The Annals of Statistics
2017, Vol. 45, No. 1, 1–38
DOI: 10.1214/15-AOS1414

TENSOR DECOMPOSITIONS AND SPARSE LOG-LINEAR MODELS

BY JAMES E. JOHNDROW*, [1], ANIRBAN BHATTACHARYA†, [2]
AND DAVID B. DUNSON*, [1]

*Duke University** and *Texas A&M University*†

Contingency table analysis routinely relies on log-linear models, with latent structure analysis providing a common alternative. Latent structure models lead to a reduced rank tensor factorization of the probability mass function for multivariate categorical data, while log-linear models achieve dimensionality reduction through sparsity. Little is known about the relationship between these notions of dimensionality reduction in the two paradigms. We derive several results relating the support of a log-linear model to nonnegative ranks of the associated probability tensor. Motivated by these findings, we propose a new collapsed Tucker class of tensor decompositions, which bridge existing PARAFAC and Tucker decompositions, providing a more flexible framework for parsimoniously characterizing multivariate categorical data. Taking a Bayesian approach to inference, we illustrate empirical advantages of the new decompositions.

1. Introduction. Parsimonious models for contingency tables are of growing interest due to the routine collection of data on moderate to large numbers of categorical variables. We study the relationship between two paradigms for inference in contingency tables: the log-linear model [1, 4, 17] and latent structure models [2, 20, 21, 23, 32, 35, 42] that induce a tensor decomposition of the joint probability mass function [3, 15]. We aim to understand situations where the joint probability corresponding to a sparse log-linear model has a low rank tensor factorization. Connecting the seemingly distinct notions of parsimony in the two parameterizations can motivate the use of factorizations having a combination of computational tractability and flexibility.

Let $V = \{1, \dots, p\}$ denote a set of p categorical variables. We use $(y_j, j \in V)$ to denote variables, with $y_j \in \mathcal{I}_j$ having $d_j = |\mathcal{I}_j|$ levels. Without loss of generality, we assume $\mathcal{I}_j = \{1, \dots, d_j\}$. Let $\mathcal{I}_V = \times_{j \in V} \mathcal{I}_j$. Elements of $\mathcal{I}_V$ are referred to as cells of the contingency table; there are $\prod_{j=1}^{p} d_j$ cells in total. We generically denote a cell by $\mathbf{i}$, with $\mathbf{i} = (i_1, \dots, i_p) \in \mathcal{I}_V$. The joint probability mass function

Received April 2014; revised November 2015.

[1] Supported in part by Grants ES017436, ES020619 and ES017240 from the National Institute of Environmental Health Sciences (NIEHS) of the US National Institutes of Health.

[2] Supported by the Office of Naval Research (ONR BAA 14-0001).

MSC2010 subject classifications. 62F15.

Key words and phrases. Bayesian, categorical data, contingency table, latent class analysis, graphical model, high-dimensional, low rank, Parafac, Tucker, sparsity.

of $\mathbf{y} = (y_1, \dots, y_p)$ is denoted by π, with

$$\pi_{i_1,\dots,i_p} = \Pr(y_1 = i_1, \dots, y_p = i_p), \qquad \mathbf{i} \in \mathcal{I}_V. \tag{1}$$

A p-way tensor $M \in \mathbb{R}^{d_1 \times \cdots \times d_p}$ is a multiway-array which generalizes matrices to higher dimensions [29]. Two common forms of tensor decomposition which extend the matrix singular value decomposition are the PARAFAC [24] and Tucker [10, 11, 43] decompositions. Note that $\pi = (\pi_{i_1,\dots,i_p})_{\mathbf{i} \in \mathcal{I}_V}$ can be identified with a $\mathbb{R}^{d_1 \times \cdots \times d_p}$-*probability tensor*, which is a nonnegative tensor with entries summing to one. Given n i.i.d. replicates of $\mathbf{y}$, let $\mathbf{n}(\mathbf{i})$ denote the cell-count of cell $\mathbf{i}$. We assume the cell counts are multinomially distributed according to the probabilities in π.

Inference for contingency tables often employs log-linear models that express the logarithms of the entries in π as a linear function of parameters related to the index of each cell. Most of these parameters relate to interactions between the variables [1]. A saturated log-linear model has as many parameters as π has cells. To reduce dimensionality, it is common to assume a large subset of the interaction parameters are zero, and estimate the model using L_1 regularization [37, 38], decomposition approaches [6] or Bayesian model averaging [12, 13, 36]. Zero interaction terms are easily interpreted in terms of conditional and marginal independence relationships among the variables. A significant literature exists on Bayesian inference for log-linear models, focusing mainly on the development of novel conjugate priors [8, 36], model selection/averaging [25, 33] and stochastic search algorithms to explore the model space (e.g., [14]).

An alternative approach is to assume that the p variables are conditionally independent given one or more discrete latent class indices, with dependence induced upon marginalization over the latent variable(s). The attractiveness of such latent class models arises partly from easy model fitting using data-augmentation, with a Bayesian nonparametric formulation allowing the number of latent classes to be learned from the data [15]. Dunson and Xing [15] showed that a single latent class model is equivalent to a reduced-rank nonnegative PARAFAC decomposition of the joint probability tensor π, while the multiple latent class model in [3] implied a Tucker decomposition. See also [44] and [30] for extensions of these models to more complex settings.

Latent class models and log-linear models can be unified within a larger class of graphical models with observed and unobserved variables (see, e.g., [26, 31]). In particular, [19] describes relationships between the number of components in a PARAFAC expansion of π and the topological structure of the corresponding parameter space of a log-linear model, with consequences for estimation and selection in latent structure models. Others have established additional connections between latent structure models and the algebraic topology of the log-linear model [16, 18, 33, 39–41].

These two classes of models impose sparsity (or parsimony) in seemingly different ways, and to best of our knowledge, no connection has been established yet in this regard. The class of sparse log-linear models is often considered a desirable data generating class in high-dimensional settings for flexibility and ease of interpretation, and it is important to determine whether there exist low rank expansions for probability tensors corresponding to sparse log-linear models. Determining whether a nontrivial relationship exists is a major focus of the paper. Working with a class of weakly hierarchical log-linear models, we provide precise bounds on the tensor ranks of sparse log-linear models. There are limited results on ranks of higher-order tensors, and the techniques developed here may be of independent interest.

The complementary goal of this work is to leverage insights from our theoretical study to develop improved classes of factorization models that provide computationally tractable alternatives to sparse log-linear models. Sparse log-linear models are appealing in terms of interpretation and flexibility but unfortunately cannot be implemented practically in high dimensions. Motivated by our theoretical results that usual latent class models require many extra parameters to characterize sparse log-linear models, we propose a new class of collapsed Tucker (c-Tucker) factorizations. These factorizations can parsimoniously characterize complex interactions in categorical data, including data generated from sparse log-linear models. We propose Bayesian methods for analyzing data under c-Tucker models, demonstrating advantages over usual PARAFAC-type latent class models.

This paper is organized as follows. Section 2 introduces notation and provides background relevant to log-linear models and latent structure models. Section 3 provides our main theoretical results on the rank of probability tensors corresponding to sparse log-linear models, and defines classes of sparse log-linear models corresponding to relatively low rank probability tensors. Section 4 introduces and motivates the proposed collapsed Tucker model. Section 5 presents a numerical study of the Bayesian collapsed Tucker model, focusing on its performance in estimation of π and the parameters of a log-linear model; we also show close agreement to an alternative method on a real data example. Section 6 gives further discussion of results and implications.

