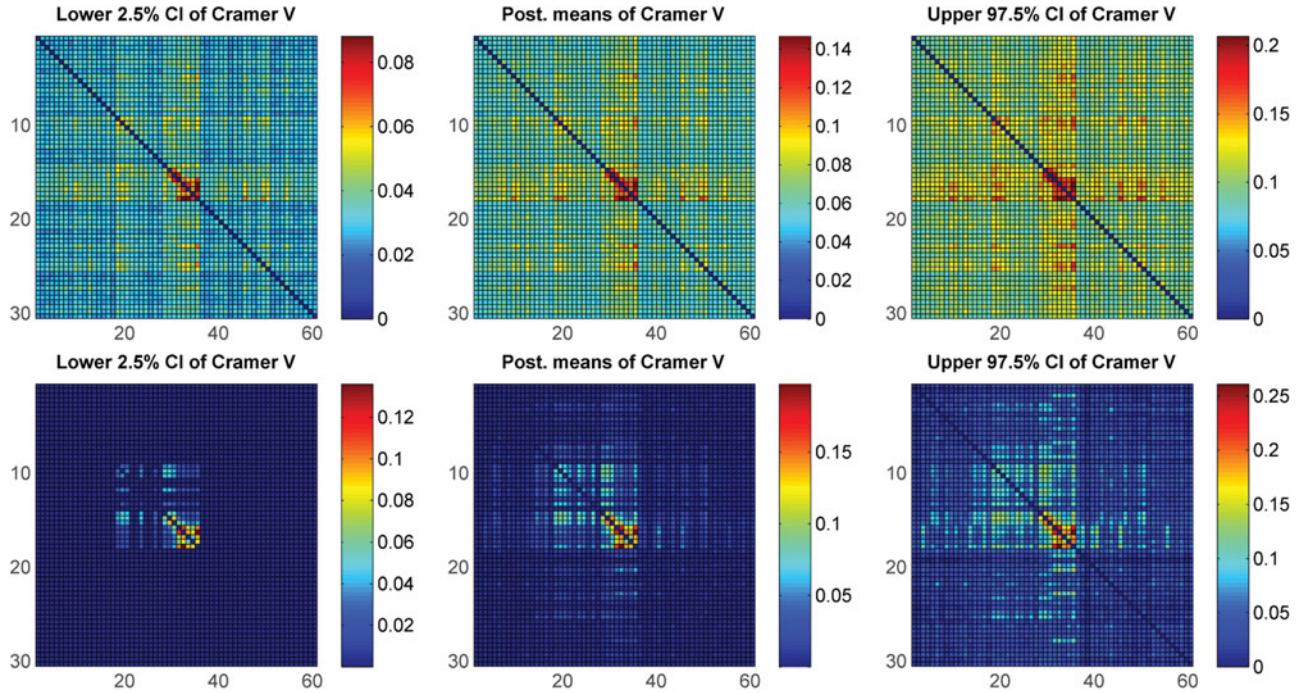

Figure 9. Posterior quantiles of Cramer's V with 120 sequences of splice data—Upper panel: under standard PARAFAC; Bottom panel: under proposed method.

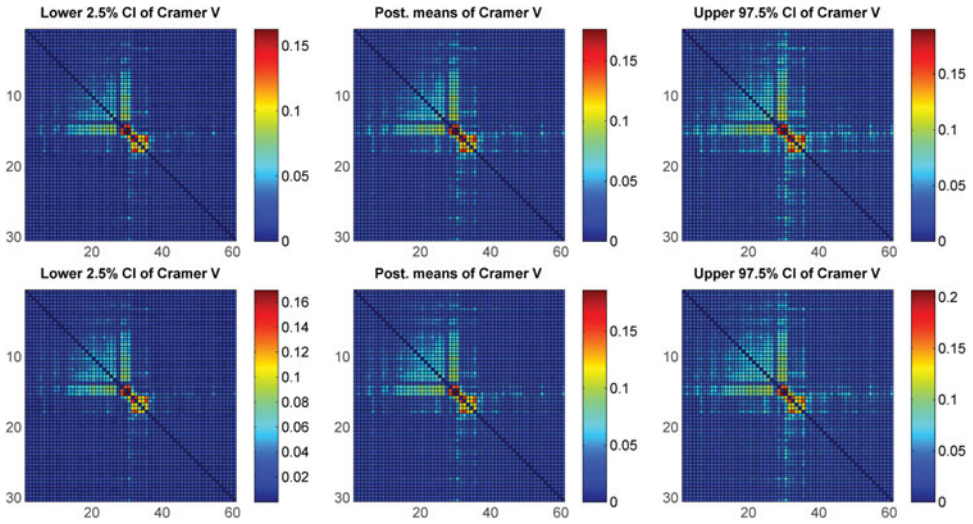

Figure 10. Posterior quantiles of Cramer's V with 3,175 sequences of splice data—Upper panel: under standard PARAFAC; Bottom panel: under proposed method.

the splice data to compare the performance with the standard PARAFAC method under a small sample case and a full sample case. A total of 100 subjects were first randomly selected to determine the association among the 44 social variables. Empirical marginal probabilities with a Dirichlet(1, . . . ,1) prior were used in our model, because we believe that the underlying independent variables are not following the discrete uniform distribution and we need to avoid the zero count problem in some categories. Comparing Figure 12 with Figure 11, the sp-FARAFAC again proves its advantage in detecting more true signals and shrinking the noise.