2. Notation and background.

We introduce some notation and background on log-linear models and tensor decompositions. Additional notation will be introduced in Section 3. See Table 1 in the Appendix for a list of notation.

2.1. *Log-linear models.*

A standard approach to contingency table analysis parametrizes π as a log-linear model satisfying certain constraints. For a subset of variables $E \subset V$, we adopt the notation of [36] to denote by $\mathbf{i}_E$ the cells in the marginal E-table, so that $\mathbf{i}_E \in \mathcal{I}_E := \bigtimes_{j \in E} \mathcal{I}_j$. Let $\theta_E(\mathbf{i}_E)$ denote the interaction among the variables in E corresponding to the levels in $\mathbf{i}_E$. With this notation, a log-linear model assumes the form

$$\log(\pi_{\mathbf{i}}) = \sum_{E \subset V} \theta_E(\mathbf{i}_E). \tag{2}$$

As a convention, $\theta_\varnothing$ corresponds to $E = \varnothing$. To identify the model, we choose the corner parameterization [1, 36], which sets $\theta_E(\mathbf{i}_E) = 0$ if there exists $j \in E$ such that $i_j = 1$. In the binary setting ($d_j = 2$ for all j) with corner parametrization, any E for which $\theta_E(\mathbf{i}_E) \neq 0$ must have every element of $\mathbf{i}_E$ equal to 2. In this case, we will represent $\theta_E(\mathbf{i}_E)$ as θ_E since there is no ambiguity. When $d > 2$, the notation θ_E refers to the collection of parameters $\{\theta_E(\mathbf{i}_E) : \mathbf{i}_E \in \mathcal{I}_E\}$, and $\theta_E = 0$ indicates $\theta_E(\mathbf{i}_E) = 0$ for all $\mathbf{i}_E \in \mathcal{I}_E$.

Let $\boldsymbol{\theta} = \{\theta_E(\mathbf{i}_E) : i_\gamma \neq 1, \forall \gamma \in E\}$ denote the collection of free model parameters and S_θ denote the collection of nonzero elements of $\boldsymbol{\theta}$. A saturated model includes all free model parameters, so that $|S_\theta| = \prod_j d_j - 1$. Although any model that is not saturated is technically sparse, when we refer to sparse log-linear models we have in mind settings where $|S_\theta| \ll \prod_j d_j - 1$. We will be primarily concerned with how the degree and structure of sparsity affects the nonnegative tensor rank of π.

An attractive feature of log-linear models is that the parameters are interpretable as defining conditional and marginal independence relationships between the y_j's. A log-linear model is hierarchical [7, 9, 36] if for every $E \subset V$ for which $\theta_E = 0$, we have $\theta_F = 0$ for all $F \supseteq E$. Here, we work with a more general class of log-linear models that contains hierarchical models. We refer to this class as weakly hierarchical.

DEFINITION 2.1. A log-linear model is weakly hierarchical when the following condition is satisfied: if $\theta_E(\mathbf{i}_E) = 0$ for $E \subset V$ and $\mathbf{i}_E \in \mathcal{I}_E$, then $\theta_F(\mathbf{i}'_F) = 0$ for every $F \supseteq E$ and $\mathbf{i}'_F \in \mathcal{I}_F$ such that $i'_j = i_j$ for all $j \in E$.

When $d_j = 2$ for all j, weakly hierarchical models and hierarchical models define identical subsets of log-linear models, but if any $d_j > 2$, the collection of hierarchical models is a proper subset of the collection of weakly hierarchical models. To see this, suppose a model is weakly hierarchical. Assume $\theta_E = 0$. Then $\theta_E(\mathbf{i}_E) = 0$ for all $\mathbf{i}_E \in \mathcal{I}_E$. Let $F \supseteq E$. For any $\mathbf{i}'_F \in \mathcal{I}_F$, $\theta_F(\mathbf{i}'_F) = 0$ by weak hierarchicality, since $\theta_E(\mathbf{i}'_E) = 0$. Since $\mathbf{i}'_F$ is arbitrary, we must have $\theta_F = 0$, proving hierarchicality.

The essential difference between hierarchical and weakly hierarchical models is illustrated by the following example. Let $V = \{1, 2, 3\}$ and $d_1 = d_2 = d_3 = 4$. Suppose

$$\begin{aligned} S_\theta = \bigl\{&\theta_{\{1\}}(2), \theta_{\{2\}}(2), \theta_{\{3\}}(2), \theta_{\{1,2\}}(2,2), \theta_{\{1,3\}}(2,2), \theta_{\{2,3\}}(2,2), \\ &\theta_{\{1,2,3\}}(2,2,2)\bigr\}. \end{aligned}$$

In other words, any interactions that correspond to all variables in E taking level 2 are nonzero, and all others are zero. This model is weakly hierarchical but not hierarchical. For a model to be hierarchical, the collection of nonzero parameters must be uniquely specified by a generator—a collection of subsets of V. For weakly hierarchical models, some interactions corresponding to a single subset E may be zero and others nonzero, so long as Definition 2.1 is satisfied.

2.2. *Tensor factorization models.* An alternative to log-linear models is latent structure analysis [2, 20, 21, 23, 32, 35, 42], which assumes the $y_1, \dots, y_p$ are conditionally independent given one or more latent class variables. In marginalizing out the latent class variables, one obtains a tensor decomposition of π. Latent structure models inducing PARAFAC and Tucker decompositions are briefly reviewed below.

2.2.1. *PARAFAC models.* An m-component nonnegative PARAFAC decomposition [24] of a probability tensor π is given by

$$\pi = \sum_{h=1}^{m} \nu_h \lambda_h^{(1)} \otimes \cdots \otimes \lambda_h^{(p)} = \sum_{h=1}^{m} \nu_h \bigotimes_{j=1}^{p} \lambda_h^{(j)}, \tag{3}$$

where $\otimes$ denotes an outer product,[3] each $\lambda_h^{(j)} \in \Delta^{(d_j-1)}$ is an element of the $(d_j - 1)$ dimensional simplex,[4] and $\nu \in \Delta^{(m-1)}$. Element wise, $\pi_{i_1,\dots,i_p} = \sum_{h=1}^{m} \nu_h \prod_{j=1}^{p} \lambda_{hi_j}^{(j)}$. By constraining ν and the $\lambda_h^{(j)}$'s to be probability vectors, it is ensured that the entries of π are nonnegative and sum to one. The vectors $\lambda_h^{(j)}$ are referred to as the arms of the tensor decomposition.

A probabilistic PARAFAC decomposition [15] of π can be induced by a single index latent class model

$$\begin{aligned} y_j \mid z &\overset{\text{ind.}}{\sim} \text{Multi}\big(\{1, \dots, d_j\}, \lambda_{z1}^{(j)}, \dots, \lambda_{zd_j}^{(j)}\big), \\ \Pr(z = h) &= \nu_h, \qquad h = 1, \dots, m. \end{aligned} \tag{4}$$

Marginalizing over the latent variable z, we obtain expression (3).

Unlike matrices, there is no unambiguous definition of the rank of a tensor. A notion of tensor rank is derived restricting attention to PARAFAC expansions. The nonnegative PARAFAC rank of a nonnegative tensor M is the minimal value of m for which there exist nonnegative vectors $\tilde{\lambda}_h^{(j)}$ such that

$$M = \sum_{h=1}^{m} \bigotimes_{j=1}^{p} \tilde{\lambda}_h^{(j)}. \tag{5}$$

We will denote the nonnegative PARAFAC rank of a tensor M as $\text{rnk}_P^+(M)$. In the case of probability tensors, the definition in (5) is equivalent to the minimum m such that (3) holds, since the weights ν_h can be absorbed into the arms $\lambda_h^{(j)}$. For

[3] $\{\bigotimes_{j=1}^{p} \lambda_h^{(j)}\}_{i_1,\dots,i_p} = \prod_{j=1}^{p} \lambda_{hi_j}^{(j)}$.

[4] $\Delta^{(r-1)} = \{x \in \mathbb{R}^r : x_j \geq 0 \, \forall j, \sum_{j=1}^{r} x_j = 1\}$.

probability tensors, we can always write a trivial PARAFAC expansion exploiting the probabilistic structure as

$$
\begin{aligned}
\pi_{i_1,\dots,i_p} &= \Pr(y_1 = i_1 | y_2 = i_2, \dots, y_p = i_p) \Pr(y_2 = i_2, \dots, y_p = i_p) \\
&= \sum_{c_2 \in \mathcal{I}_2} \cdots \sum_{c_p \in \mathcal{I}_p} \Pr(y_1 = i_1 | y_2 = c_2, \dots, y_p = c_p) \mathbb{1}_{(c_2 = i_2, \dots, c_p = i_p)} \\
&\quad \times \Pr(y_2 = c_2, \dots, y_p = c_p).
\end{aligned} \tag{6}
$$

To see the correspondence with (3), introduce one level of h for each distinct value of the multiindex $(c_2, \dots, c_p)$ so that $m = \prod_{j=2}^{p} d_j$, and set $\nu_h = \Pr(y_2 = c_2, \dots, y_p = c_p)$, $\lambda_{hi_1}^{(1)} = \Pr(y_1 = i_1 | y_2 = c_2, \dots, y_p = c_p)$ and $\lambda_{hi_j}^{(j)} = \mathbb{1}_{(i_j = c_j)}$ for $j = 2, \dots, p$. As a consequence, we obtain an upper bound of d^{p-1} on the nonnegative PARAFAC rank $\operatorname{rnk}_P^+(\pi)$ when $d_j = d$ for all j. Thus, every nonnegative tensor has finite nonnegative PARAFAC rank, and the single latent class model has full support.

2.2.2. *Tucker models.* An m-component nonnegative Tucker decomposition [10, 43] alternatively expresses the entries in π as

$$
\pi_{c_1,\dots,c_p} = \sum_{h_1=1}^{m} \cdots \sum_{h_p=1}^{m} \phi_{h_1,\dots,h_p} \prod_{j=1}^{p} \lambda_{h_j c_j}^{(j)}, \tag{7}
$$

where ϕ is an m^p *core* probability tensor and $\lambda_h^{(j)} \in \Delta^{d_j - 1}$ for every h and j. The Tucker decomposition can be thought of as a weighted sum of m^p tensors each having PARAFAC rank one with weights given by the entries in ϕ; conversely, the PARAFAC is a special case of the Tucker decomposition where the core is an $m \times 1$ probability vector.

A probabilistic Tucker expansion of a probability tensor π can be induced by a latent class model with a vector of latent class indicators $z = (z_1, \dots, z_p)$,

$$
\begin{aligned}
y_j | z &\overset{\text{ind.}}{\sim} \operatorname{Multi}(\{1, \dots, d_j\}, \lambda_{z_j 1}^{(j)}, \dots, \lambda_{z_j d_j}^{(j)}), \\
\Pr(z_1 = h_1, \dots, z_p = h_p) &= \phi_{h_1,\dots,h_p}.
\end{aligned} \tag{8}
$$

From this, it is clear that ϕ parametrizes the joint distribution of the latent variables $z_1, \dots, z_p$. See [3] for a class of hierarchical models that induce a structured Tucker decomposition of a probability tensor.

The Tucker decomposition gives rise to an alternative definition of the nonnegative tensor rank of a tensor M as the minimal value of m such that M can be expressed exactly by an expansion of the form in (7). We will denote the nonnegative Tucker rank of a tensor M as $\operatorname{rnk}_T^+(M)$. In the case where $d_j = d$ for all j, an argument similar to the one in (6) shows that for probability tensors π,

$\mathrm{rnk}_T^+(\pi) \le d$. The scale of Tucker ranks is quite different from that of PARAFAC ranks because the core itself has dimension m^p. Therefore, in modeling it is common to choose a parsimonious representation of the core, an issue we revisit in Section 4.

3. Main results: PARAFAC rank of sparse log-linear models.

3.1. *PARAFAC rank result for general p and d.* We now provide bounds on the nonnegative PARAFAC rank of joint probability tensors. There are few results on ranks of tensors beyond three dimensions and the techniques developed here are likely to be of independent interest. All proofs are deferred to the Appendix. In addition to the bounds developed in this section based on probabilistic arguments, we provide algebraic constructions in the two-dimensional case in a supplementary document (see [28]).

In the results that follow, we exploit the fact that a PARAFAC expansion of a probability tensor has a dual representation as a latent variable model (4), and the PARAFAC rank of a probability tensor can be defined in terms of the support of the corresponding latent class variable. Remark 3.1 re-expresses an observation from [34] that formalizes this relationship. For a nonnegative integer-valued random variable w, denote $\mathrm{spt}(w) = \{h : \Pr(w = h) > 0\}$.

REMARK 3.1. Suppose π is a $\prod_{j=1}^p d_j$ probability tensor, and let $y_1, \dots, y_p$ be categorical random variables with joint distribution defined by π. Then $\mathrm{rnk}_P^+(\pi) = \bigwedge_{z \in \mathcal{Z}} |\mathrm{spt}(z)|$, where $\mathcal{Z}$ is the collection of all finitely-supported, discrete latent variables z such that

$$\Pr(y_1 = i_1, \dots, y_p = i_p | z = h) = \prod_{j=1}^{p} \Pr(y_j = i_j | z = h), \tag{9}$$

for all $h \in \mathrm{spt}(z)$ and $\mathbf{i} \in \mathcal{I}_V$.

Therefore, if a latent variable z satisfying (9) can be constructed, then the rank of π can be at most $|\mathrm{spt}(z)|$. Our recipe to create such discrete random variables z is to partition the probability space $\mathcal{Y}$ on which $(y_1, \dots, y_p)$ is defined and assign z a constant value on each partition set. Since $\mathbf{y}$ is a mapping from $\mathcal{Y}$ to $\mathcal{I}_V$, for any partition of $\mathcal{I}_V$, the inverse images of the partition sets under the mapping $\mathbf{y}$ define a partition of $\mathcal{Y}$. We shall restrict our attention to such partitions of $\mathcal{Y}$. As a convention to simplify notation, we shall continue to use Pr to denote probabilities under the probability measure induced on $\mathcal{I}_V$ via the measurable map $\mathbf{y}$. For subsets $B_j \subset \mathcal{I}_j$, it follows from a standard property that $\Pr(y_1 \in B_1, \dots, y_p \in B_p) = \Pr(\bigtimes_{j=1}^p B_j)$, with the first probability defined on the σ-algebra on $\mathcal{Y}$ and the second on the product σ-algebra on $\mathcal{I}_V$. We shall henceforth identify the event $\{y_1 \in B_1, \dots, y_p \in B_p\}$ in $\mathcal{Y}$ with the event

$\times_{j=1}^{p} B_j$ in $\mathcal{I}_V$. For a set $A \in \mathcal{I}_V$, $\Pr(y_1 \in B_1, \dots, y_p \in B_p | A)$ is defined as $\Pr[(\times_{j=1}^{p} B_j) \cap A]/\Pr(A)$.