7. DISCUSSION

We have proposed a sparse modification to the widely-used PARAFAC tensor factorization, and have applied this in a Bayesian context to improve analyses of ultra sparse huge contingency tables. Given the compelling success in this application area, we hope that the proposed notion of sparsity will have a major impact in other areas, including tensor completion problems in machine learning. There is an enormous literature on low rank and sparse matrix factorizations, and the sp-PARAFAC should facilitate scaling of such approaches to many-way tables while dealing with the inevitable curse of dimensionality.

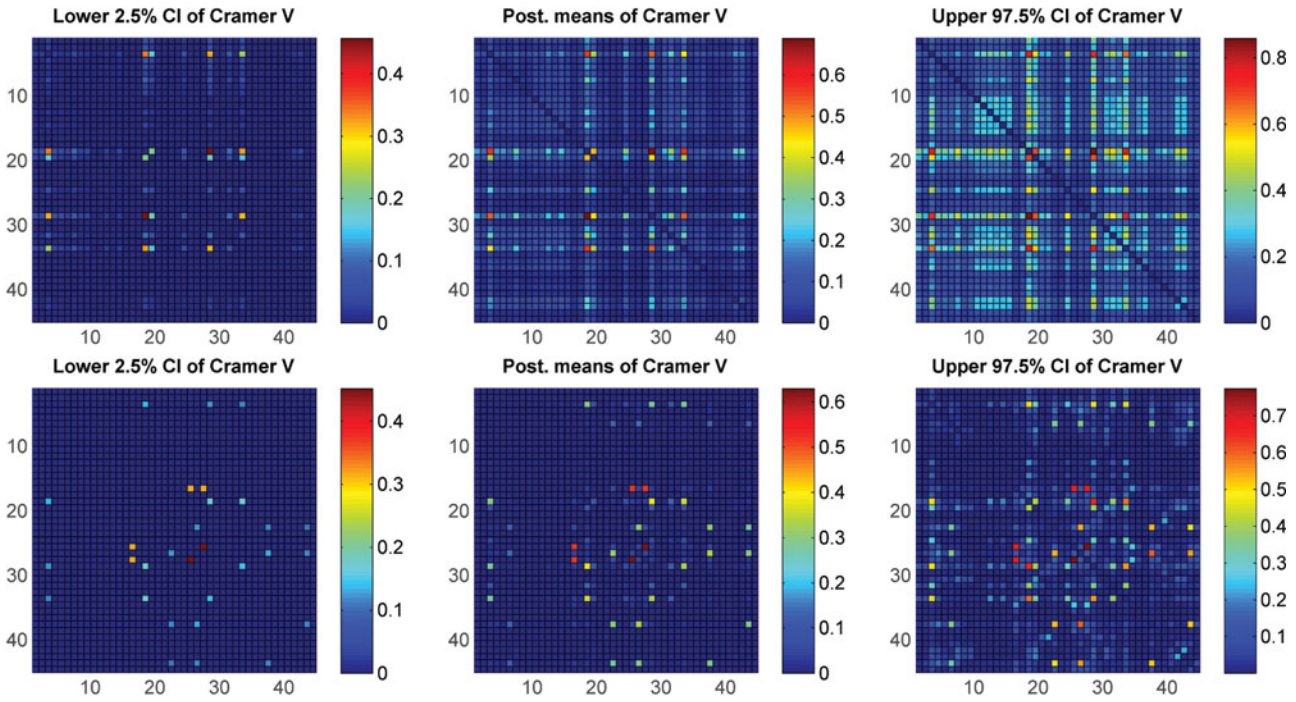

Figure 11. Posterior quantiles of Cramer's V with 100 subjects of PUMS – Upper panel: under standard PARAFAC; Bottom panel: under proposed method.