We now elaborate on the construction of z. For a partition $\mathcal{P}$ of $\mathcal{I}_V$, with $\{A_1, \dots, A_{|\mathcal{P}|}\}$ denoting an (arbitrary) enumeration of the sets in $\mathcal{P}$, we define a discrete random variable $z = z_{\mathcal{P}}$ on $\mathcal{Y}$ corresponding to $\mathcal{P}$ as

$$z = h \mathbb{1}_{A_h}(\mathbf{y}), \qquad h = 1, \dots, |\mathcal{P}|. \tag{10}$$

In particular, for partitions $\mathcal{P}_j$ of $\mathcal{I}_j$, we can define the product partition $\mathcal{P}$ as

$$\mathcal{P} = \mathop{\times}_{j=1}^{p} \mathcal{P}_j := \left\{ \mathop{\times}_{j=1}^{p} B_j : B_j \in \mathcal{P}_j \right\}. \tag{11}$$

It follows from properties of the Cartesian product that $\mathcal{P}$ indeed forms a partition of $\mathcal{I}_V$ and $|\mathcal{P}| = \prod_{j=1}^{p} |\mathcal{P}_j|$.

Clearly, for any z as in (10), (9) is equivalent to

$$\Pr(y_1 = i_1, \dots, y_p = i_p | A_h) = \prod_{j=1}^{p} \Pr(y_j = i_j | A_h), \tag{12}$$

for all $h = 1, \dots, |\mathcal{P}|$ and $\mathbf{i} \in \mathcal{I}_V$. We now proceed to create partitions $\mathcal{P}$ satisfying (12). First, observe that the trivial PARAFAC expansion in (6) corresponds to the product partition (11) with $\mathcal{P}_1 = \mathcal{I}_1$ and $\mathcal{P}_j = \{\{c_j\} : c_j \in \mathcal{I}_j\}$ for $j \geq 2$, so that the event $\{z = h\}$ for each h designates an event of the form $\mathcal{I}_1 \times \{c_2\} \times \cdots \times \{c_p\}$. Clearly, $|\mathcal{P}| = d^{p-1}$; the trivial upper bound. Our main target is to show that much tighter bounds can be achieved under the assumption of weak hierarchicality.

We introduce some additional notation here. For a variable $j \in V$, let $C_\theta^{(j)}$ denote the levels of variable j that share a nonzero two-way or higher order interaction with at least one other variable. For weakly hierarchical models, it is sufficient to only search over the nonzero two-way interactions, so that $C_\theta^{(j)} = \{c_j \in \mathcal{I}_j :$ there exists $j' \neq j$ and $c_{j'} \in \mathcal{I}_{j'}$ such that $\theta_{\{j,j'\}}(c_j, c_{j'}) \neq 0\}$. For any $\boldsymbol{\theta}$, let $C_\theta := \{(E, \mathbf{i}_E) : |E| \geq 2, \theta_E(\mathbf{i}_E) \neq 0\}$ and $C_{\theta,2} := \{(E, \mathbf{i}_E) : |E| = 2, \theta_E(\mathbf{i}_E) \neq 0\}$. Note that C_θ is not the collection of nonzero second or higher order interactions; elements of C_θ are tuples $(E, \mathbf{i}_E)$ such that there is a nonzero interaction among variables in E corresponding to the levels in $\mathbf{i}_E$. $C_{\theta,2}$ is constructed similarly for the nonzero two-way interactions only.

If the model is weakly hierarchical, it follows from Definition 2.1 that for any subset C' of $(C_\theta^{(j)})^c$, $y_j \mathbb{1}_{C'}(y_j) \perp y_{[-j]}$, where $y_{[-j]} = (y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_p)$ and for random variables x_1, x_2, $x_1 \perp x_2$ indicates marginal independence. Thus, instead of having to let the levels of z vary over all events of the form $\{\{c_2\} \cap \cdots \cap \{c_p\}\}$, one can coarsen the partition $\mathcal{P}$ in (11) by pooling together all the levels in $(C_\theta^{(j)})^c$ to form a single element of $\mathcal{P}_j$. Further improvement can be achieved by scanning through the variables in a particular order and only considering the subset of $C_\theta^{(j)}$ that correspond to nonzero two-way interactions with variables that appear later in the ordering. We formalize this observation in Theorem 3.1 below.

THEOREM 3.1. *Suppose π is a d^p probability tensor corresponding to a weakly hierarchical log-linear model. Let σ be a permutation on V. For each $j = 1, \dots, p-1$, denote $G_\sigma^{(j)} = \{\sigma(j+1), \dots, \sigma(p)\}$ and define $B_{\sigma(j)}$ to be the following subset of $C_\theta^{(j)}$:*

$$B_{\sigma(j)} = \{i_{\sigma(j)} \in \mathcal{I}_{\sigma(j)} : \exists f \in G_\sigma^{(j)} \text{ and } i_f \in \mathcal{I}_f \text{ s.t. } \theta_{\{\sigma(j),f\}}(i_{\sigma(j)}, i_f) \neq 0\}.$$

Then the PARAFAC rank $\operatorname{rnk}_P^+(\pi)$ of π is at most

$$\bigwedge_\sigma \prod_{j=1}^{p-1} (|B_{\sigma(j)}| + 1).$$

The bound in Theorem 3.1 gives the correct upper bound d^{p-1} when the model is saturated, since then for any permutation σ we have $|B_{\sigma(j)}| = (d-1)$ for $j = 1, \dots, p-1$. More importantly, it is easy to compute and provides a useful estimate of the order of the PARAFAC rank in d and/or p when the interactions are *uniformly* spread. However, if the interactions are highly structured, Theorem 3.1 may yield the trivial upper bound irrespective of the true rank, as seen in Example 3.3 below.

Our next result provides sharper bounds on the PARAFAC rank. In the first part of Theorem 3.2, we provide a "dimension-free" upper bound that is unaffected by increasing d as long as the true PARAFAC rank is constant. We then present a tight upper bound in the second part of Theorem 3.2 which cannot be globally improved in the class of weakly hierarchical log-linear models.

THEOREM 3.2. *Suppose π is a probability tensor corresponding to a weakly hierarchical log-linear model. Let $H = \{H_1, \dots, H_p\}$ denote collections of sets of indices, where each $H_j \subset \mathcal{I}_j$. Given H, define $T_{(C_\theta, H)} = \{(E, \mathbf{i}_E) \in C_\theta : i_j \in H_j \text{ for some } j \in E\}$ and let*

$$\mathscr{H} = \{H : T_{(C_\theta, H)} = C_\theta\}. \tag{13}$$

Assume $C_\theta^{(j)} \neq \varnothing$ for all j. Then

$$\operatorname{rnk}_P^+(\pi) \leq \bigwedge_{H \in \mathscr{H}} \Bigl(\prod_{j \in V} (|H_j| + 1) \Bigr). \tag{14}$$

For any $l \in V$, set $W_l = \{j \in V \setminus \{l\} : |H_j| = d - 1\}$ and $\bar{W}_l = V \setminus W_l$. Then a tight upper bound on $\operatorname{rnk}_P^+(\pi)$ is

$$\bigwedge_{H \in \mathscr{H}} \bigwedge_{l \in V} \Bigl(\prod_{j \in V} (|H_j| + 1) - \Bigl[\prod_{j \in W_l} (|H_j| + 1) \Bigr] \Bigl[\prod_{j \in \bar{W}_l} |H_j| \Bigr] \Bigr). \tag{15}$$

The full proof of Theorem 3.2 is provided in the Appendix; Example 3.4 illustrates the main ideas of the proof.