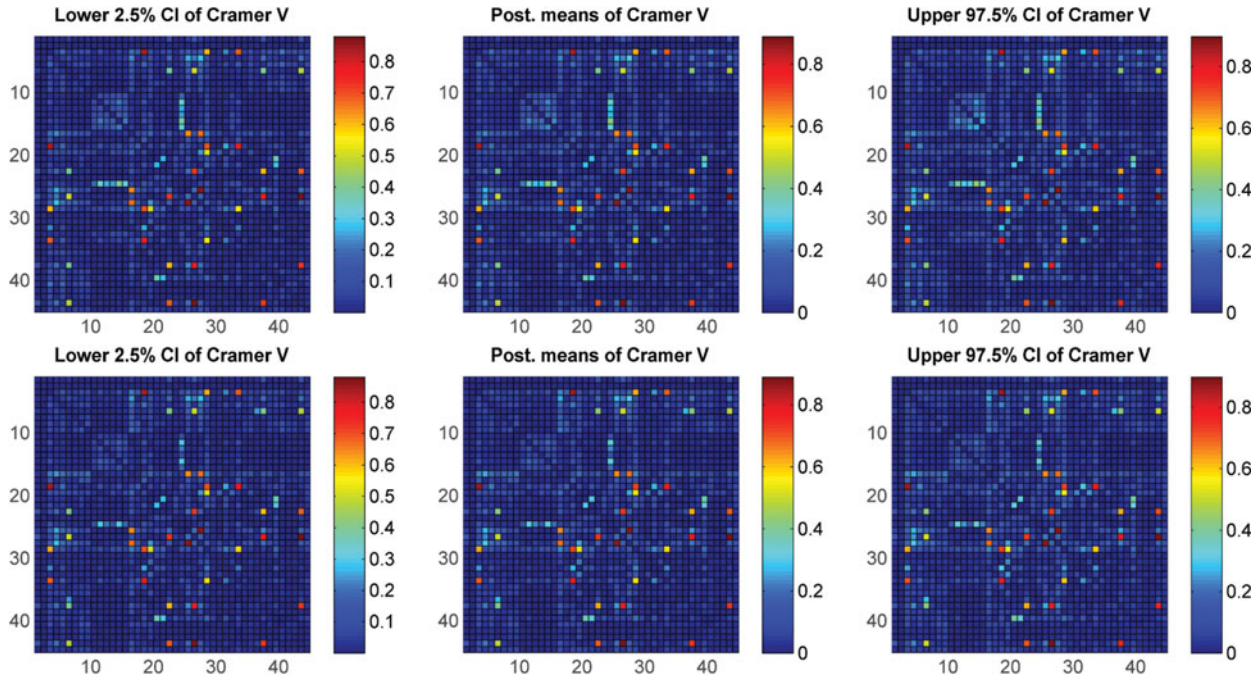

Figure 12. Posterior quantiles of Cramer's V with 38,549 subjects of PUMS—Upper panel: under standard PARAFAC; Bottom panel: under proposed method.

Although we take a Bayesian approach, we believe that frequentist penalized optimization methods can also exploit our same concept of sparsity in learning a compressed characterization of a huge array based on limited data.

APPENDIX A

Transformation to Log-Linear Parameterization

Let $V := \{1, \ldots, p\}$ denote the set of variables. Since we focus on binary variables, we can assume without loss of generality that $y_j \in \{1, 2\}$ for all $j \in V$. We summarize some basic facts regarding the log-linear parameterization of the joint probability tensor π of binary variables $y_1, \ldots, y_p$. For any $F \subseteq V$, let $c^{(F)}$ denote the cell with $c_j^{(F)} = 2$ if $j \in F$ and $c_j^{(F)} = 1$ if $j \in F^c$. We specifically denote the cell corresponding to $F = \emptyset$ by c^*, so that $c^* = (1, 1, \ldots, 1)$. In the corner parameterization (Massam, Liu, and Dobra 2009, Sec. 2),

$$\log \pi_{c^{(S)}} = \sum_{E \subseteq S} \beta_E,$$

for any $S \subseteq V$, where β_E denotes the interaction term corresponding to the variables in E. Clearly, in this parameterization, we have p main effects ($|E| = 1$), $\binom{p}{2}$ two-way interactions ($|E| = 2$) and so on. In this corner parameterization, the main effects and interactions can be recovered from the joint probability tensor as

$$\beta_S = \sum_{E \subseteq S} (-1)^{|S \setminus E|} \log \pi_{c^{(E)}}. \tag{A.1}$$

It follows from (A.1) that any main effect $\beta_j = \log\{\pi_{c^{(j)}}/\pi_{c^*}\}$ and any two-way interaction

$$\beta_{jj'} = \log\left\{\frac{\pi_{c^{(\{j,j'\})}}\pi_{c^*}}{\pi_{c^{(j)}}\pi_{c^{(j')}}}\right\}.$$

For example, in our $p = 3$ example in Section 2.2, $\beta_1 = \log\{\pi_{211}/\pi_{111}\}$ and $\beta_{12} = \log\{\frac{\pi_{221}\pi_{111}}{\pi_{121}\pi_{211}}\}$.

Proof of Proposition 2.1. We first calculate $\mathbb{P}(\beta_j = 0)$ for $j \in V$. Since the induced prior on the main effects are exchangeable, the probabilities are the same for all $j \in V$ and it suffices to calculate $\mathbb{P}(\beta_1 = 0)$. From the previous subsection, we have $\beta_1 = \log(\pi_{c^{(\{1\})}}) - \log(\pi_{c^*})$, where $c^{(\{1\})}$ is the cell $(2, 1, \ldots, 1)$. Clearly, the cells $c^{(\{1\})}$ and c^* only differ in the first coordinate corresponding to the first variable. Recalling that the stick-breaking prior is assumed to be truncated to K components, we have from (5) that

$$\pi_{c^{(\{1\})}} = \pi_{21\ldots1} = \sum_{h=1}^{K} \eta_h \lambda_{h2}^{(1)},$$
$$\pi_{c^*} = \pi_{11\ldots1} = \sum_{h=1}^{K} \eta_h \lambda_{h1}^{(1)},$$

where $\eta_h = \nu_h \prod_{j \neq 1} \lambda_{h1}^{(j)}$ is the same in both expressions since $c^{(\{1\})}$ and c^* only differ in the first coordinate. From the above display, it is clear that $\beta_1 = 0$ (with positive probability) if and only if variable 1 is assigned to the baseline group for all $h = 1, \ldots, K$, whence $\lambda_{h1}^{(1)} = \lambda_{01}^{(1)} = 1/2$ and $\lambda_{h2}^{(1)} = \lambda_{02}^{(1)} = 1/2$. Letting $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$, we therefore have $\mathbb{P}(\beta_1 = 0 \mid \boldsymbol{\tau}) = \prod_{h=1}^{K}(1 - \tau_h)$, since given $\boldsymbol{\tau}$, each variable is assigned to the baseline group inside the hth component with probability $(1 - \tau_h)$. The unconditional probability $\mathbb{P}(\beta_1 = 0) = \mathbb{E}\mathbb{P}(\beta_1 = 0 \mid \boldsymbol{\tau}) = \{\gamma/(1 + \gamma)\}^K$, since τ_h's are independent Beta$(1, \gamma)$. □

We next calculate $\mathbb{P}(\beta_{jj'} = 0)$ for $j \neq j' \in V$. Using the exchangeability argument, it is enough to calculate $\mathbb{P}(\beta_{12} = 0)$. We have from (A.1) that $\beta_{12} = \log(\pi_{c^{(\{1,2\})}}) - \log(\pi_{c^{(\{1\})}}) - \log(\pi_{c^{(\{2\})}}) + \log(\pi_{c^*})$. We can write

$$\pi_{c^{(\{1,2\})}} = \pi_{221\ldots1} = \sum_{h=1}^{K} \eta'_h \lambda_{h2}^{(1)} \lambda_{h2}^{(2)},$$
$$\pi_{c^*} = \pi_{111\ldots1} = \sum_{h=1}^{K} \eta'_h \lambda_{h1}^{(1)} \lambda_{h1}^{(2)}$$

$$\pi_{c^{([1])}} = \pi_{211\ldots1} = \sum_{h=1}^{K} \eta'_h \lambda_{h2}^{(1)} \lambda_{h1}^{(2)},$$
$$\pi_{c^{([2])}} = \pi_{121\ldots1} = \sum_{h=1}^{K} \eta'_h \lambda_{h1}^{(1)} \lambda_{h2}^{(2)},$$

where $\eta'_h = \nu_h \prod_{j\neq 1,2} \lambda_{h1}^{(j)}$. We claim that $\mathbb{P}(\beta_{12} = 0) = \mathbb{P}(A_1 \cup A_2)$, where A_l, $l = 1, 2$ denotes the event that variable l is assigned to the baseline group for all $h = 1, \ldots, K$. The identity follows since we can write (a) $\beta_{12} = [\log(\pi_{c^{([1,2])}}) - \log(\pi_{c^{([2])}})] - [\log(\pi_{c^{([1])}}) - \log(\pi_{c^*})]$ or (b) $\beta_{12} = [\log(\pi_{c^{([1,2])}}) - \log(\pi_{c^{([1])}})] - [\log(\pi_{c^{([2])}}) - \log(\pi_{c^*})]$. If A_1 holds, both $[\log(\pi_{c^{([1,2])}}) - \log(\pi_{c^{([2])}})]$ and $[\log(\pi_{c^{([1])}}) - \log(\pi_{c^*})]$ are zero, while if A_2 holds, both $[\log(\pi_{c^{([1,2])}}) - \log(\pi_{c^{([1])}})]$ and $[\log(\pi_{c^{([2])}}) - \log(\pi_{c^*})]$ are zero. We have $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \{\gamma/(1+\gamma)\}^K$ from the first part, and $\mathbb{P}(A_1 \cap A_2) = \mathbb{E}\prod_{h=1}^{K}(1-\tau_h)^2 = \{\gamma/(2+\gamma)\}^K$.

It follows from the above proof that $\mathbb{P}(\beta_1 = 0) = \mathbb{P}(A_1)$, while $\mathbb{P}(\beta_{12} = 0) = \mathbb{P}(A_1 \cup A_2)$, which clearly implies that $\mathbb{P}(\beta_1 = 0) \leq \mathbb{P}(\beta_{12} = 0)$, that is, the two-way interactions are shrunk more heavily compared to the main effects. The proof also clearly suggests that this phenomena extends to any order of interaction.

APPENDIX B

Proof of Theorem 3.1. We verify the conditions of Theorem 4 in Yang and Dunson (2013), which is a minor modification of Theorem 2.1 appearing in Ghosal, Ghosh, and van der Vaart (2000). Let $\epsilon_n \to 0$ be such that $n\epsilon_n^2 \to \infty$ and $\sum_{n\geq} \exp(-n\epsilon_n^2) \leq \infty$. Suppose there exist a sequence of sets $\mathcal{P}_n \subset \mathcal{F}_n$ and a constant $C > 0$ such that the following hold: [4]

1. $\log N(\epsilon_n; \mathcal{P}_n, \|\cdot\|_1) \leq n\epsilon_n^2$;
2. $\mathbb{P}_n(\mathcal{F}_n \cap \mathcal{P}_n^c) \leq \exp\{-(2+C)n\epsilon_n^2\}$; and
3. $\mathbb{P}_n(\pi : \|\log \frac{\pi}{\pi^{(0n)}}\|_\infty \leq \epsilon_n^2) \geq \exp(-Cn\epsilon_n^2)$.

Then, the posterior contracts at the rate ϵ_n, that is, (8) is satisfied. We now proceed to verify conditions (1)–(3). We define,

$$\mathcal{P}_n = \left\{ \pi \in \mathcal{F}_n : \pi_{c_1\ldots c_p} = \sum_{h=1}^{k_n} \nu_h^* \prod_{j\in S_h^*} \lambda_{hc_j}^{(*j)} \prod_{j\in S_h^{*c}} \lambda_{0c_j}^{(j)};\ \nu \in \mathcal{S}^{(k_n-1)},\ |S_h^*| \leq As_n, h = 1, \ldots, k_n \right\}, \tag{B1}$$

where $\mathcal{S}^{(r-1)}$ denotes the $(r-1)$-dimensional probability simplex and $A > 0$ is an absolute constant. We shall use C to denote an absolute constant whose meaning may change from one line to the next.

To estimate $N(\epsilon_n; \mathcal{P}_n, \|\cdot\|_1)$, we make use of the following Lemma, which follows in a straightforward manner by repeated uses of the triangle inequality.