REMARK 3.2. By definition, $T_{(C_\theta,H)} \subset C_\theta$, so the condition $T_{(C_\theta,H)} = C_\theta$ in the definition of $\mathscr{H}$ in (13) equivalently requires that for every $(E, \mathbf{i}_E) \in C_\theta$, $i_j \in H_j$ for some $j \in E$. Moreover, for weakly hierarchical models, $T_{(C_\theta,H)} = C_\theta \Leftrightarrow T_{(C_{\theta,2},H)} = C_{\theta,2}$.

REMARK 3.3. Theorem 3.2 assumes $C_\theta^{(j)} \neq \varnothing$ for all j, that is, every variable shares at least one second-order interaction. Clearly, the set of variables which do not satisfy the condition are marginally independent of all other variables and do not contribute to the rank. Letting $U = \{j : C_\theta^{(j)} = \varnothing\}$, the statement of Theorem 3.2 will continue to hold without this assumption as long as we replace all instances of V by $V^* = V \setminus U$.

3.2. *Illustrative examples.* In this subsection, we present two examples to highlight the refinement of the bounds in Theorem 3.2 over Theorem 3.1 and illustrate the main ideas behind the proof of Theorem 3.2.

In the setting of Example 3.3 below, the expressions in (14) and (15) can be explicitly calculated to illustrate the improvement over Theorem 3.1.

EXAMPLE 3.3. Suppose $p = 2$ and $d_1 = d_2 = d$. Assume $\theta_{\{1,2\}}(2, c_2) \neq 0$ for all $c_2 \geq 2$, $\theta_{\{1,2\}}(c_1, 2) \neq 0$ for all $c_1 \geq 2$ and $\theta_{\{1,2\}}(c_1, c_2) = 0$ otherwise. Thus, level 2 of variable 1 interacts with all levels except 1 of variable 2, and similarly, level 2 of variable 2 interacts with all levels except 1 of variable 1. In addition, for convenience of illustration, also assume that all main effects are zero,[5] so that

$$\log \pi_{i_1 i_2} = \theta_0 + \theta_{\{1,2\}}(i_1, i_2)\mathbb{1}_{(i_i=2, i_2\geq 2)} + \theta_{\{1,2\}}(i_1, i_2)\mathbb{1}_{(i_i\geq 2, i_2=2)}.$$

Letting J_d denote the $d \times d$ matrix given by $v_1 \otimes v_2$, where $\{v_1\}_{i_1} = \mathbb{1}_{i_1\neq 2}$ and $\{v_2\}_{i_2} = \mathbb{1}_{i_2\neq 2}$, we can write $\pi = e^{\theta_0} J_d + \tilde{\pi}$, where $\tilde{\pi}$ is a $d \times d$ nonnegative matrix with entries

$$\tilde{\pi}_{i_1 i_2} = e^{\theta_0 + \theta_{\{1,2\}}(i_1, 2)\mathbb{1}_{(i_2=2)} + \theta_{\{1,2\}}(2, i_2)\mathbb{1}_{(i_1=2)}}\mathbb{1}_{(i_i=2 \text{ or } i_2=2)}.$$

Note that $\tilde{\pi}$ is everywhere zero except in the second row and column. In the case of nonnegative matrices, $\operatorname{rnk}_P^+(A)$ equals the ordinary matrix rank $\operatorname{rnk}(A)$ when $\operatorname{rnk}(A) \leq 2$ (see [22]). It is easy to see that the ordinary matrix rank of $\tilde{\pi}$ is 2, since there are at most two linearly independent columns. Hence, $\operatorname{rnk}_P^+(\tilde{\pi}) = 2$ and

[5] Here and in several later examples, we assume that the main effects $\{\theta_E(\mathbf{i}_E) : |E| = 1\}$ are zero for notational brevity. While formally these models are not weakly hierarchical, the inclusion of nonzero main effects do not influence the PARAFAC rank, and hence this assumption can be made without loss of generality.

applying Lemma A.1 in the Appendix, we conclude $\mathrm{rnk}_P^+(\pi) \le 1 + \mathrm{rnk}_P^+(\tilde{\pi}) \le 3$. Barring pathological cases, the ordinary rank $\mathrm{rnk}(\pi)$ will always be 3, and since $\mathrm{rnk}_P^+(A) \ge \mathrm{rnk}(A)$ for matrices [5], $\mathrm{rnk}_P^+(\pi)$ will also be exactly 3.

In applying Theorem 3.1, we have $|B_1| = |B_2| = d - 1$, so that we always get the trivial upper bound d irrespective of the choice of σ.

Next, apply Theorem 3.2. Observe that $H = \{\{2\}, \{2\}\} \in \mathscr{H}$, since all of the interaction terms have either $c_1 = 2$ or $c_2 = 2$, and hence the upper bound in (14) is reduced to 4 irrespective of the value of d. With this choice of H, the expression inside the minimum in (15) becomes $(|H_1| + 1)(|H_2| + 1) - |H_1||H_2| = 4 - 1 = 3$, which returns the exact rank.

As in case of Theorem 3.1, the main strategy of proving Theorem 3.2 is to carefully construct a partition $\mathcal{P}$ of $\mathcal{I}_V$ and define z as in (10). In this case, generate a partition utilizing the sets H_j and establish the conditional independence (12) exploiting the definition of $\mathscr{H}$. Let $\bar{H}_j = \mathcal{I}_j \setminus H_j$ and let $\mathcal{P}_{H,j}$ denote the partition of $\mathcal{I}_j$ consisting of the singleton sets $\{i_j\}$ for $i_j \in H_j$ and the set $\bar{H}_j$. Define a partition $\mathcal{P}_H^0$ of $\mathcal{I}_V$ as the Cartesian product (11) of the partitions $\mathcal{P}_{H,j}$. It is then immediate that $|\mathcal{P}_j| = |H_j| + 1$, and hence $|\mathcal{P}| = \prod_{j=1}^{p}(|H_j| + 1)$. The nontrivial aspect of the proof of (14) is to show that for any $H \in \mathscr{H}$, $y_1, \dots, y_p$ are conditionally independent given any set A in $\mathcal{P}_H^0$. The tight upper bound in (15) of Theorem 3.2 exploits that certain sets in $\mathcal{P}_H^0$ can be merged without sacrificing conditional independence. Although detailed proofs of these facts are provided in the Appendix, we highlight the salient features in Example 3.4, which is an extension of Example 3.3 to higher dimensions with a more complicated interaction structure.