Lemma B.1. Let $\pi^{(1)}, \pi^{(2)} \in \mathcal{F}_n$ with

$$\pi^{(i)} = \sum_{h=1}^{k_n} \nu_{ih} \lambda_{ih}^{(1)} \otimes \ldots \otimes \lambda_{ih}^{(p_n)},\ i = 1, 2.$$

Then,

$$\left\|\pi^{(1)} - \pi^{(2)}\right\|_1 \leq \sum_{h=1}^{k_n} |\nu_{1h} - \nu_{2h}| + \sum_{h=1}^{k_n} \nu_{2h} \left(\sum_{j=1}^{p_n} \sum_{c=1}^{d} |\lambda_{1hc}^{(j)} - \lambda_{2hc}^{(j)}| \right).$$

Lemma B.1 implies that if $\pi^{(1)}, \pi^{(2)} \in \mathcal{P}_n$ with $S_{1h}^* = S_{2h}^* = S_h^*$, then

$$\left\|\pi^{(1)} - \pi^{(2)}\right\|_1 \leq \sum_{h=1}^{k_n} |\nu_{1h} - \nu_{2h}| + \sum_{h=1}^{k_n} \nu_{2h} \left(\sum_{j\in S_h^*} \sum_{c=1}^{d} |\lambda_{1hc}^{(j)} - \lambda_{2hc}^{(j)}| \right).$$

Based on the above observation, we create an ϵ_n-net of $\mathcal{P}_n$ as follows: In (B.1), (i) vary S_h^* over all possible subsets of $\{1, \ldots, p_n\}$ with $|S_h^*| \leq As_n$ for $h = 1, \ldots, k_n$, (ii) for $h \in \{1, \ldots k_n\}$ and $j \in S_h^*$, vary $\lambda_h^{(*j)}$ over an $\epsilon_n/(2Ads_n)$-net of $\mathcal{S}^{(d-1)}$ and (iii) vary ν^* over an $\epsilon_n/(2k_n)$-net of $\mathcal{S}^{(k_n-1)}$.

For a fixed h, there are $\sum_{s=0}^{As_n} \binom{p}{s}$ subsets of size smaller then or equal to As_n. Using the inequality $\binom{p}{s} \leq (pe/s)^s$ for $s \leq p/2$, the number of possible subsets in (i) can be bounded above by $\exp(Ck_n s_n \log p_n)$. Hence,

$$\begin{aligned} &N(\epsilon_n; \mathcal{P}_n, \|\cdot\|_1) \\ &\quad\leq \exp(Ck_n s_n \log p_n)\, N(\epsilon_n/(2Ads_n); \mathcal{S}^{d-1}, \|\cdot\|_1)^{2Ads_n k_n} \\ &\qquad\times N(\epsilon_n/(2k_n); \mathcal{S}^{k_n-1}, \|\cdot\|_1). \end{aligned}$$

Using the fact that $N(\delta, \mathcal{S}^{r-1}, \|\cdot\|_1) \leq (C/\delta)^r$ (Vershynin 2010), the right-hand side in the above display can be bounded above by $\exp(Cs_n \log p_n) = \exp(n\epsilon_n^2)$, since $k_n = O(1)$.

We now bound $\mathbb{P}_n(\mathcal{F}_n \cap \mathcal{P}_n^c)$. Recall that in the sp-PARAFAC model, the induced prior on the subset size $|S_h|$ is $\text{Bin}(p_n, \tau_h)$, with $\tau_h \sim \text{Beta}(1, \gamma)$. Now,

$$\begin{aligned} \mathbb{P}_n(\mathcal{F}_n \cap \mathcal{P}_n^c) &\leq \Pr(\exists\, h \in \{1, \ldots, k_n\} \text{ s.t. } |S_h| \geq As_n) \\ &\leq k_n P(|S_1| > As_n). \end{aligned}$$

Integrating τ_1, the distribution of $|S_1|$ is a beta-Bernoulli distribution with probability mass function

$$\begin{aligned} \Pr(|S_1| = s) &= \binom{p}{s} \frac{1}{\text{B}(1,\gamma)} \int_{\tau=0}^{1} \tau^s (1-\tau)^{p_n - s} (1-\tau)^{\gamma-1} d\tau \\ &= \binom{p_n}{s} \frac{\text{B}(1+s, \gamma + p_n - s)}{\text{B}(1,\gamma)} \\ &= \frac{1}{\gamma} \frac{p_n!}{(p_n - s)!} \frac{(\gamma + p_n - s - 1)!}{(\gamma + p_n)!}, \end{aligned}$$

for $s = 0, 1, \ldots, p_n$. $\text{B}(\cdot,\cdot)$ denotes the Beta function in the above display. Hence, for $s \geq 1$,

$$\frac{\Pr(|S_1| = s)}{\Pr(|S_1| = s-1)} = \frac{(p_n - s + 1)}{(p_n - s + \gamma)}.$$

Now, letting $\gamma = p_n^2$, one has for any $p_n \geq 2$ and $1 \leq s \leq p_n/2$,

$$\frac{1}{4p_n} \leq \frac{(p_n - s + 1)}{(p_n - s + \gamma)} \leq \frac{1}{p_n}.$$

In general, for $\gamma = \kappa p_n^2$, we can bound this from both sides by C/p_n. Noting that $\Pr(|S_1| = 0) = C/p_n^3$, we have

$$\Pr(|S_1| = s) = \frac{C}{p_n^3} \prod_{j=1}^{s} \frac{\Pr(|S_1| = j)}{\Pr(|S_1| = j-1)},$$

implying there exists constants $c_1, c_2 > 0$ such that

$$e^{-c_1(s+3)\log p_n} \leq \Pr(|S_1| = s) \leq e^{-c_2(s+3)\log p_n}, \tag{B2}$$

for $0 \leq s \leq p_n/2$. In particular, the upper bound holds for all $0 \leq s \leq p_n$, since $(p_n - s + 1)/(p_n - s + \gamma) \leq C/p_n$ for all s. Hence, for n

[4] Given a metric space $(\mathcal{X}, d)$, let $N(\epsilon; \mathcal{X}, d)$ denote its ϵ-covering number, that is, the minimum number of d-balls of radius ϵ needed to cover $\mathcal{X}$.

large enough so that $s_n \geq 3$,

$$\mathrm{P}(|S_1| > As_n) \leq \sum_{j=As_n+1}^{p_n} \exp(-Cj \log p_n) \leq \exp(-Cs_n \log p_n) \leq \exp(-n\epsilon_n^2).$$

We finally show that (3) holds. Recall the decomposition of $\pi^{(0n)}$ from (9). A probability tensor π following a sp-PARAFAC model with a truncated stick-breaking prior on ν can be parameterized as

$$\theta_\pi = \left(\nu, \{S_h\}_{1\leq h\leq k_n}, \{\lambda_h^{(j)}\}_{1\leq h\leq k_n, j\in S_h}\right),$$

where $\nu \in \mathcal{S}^{k_n-1}$, $S_h \subset \{1,\ldots,p_n\}$, $\lambda_h^{(j)} \in \mathcal{S}^{d-1}$. Consider the following subset $\mathcal{A}$ of the parameter space,

$$\mathcal{A} = \left\{ S_h = S_0, 1 \leq h \leq k_n; \sum_{h=1}^{k_n} |\nu_h - \nu_{0h}| \leq \frac{\epsilon_n^2}{2e^{c_0 s_n}}; \right. \\ \left. \times \sum_{c=1}^{d} \left|\lambda_{hc}^{(j)} - \bar{\lambda}_{hc}^{(0j)}\right| \leq \frac{\epsilon_n^2 \varepsilon_0}{4q_n}, 1 \leq h \leq k_n, j \in S_0 \right\}.$$

We now show that $\theta_\pi \in \mathcal{A}$ implies $\log \|\pi/\pi^{(0n)}\|_\infty \leq \epsilon_n^2$, so that $\mathbb{P}_n(\log \|\pi/\pi^{(0n)}\|_\infty \leq \epsilon_n^2)$ can be bounded below by $\mathbb{P}_n(\mathcal{A})$. First, observe that since $S_h = S_0$ for all h on $\mathcal{A}$, $\pi/\pi^{(0n)} = \psi/\psi^{(0n)}$, where $\psi^{(0n)}$ is as in (10) and ψ is the d^{q_n} joint probability tensor implied by the sp-PARAFAC model for the variables $\{y_{ij} : j \in S_0\}$,

$$\psi_{c_1\ldots c_{q_n}} = \sum_{h=1}^{k_n} \nu_h \prod_{j\in S_0} \lambda_{hc_j}^{(e_j)}.$$

Hence,

$$\log \left\| \frac{\pi}{\pi^{(0n)}} \right\|_\infty = \log \left\| \frac{\psi}{\psi^{(0n)}} \right\|_\infty \\ \leq \log\left(1 + \left\| \left(\frac{\psi}{\psi^{(0n)}} - 1\right) \right\|_\infty\right) \\ \leq \left\| \left(\frac{\psi}{\psi^{(0n)}} - 1\right) \right\|_\infty,$$

where the penultimate step follows from an application of triangle inequality and the last step uses $\log(1+x) \leq x$ for $x \geq 0$. For any $c_1,\ldots,c_{s_n}$, by an application of triangle inequality,

$$\left|\psi_{c_1\ldots c_{s_n}} - \psi^{(0n)}_{c_1\ldots c_{s_n}}\right| \leq \sum_{h=1}^{k_n} |\nu_h - \nu_{0h}| \\ + \sum_{h=1}^{k_n} \nu_{0h} \left| \prod_{j=1}^{q_n} \lambda_{hc_j}^{(e_j)} - \prod_{j=1}^{q_n} \bar{\lambda}_{hc_j}^{(0e_j)} \right|. \quad \text{(B3)}$$

We now state a Lemma to facilitate bounding the second term of the above display.

Lemma B.2. Let $v_1,\ldots v_r \in (\varepsilon_0, 1-\varepsilon_0)$ for some $\varepsilon_0 > 0$. Let $\delta > 0$ be such that $r\delta < \varepsilon_0/2$. Then, if $u_1,\ldots,u_r$ satisfy $|u_j - v_j| \leq \delta$ for all $j = 1,\ldots,r$, then

$$|u_1\ldots u_r - v_1\ldots v_r| \leq \frac{2r\delta}{\varepsilon_0} v_1\ldots v_r.$$

Apply Lemma B.2 with $r = q_n$, $u_j = \bar{\lambda}_{hc_j}^{(0e_j)}$ and $\delta = \epsilon_n^2\varepsilon_0/(4q_n)$ (clearly $r\delta/\varepsilon_0 = \epsilon_n^2/4 < 1/2$) to obtain that for any $1 \leq h \leq q_n$, $|\prod_{j=1}^{q_n} \lambda_{hc_j}^{(e_j)} - \prod_{j=1}^{q_n} \bar{\lambda}_{hc_j}^{(0e_j)}| \leq (\epsilon_n^2/2)\prod_{j=1}^{q_n} \bar{\lambda}_{hc_j}^{(0e_j)}$. Substituting this bound in (B.3), we have on $\mathcal{A}$,

$$\frac{|\psi_{c_1\ldots c_{s_n}} - \psi^{(0n)}_{c_1\ldots c_{s_n}}|}{\psi^{(0n)}_{c_1\ldots c_{s_n}}} \leq \frac{\sum_{h=1}^{k_n} |\nu_h - \nu_{0h}|}{e^{-c_0 s_n}} \\ + \left(\epsilon_n^2/2\right) \frac{\sum_{h=1}^{k_n} \nu_{0h} \prod_{j=1}^{q_n} \bar{\lambda}_{hc_j}^{(0e_j)}}{\psi^{(0n)}_{c_1\ldots c_{s_n}}} \leq \epsilon_n^2.$$

For the two terms in the above display after the first inequality, we used the lower bound (11) for the first term along with $\sum_{h=1}^{k_n} |\nu_h - \nu_{0h}| \leq \epsilon_n^2/(2e^{c_0 s_n})$ on $\mathcal{A}$, and by definition of $\psi^{(0n)}$, the second term is $\epsilon_n^2/2$.