EXAMPLE 3.4. Let $p = 5$ with $d \ge 4$ and suppose S_θ is given by

$$\begin{aligned}
2\theta_{\{1,2\}}(2, c_2) &\neq 0 \quad \text{for } c_2 \ge 2, & \theta_{\{2,3\}}(2, c_3) &\neq 0 \quad \text{for } c_3 \ge 2,\\
\theta_{\{3,4\}}(2, c_4) &\neq 0 \quad \text{for } c_4 \ge 2, & \theta_{\{4,5\}}(2, c_5) &\neq 0 \quad \text{for } c_5 \ge 2,\\
\theta_{\{1,5\}}(c_1, 2) &\neq 0 \quad \text{for } c_1 \ge 2, & \theta_{\{2,4\}}(2, c_4) &\neq 0 \quad \text{for } c_4 \ge 2,\\
\theta_{\{1,4\}}(2, c_4) &\neq 0 \quad \text{for } c_4 \ge 2, & \theta_{\{1,2,4\}}(2, 2, 4) &\neq 0,\\
\theta_{\{2,5\}}(2, c_5) &\neq 0 \quad \text{for } c_5 \ge 2, & \theta_{\{1,5\}}(2, c_5) &\neq 0 \quad \text{for } c_5 \ge 2,\\
\theta_{\{1,2,5\}}(2, 2, 4) &\neq 0,
\end{aligned}$$

so there are two nonzero three-way interactions. It is not difficult to see that Theorem 3.1 gives the trivial bound of d^4 for all $5! = 120$ permutations. Now, let $H_j = \{2\}$ for each j, so that $H = \{\{2\}, \{2\}, \{2\}, \{2\}, \{2\}\}$. From (3.2), we can verify that $H \in \mathscr{H}$. Hence, the conclusion of (14) holds and $\mathrm{rnk}_P^+(\pi) \le 2^5 = 32$, a massive reduction.

As an illustration of the proof technique, we now show that:

1. (12) holds with a specific $A \in \mathcal{P}_H^0$ and a specific cell $\mathbf{i} \in A$, providing intuition for the proof of (14);

2. (12) continues to hold when two example sets in $\mathcal{P}_H^0$ that have $(|V|-1)$ identical coordinate projections that are singleton sets are merged, providing intuition for the proof of (15); and,

3. that (12) fails when two example sets in $\mathcal{P}_H^0$ that do not have $(|V|-1)$ identical coordinate projections that are singleton sets are merged, providing a heuristic for the tightness of (15).

Since $H_j = \{2\}$, $\bar{H}_j = \{1, 3, \dots, d\}$; we shall denote this by $\{\neq 2\}$ for brevity. The partition $\mathcal{P}_{H,j}$ of $\mathcal{I}_j$ therefore consist of the two sets $\{2\}$ and $\{\neq 2\}$ for each $j = 1, \dots, 5$ and the partition $\mathcal{P}_H^0$ has 32 elements.

Part 1. Consider the event $A = \{2\} \times \{2\} \times \{2\} \times \{\neq 2\} \times \{\neq 2\} \in \mathcal{P}_H^0$ and the cell $\mathbf{i} = (2, 2, 2, 4, 4)$. We show that (12) holds with A and $\mathbf{i}$, that is, if A^* denotes the event $\{\mathbf{y} = \mathbf{i}\}$ then

$$\begin{aligned} 2\Pr(A^*|A) &= \Pr(y_1 = 2|A)\Pr(y_2 = 2|A)\Pr(y_3 = 2|A) \\ &\quad \times \Pr(y_4 = 4|A)\Pr(y_5 = 4|A) \\ &= 1 \times 1 \times 1 \times \Pr(y_4 = 4|A)\Pr(y_5 = 5|A). \end{aligned} \tag{16}$$

Now notice that

$$\Pr(y_4 = 4|A) = \sum_{c_5 \neq 2} \frac{\pi_{2224c_5}}{\Pr(A)} = \Pr(A^*|A) \sum_{c_5 \neq 2} \frac{\pi_{2224c_5}}{\pi_{22244}}$$

and similarly

$$\Pr(y_5 = 4|A) = \sum_{c_4 \neq 2} \frac{\pi_{222c_44}}{\Pr(A)} = \Pr(A^*|A) \sum_{c_4 \neq 2} \frac{\pi_{222c_44}}{\pi_{22244}}.$$

So (16) is equivalent to showing

$$\frac{\Pr(A)}{\pi_{22244}} = \sum_{c_4 \neq 2} \sum_{c_5 \neq 2} \frac{\pi_{2224c_5}}{\pi_{22244}} \frac{\pi_{222c_44}}{\pi_{22244}}.$$

Since

$$\frac{\Pr(A)}{\pi_{22244}} = \sum_{c_4 \neq 2} \sum_{c_5 \neq 2} \frac{\pi_{222c_4c_5}}{\pi_{22244}} = \sum_{c_4 \neq 2} \sum_{c_5 \neq 2} \frac{\pi_{222c_4c_5}}{\pi_{222c_44}} \frac{\pi_{222c_44}}{\pi_{22244}},$$

we need to show that

$$\frac{\pi_{222c_4c_5}}{\pi_{222c_44}} = \frac{\pi_{2224c_5}}{\pi_{22244}}. \tag{17}$$

All main effects and interactions that correspond to variables $y_1, \dots, y_4$ will be eliminated in the ratios on both sides, so we focus only on those involving y_5. This gives us that the LHS of (17)—assuming $c_5 \neq 4$—is

$$\exp\bigl(\theta_{\{5\}}(c_5) - \theta_{\{5\}}(4) + \theta_{\{1,5\}}(2, c_5) - \theta_{\{1,5\}}(2,4) + \theta_{\{2,5\}}(2, c_5) \\ - \theta_{\{2,5\}}(2,4) - \theta_{\{1,2,5\}}(2,2,4)\bigr).$$

The RHS differs only in the value of y_4, but since there are no $\{4, 5\}$ interactions at these levels of the variables and the level of y_4 is the same in the numerator and denominator on the RHS, equality holds in (17), despite the fact that there are nonzero three-way interactions. Note that $\theta_{\{1,2,4\}}(2, 2, 4)$ cancelled on the RHS and was either zero or cancelled on the LHS as well (the latter occurring when $c_4 = 4$).

Part 2. Fix $l = 5$. The partition $\mathcal{P}_H^0$ contains the sets

$$A^\delta = \bigl\{\{2\} \times \{2\} \times \{2\} \times \{2\} \times \{2\}\bigr\},$$
$$A^\beta = \bigl\{\{2\} \times \{2\} \times \{2\} \times \{2\} \times \{\neq 2\}\bigr\}.$$

These sets share $|V| - 1 = 4$ coordinate projections that are singleton sets, for example, the set $\{2\}$ corresponding to variables 1 through 4. Now set

$$A^\varepsilon = A^\delta \cup A^\beta = \bigl\{\{2\} \times \{2\} \times \{2\} \times \{2\} \times \mathcal{I}_5\bigr\}$$

and put $\mathcal{P}_H^1 = (\mathcal{P}_H^0 \setminus A^\beta, A^\delta) + A^\varepsilon$. Following the argument in the display after (11), we have conditional independence given A^ε. Since this is the only set in $\mathcal{P}_H^1$ that is not in $\mathcal{P}_H^0$, $\mathcal{P}_H^1$ satisfies (12). Therefore, we see that it is possible to merge two sets that have $(|V| - 1)$ identical coordinate projections that are singleton sets to create a coarser partition that continues to satisfy (12). Though we do not show it rigorously in this example, it is only possible to merge two sets of this form in $\mathcal{P}_H^0$ while maintaining conditional independence, giving us the upper bound $\operatorname{rnk}_P^+(\pi) = 2^5 - 1 = 31$. The same value is given by (15).