It thus remains to lower bound $\mathbb{P}_n(\mathcal{A})$. By independence across h, $\Pr(S_h = S_0, 1 \leq h \leq k_n) = \Pr(S_1 = S_0)^{k_n}$. Further, by exchangeability of the prior on S_1, since all subsets of a particular size receive the same prior probability, $\Pr(S_1 = S_0) = \Pr(|S_1| = q_n)/\binom{p_n}{q_n}$. From (B.2), $\Pr(|S_1| = q_n \geq \exp(-Cs_n \log p_n)$. Using $\binom{p_n}{q_n} \leq (p_n e/q_n)^{q_n}$, we conclude that $\Pr(S_1 = S_0) \geq \exp(-Cs_n \log p_n)$.

Recall that $\nu_h = \nu_h^* \prod_{l<h}(1-\nu_l^*)$, where $\nu_l^* \sim \mathrm{Beta}(1,\alpha)$ independently. Find numbers $\{\nu_{0h}^*\}$ such that $\nu_{0h} = \nu_{0h}^* \prod_{l<h}(1-\nu_{0l}^*)$. It is easy to see that there exists a constant $C > 0$ such that $|\nu_h^* - \nu_{0h}^*| \leq \epsilon_n/(Ck_n)$ for all $h = 1,\ldots,k_n$ implies $\sum_{h=1}^{k_n} |\nu_h - \nu_{0h}| \leq \epsilon_n$. Hence, using a general result on small ball probability estimate of Dirichlet random vectors (Lemma 6.1 of Ghosal, Ghosh, and van der Vaart 2000), one has

$$\Pr\left(\sum_{h=1}^{k_n} |\nu_h - \nu_{0h}| \leq \frac{\epsilon_n^2}{2e^{c_0 s_n}}\right) \geq \exp\{-Cs_n \log(1/\epsilon_n)\}.$$

Another application of Lemma 6.1 of Ghosal, Ghosh, and van der Vaart (2000) yields,

$$\Pr\left(\sum_{c=1}^{d} |\lambda_{hc}^{(j)} - \bar{\lambda}_{hc}^{(0j)}| \leq \frac{\epsilon_n^2\varepsilon_0}{4q_n}\right) \geq \exp\{-C \log(s_n/\epsilon_n)\}.$$

Combining, we get $\Pr(\mathcal{A}) \geq \exp(-Cs_n \log p_n) \geq \exp(-n\epsilon_n^2)$. Hence, we have established (1) – (3), completing the proof.

Proof of Lemma B.2. Observe that

$$|u_1\ldots u_r - v_1\ldots v_r| \\ = |v_1\ldots v_r| \left| \frac{u_1\ldots u_r}{v_1\ldots v_r} - 1 \right| \\ = v_1\ldots v_r \max\left\{ \frac{u_1\ldots u_r}{v_1\ldots v_r} - 1, 1 - \frac{u_1\ldots u_r}{v_1\ldots v_r} \right\}.$$

Now, since $u_h \leq v_h + \delta$ for all h,

$$\frac{u_1\ldots u_r}{v_1\ldots v_r} \leq \prod_{h=1}^{r}(1 + \delta/v_h) \leq (1+\delta/\varepsilon_0)^r.$$

Using the binomial theorem, $(1+\delta/\varepsilon_0)^r - 1 = r\delta/\varepsilon_0 + \sum_{h=2}^{r}\binom{r}{h}(\delta/\varepsilon_0)^h$. Next, bound $\binom{r}{h} \leq r^h$ and use the fact that $r\delta/\varepsilon_0 < 1/2$ to conclude that $\sum_{h=2}^{r}\binom{r}{h}(\delta/\varepsilon_0)^h \leq \sum_{h=1}^{\infty}(r\delta/\varepsilon_0)^h \leq 2r\delta/\varepsilon_0$.

On the other hand, using $u_h \geq v_h - \delta$ for all h,

$$\frac{u_1\ldots u_r}{v_1\ldots v_r} \geq \prod_{h=1}^{r}(1 - \delta/v_h) \geq (1-\delta/\varepsilon_0)^r \geq 1 - r\delta/\varepsilon_0.$$

The proof is concluded by observing that

$$\max\left\{ \frac{u_1\ldots u_r}{v_1\ldots v_r} - 1, 1 - \frac{u_1\ldots u_r}{v_1\ldots v_r} \right\} \leq 2r\delta/\varepsilon_0.$$

□

[Received June 2013. Revised September 2014.]