Part 3. We now utilize the same setup to demonstrate the key argument as to why (15) is tight. This principle can be described succinctly as the failure of conditional independence upon replacing sets in the partition $\mathcal{A}_H^0$ with their union when these sets do not have in common at least $|V| - 1$ identical singleton events. Let A^β and A^* be as in Parts 2 and 1, respectively, and set

$$A^\gamma = \bigl\{\{2\}, \{2\}, \{2\}, \{\neq 2\}, \{\neq 2\}\bigr\}$$

and note that A^γ and A^β share $3 = |V| - 2$ identical coordinate projections which are singleton events. Then

$$A^\gamma \cup A^\beta = \bigl\{\{2\}, \{2\}, \{2\}, \mathcal{I}_4, \{\neq 2\}\bigr\}.$$

Since A^{γ} and A^{β} share only $|V|-2$ singleton coordinate projections, (12) should fail if we merge these sets. So we want to show that

$$\Pr(A^*|A) \neq \Pr(\mathcal{I}_4|A)\Pr(\{\neq 2\}|A).$$

This will be true iff

$$\frac{\pi_{222c_4c_5}}{\pi_{222c_44}} \neq \frac{\pi_{2224c_5}}{\pi_{22244}} \tag{18}$$

for one or more values of $c_4 \in A_4, c_5 \in A_5$. Here, unlike our previous example using this setup, c_4 can take any value in $\mathcal{I}_4$, *including the value 2*. However, $\theta_{\{4,5\}}(2, c_5) \neq 0$ for any $c_5 \geq 2$. So now on the LHS of (18) we get

$$\begin{aligned}\exp\{&\theta_{\{5\}}(c_5) - \theta_{\{5\}}(4) + \theta_{\{1,5\}}(2, c_5) - \theta_{\{1,5\}}(2, 4) + \theta_{\{2,5\}}(2, c_5)\\ &- \theta_{\{2,5\}}(2, 4) - \theta_{\{1,2,5\}}(2, 2, 4) + \theta_{\{4,5\}}(2, c_5) - \theta_{\{4,5\}}(2, 4)\},\end{aligned}$$

when $c_4 = 2$ and $c_5 \neq 4$. But on the RHS we still get

$$\begin{aligned}\exp\{&\theta_{\{5\}}(c_5) - \theta_{\{5\}}(4) + \theta_{\{1,5\}}(2, c_5) - \theta_{\{1,5\}}(2, 4)\\ &+ \theta_{\{2,5\}}(2, c_5) - \theta_{\{2,5\}}(2, 4) - \theta_{\{1,2,5\}}(2, 2, 4)\}\end{aligned}$$

always, so there are events contained in A where the equality fails and, therefore, conditional independence does not hold.

As a concluding comment, in all the examples where we could calculate the exact rank explicitly, the bound in (15) produced the exact rank. However, to show that (15) provides the exact rank, we need an additional condition; see Remark 3.4 below (a proof is provided in the supplement [28]).

REMARK 3.4. Suppose for every $H \in \mathscr{H}$ for which there exists $H^* \in \mathscr{H}$ such that $H_j^* \subseteq H_j$ for every j, the smallest partition $\mathcal{P}_H^{\inf}$ satisfying (9) that can be formed from unions of the events in $\mathcal{P}_H^0$ satisfies $|\mathcal{P}_H^{\inf}| \geq |\mathcal{P}_{H^*}^{\inf}|$. Then (15) gives the exact value of $\operatorname{rnk}_P^+(\pi)$.

3.3. *Practical consequences of rank results*. We provide corollaries to Theorem 3.1 that give insight into cases where a relatively low PARAFAC rank can be expected. These corollaries motivate subsequent analysis of the statistical properties of latent class models. The number of parameters in a PARAFAC decomposition is given by $(k-1) + k\sum_{j=1}^{p} d_j$, where $k = \operatorname{rnk}_P^+(\pi)$. Hence, the PARAFAC rank determines precisely the parameter complexity of the related latent class model, and bounding the rank is sufficient to bound parameter complexity.

The results in this section make some additional assumptions about the support of the log-linear model. As a basis for comparison across the different cases, we will consider the order of the PARAFAC rank as a function of p or d under different scenarios for the support of the log-linear model. This provides a rough indication of the extent of dimension reduction that is achievable with PARAFAC decompositions in different cases.

Corollary 3.5 shows that when the maximum number of interacting levels of all variables is small relative to d the rank will be substantially reduced.

COROLLARY 3.5. *If* $|C_\theta^{(j)}| < \eta - 1$ *for all* j, $\operatorname{rnk}_P^+(\pi) < \eta^{p-1}$.

PROOF. This follows immediately from Theorem 3.1 by noting that the condition $|C_\theta^{(j)}| < \eta - 1$ implies that $|B_{\sigma(j)}| < \eta - 1$ for every permutation σ and every j. □

In the case where $\eta \ll d$, the condition in Corollary 3.5 reduces the PARAFAC rank by a factor of $(d/\eta)^{p-1}$. However, the PARAFAC rank is still exponential in p, so this assumption is unhelpful in controlling the PARAFAC rank as a function of p. By Theorem 3.2, the exact rank is also exponential in p, so in general the order of the exact PARAFAC rank as a function of p is the same as that given by Corollary 3.5, which relies on Theorem 3.1.

If we also assume that certain types of conditional independence exist, useful bounds on the PARAFAC rank as a function of both d and p can be obtained. Corollary 3.6 gives one such result.

COROLLARY 3.6. *Suppose that the conditions in Corollary* 3.5 *hold and for* $J \subset V$, *set* $y_{(J)} = \{y_j : j \in J\}$. *Then if* $y_{(J^c)}$ *are independent given the variables* $y_{(J)}$, $\operatorname{rnk}_P^+(\pi) \leq \eta^{|J|}$.

The simplest such case is represented by the graphical log-linear model in Example 1 of Figure 1: a single star-graph, where y_7 is the hub variable.[6] More generally, if we consider the special case of graphical models, the setting in Corollary 3.6 has a graphical representation where all edges involve at least one of the variables in J. The PARAFAC rank is then exponential in $|J|$, not p. With $\eta \leq \log d$ and $|J| \leq \log p$, we obtain $\operatorname{rnk}_P^+(\pi) \leq (\log d)^{\log p}$, so the rank becomes at most exponential in $\log p$.

Similar bounds can be obtained when marginal independence exists, which is represented by the graphical model in Example 2 in Figure 1 and formalized for general weakly hierarchical models in Corollary 3.7.

COROLLARY 3.7. *Suppose the conditions of Corollary* 3.5 *hold, and suppose there exists* $J \subset V$ *with the property that* $j \in J^c \Rightarrow y_j \perp y_{[-j]}$. *Then* $\operatorname{rnk}_P^+(\pi) \leq \eta^{(|J|)}$.

[6] While we use graphical representations to simplify exposition, none of the results presented in this section require that the log-linear model is graphical; it is sufficient that it be weakly hierarchical.