REFERENCES

Agresti, A. (2002), *Categorical Data Analysis* (vol. 359), New York: Wiley-Interscience. [1563]

Arias-Castro, E., and Lounici, K. (2012), "Variable Selection With Exponential Weights and l_0-Penalization," *Electronic Journal of Statistics* 8, 328–354. [1568]

Armagan, A., Dunson, D., and Lee, J. (2013a), "Generalized Double Pareto Shrinkage," *Statistica Sinica*, 23, 119–143. [1564]

Armagan, A., Dunson, D.B., Lee, J., Bajwa, W.U., and Strawn, N. (2013b), "Posterior Consistency in Linear Models Under Shrinkage Priors," *Biometrika*, 100, 1011–1018. [1566]

Belitser, E., and Ghosal, S. (2003), "Adaptive Bayesian Inference on the Mean of an Infinite-Dimensional Normal Distribution," *The Annals of Statistics*, 31, 536–559. [1566]

Bhattacharya, A., and Dunson, D. (2011), "Sparse Bayesian Infinite Factor Models," *Biometrika*, 98, 291–306. [1563]

Bontemps, D. (2011), "Bernstein–von Mises Theorems for Gaussian Regression With Increasing Number of Regressors," *The Annals of Statistics*, 39, 2557–2584. [1566]

Bro, R. (1997), "PARAFAC. Tutorial and Applications," *Chemometrics and Intelligent Laboratory Systems*, 38, 149–171. [1562]

Carvalho, C., Lucas, J., Wang, Q., Nevins, J., and West, M. (2008), "High-Dimensional Sparse Factor Modelling: Applications in gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456. [1562,1563]

Carvalho, C., Polson, N., and Scott, J. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [1564]

Castillo, I., and van der Vaart, A. (2012), "Needles and Straws in a Haystack: Posterior Concentration for Possibly Sparse Sequences," *The Annals of Statistics*, 40, 2069–2101. [1566,1567,1568]

Dunson, D.B., and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *Journal of the American Statistical Association*, 104, 1042–1051. [1563,1570]

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. [1563]

Ge, Y., and Jiang, W. (2006), "On Consistency of Bayesian Inference With Mixtures of Logistic Regression," *Neural Computation*, 18, 224–243. [1568]

Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008), "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models," *Annals of Applied Statistics*, 2, 1360–1383. [1564]

Ghosal, S. (1999), "Asymptotic Normality of Posterior Distributions in High-Dimensional Linear Models," *Bernoulli*, 5, 315–331. [1566]

——— (2000), "Asymptotic Normality of Posterior Distributions for Exponential Families When the Number of Parameters Tends to Infinity," *Journal of Multivariate Analysis*, 74, 49–68. [1566]

Ghosal, S., Ghosh, J., and van der Vaart, A. (2000), "Convergence Rates of Posterior Distributions," *The Annals of Statistics*, 28, 500–531. [1566,1574,1575]

Hans, C. (2011), "Elastic Net Regression Modeling With the Orthant Normal Prior," *Journal of the American Statistical Association*, 106, 1383–1393. [1564]

Harshman, R. (1970), "Foundations of the PARAFAC Procedure: Models and Conditions for an 'Explanatory' Multi-Modal Factor Analysis," *UCLA Working Papers in Phonetics*, 16, 1–84. [1562]

Jiang, W. (2007), "Bayesian Variable Selection for High Dimensional Generalized Linear Models: Convergence Rates of the Fitted Densities," *The Annals of Statistics*, 35, 1487–1511. [1566]

Kolda, T., and Bader, B. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [1562]

Lim, L., and Comon, P. (2009), "Nonnegative Approximations of Nonnegative Tensors," *Journal of Chemometrics*, 23, 432–441. [1567]

Lopes, M. (2013), "Estimating Unknown Sparsity in Compressed Sensing," in *Proceedings of the 30th International Conference on Machine Learning* (ICML-13), pp. 217–225. [1565]

Lucas, J.E., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006), "Sparse Statistical Modelling in Gene Expression Genomics," in *Bayesian Inference for Gene Expression and Proteomics*, eds. K. Do, P. Müller, and M. Vannucci, Cambridge, UK: Cambridge University Press, pp. 155–176. [1563]

Massam, H., Liu, J., and Dobra, A. (2009), "A Conjugate Prior for Discrete Hierarchical Log-Linear Models," *The Annals of Statistics*, 37, 3431–3467. [1564,1573]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [1564]

Pati, D., Bhattacharya, A., Pillai, N.S., and Dunson, D.B. (2014), "Posterior Contraction in Sparse Bayesian Factor Models for Massive Covariance Matrices," *The Annals of Statistics*, 42, 1102–1130. [1563]

Pati, D., Dunson, D.B., and Tokdar, S.T. (2013), "Posterior Consistency in Conditional Distribution Estimation," *Journal of Multivariate Analysis*, 116, 456–472. [1568]

Polson, N., and Scott, J. (2010), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, New York: Oxford University Press, pp. 501–538. [1564]

Scott, J., and Berger, J. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *The Annals of Statistics*, 38, 2587–2619. [1562,1566]

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650. [1563]

Shen, W., Tokdar, S.T., and Ghosal, S. (2013), "Adaptive Bayesian Multivariate Density Estimation With Dirichlet Mixtures," *Biometrika*, 100, 623–640. [1568]

Talagrand, M. (1996), "A New Look at Independence," *The Annals of Probability*, 24, 1–34. [1564]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [1562]

van der Vaart, A., and van Zanten, J. (2008), "Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors," *The Annals of Statistics*, 36, 1435–1463. [1567]

Vershynin, R. (2010), "Introduction to the Non-Asymptotic Analysis of Random Matrices," *Arxiv preprint arxiv:1011.3027*, unpublished note. [1574]

West, M. (2003), "Bayesian Factor Regression Models in the 'Large p, Small n' Paradigm," in *Bayesian Statistics 7*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith. and M. West, New York: Oxford University Press, pp. 733–742. [1562,1563]

Yang, Y., and Dunson, D.B. (2013), "Bayesian Conditional Tensor Factorizations for High-dimensional Classification," *arXiv preprint arXiv:1301.4950*. [1574]