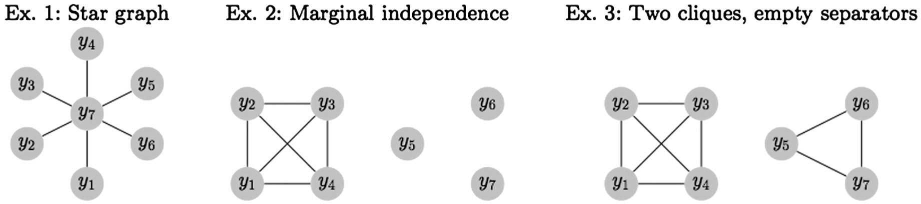

Ex. 1: Star graph
Ex. 2: Marginal independence
Ex. 3: Two cliques, empty separators
y_1
y_2
y_3
y_4
y_5
y_6
y_7

FIG. 1. *Graphical representations of certain sparse log-linear models. Example* 1 *and Example* 2 *are graphs associated with sparse weakly hierarchical log-linear models that have low PARAFAC rank. The model need not be graphical for the rank to be low*; *any weakly hierarchical log-linear model with these dependence graphs will have low rank relative to the maximal rank. Example* 1 *is a canonical example of extensive conditional independence*, *which*, *by Corollary* 3.6 *leads to low PARAFAC rank. Example* 2 *has extensive marginal independence*, *as discussed in Corollary* 3.7. *Example* 3 *corresponds to a sparse log-linear model that has high PARAFAC rank* (*one half of the maximal rank*).

Thus, in this case the PARAFAC rank will depend only on the number of variables that are not marginally independent; the same result that we obtained in Corollary 3.6 with conditional independence. It follows we can also achieve the $(\log d)^{\log p}$ order of the PARAFAC rank in p and d with the same assumptions on η and $|J|$.

The previous results in this section were corollaries to Theorem 3.1, which provides a relatively easy way to calculate bounds on the PARAFAC rank and allows us to clarify cases in which the PARAFAC rank of weakly hierarchical log-linear models will be small. However, this bound is not tight, as illustrated in Example 3.3, and thus when a specific weakly hierarchical interaction structure or class of structures is under consideration, it is necessary to utilize Theorem 3.2 to obtain a tight bound on the rank. We illustrate below through a concrete example that the conclusion of Theorem 3.2 is not simply of theoretical importance, the posterior distribution on the number of components indeed increasingly concentrates on the upper bound implied by Theorem 3.2 as sample size increases.

EXAMPLE 3.8. Set $p = 5$ and $d_j = d = 5$, so that we have a $5^5 = 3125$ cell tensor. Let $\mathbf{n} \sim \text{Multinomial}(N, \pi_0)$, where π_0 corresponds to the weakly hierarchical log-linear model with all main effects nonzero and

$$\begin{aligned} \theta_{\{1,2\}}(2, c_2) &\neq 0 \quad \text{for all } c_2 \geq 2, & \theta_{\{1,2\}}(c_1, 2) &\neq 0 \quad \text{for all } c_1 \geq 2, \\ \theta_{\{1,3\}}(2, c_3) &\neq 0 \quad \text{for all } c_3 \geq 2, & \theta_{\{2,3\}}(2, c_3) &\neq 0 \quad \text{for all } c_3 \geq 2, \\ \theta_{\{1,2,3\}}(2, 2, c_3) &\neq 0 \quad \text{for all } c_3 \geq 2, \end{aligned}$$

with all other interaction terms identically zero and $\theta_{\{\varnothing\}} = 0$ for identification. It can be verified that the minimal H for this model is $\{\{2\}, \{2\}, \varnothing, \varnothing, \varnothing\}$, so the PARAFAC rank is at most 4.

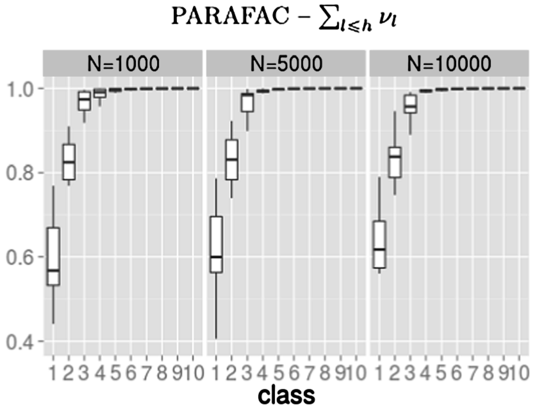
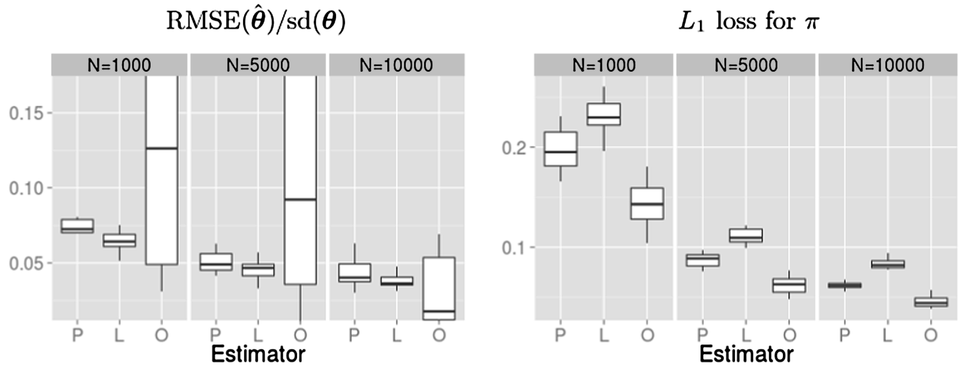


FIG. 2. *Boxplot of posterior mean of cumulative sum of largest* h *class probabilities for* $h = 1, \dots, 10$ *from PARAFAC model estimated on data generated from ten replicate simulations from the log-linear model in Example* 3.8. *The boxes within each panel are the posterior means for* $\sum_{l\leq h}\nu_l$ *for* $h = 1, \dots, 10$ *and the different panels represent sample sizes* $N = 1000$ (*left*), $N = 5000$ (*center*) *and* $N = 10{,}000$ (*right*).

A simulation study was performed to assess performance of the Bayes' PARAFAC model when the data are generated by the sparse weakly hierarchical log-linear model in Example 3.8. The nonzero entries of $\boldsymbol{\theta}$ were sampled from $N(0, 1)$, truncated to lie in the set $(-\infty, -0.2] \cup [0.2, \infty)$. The sampling of the $\boldsymbol{\theta}$ parameters was repeated ten times, and for each sample of the log-linear model parameters, $\mathbf{n}$ was sampled independently for $N = 1000$, 5000 and 10,000—sample sizes that range from about one-third of the number of cells in the table to about three times the number of cells. We then performed MCMC computation for the Bayes' PARAFAC model using the Gibbs' sampling algorithm in [15]. For comparison, we also fit a regularized log-linear model using Lasso with ten-fold cross-validation to select the penalty, as implemented in the `glmnet` package for `R`, and the oracle model—that is, a log-linear model for only the nonzero entries of $\boldsymbol{\theta}$—by maximum likelihood. These comparison methods are used in all subsequent simulation examples.

Figure 2 shows, on the left, a boxplot of the cumulative sum for the largest ten class probabilities (for the class probabilities in descending order of magnitude). The first five class probabilities nearly sum to one in every simulation, with the first four summing to at least 0.95 in each case. Thus, the posterior for the PARAFAC rank concentrates around the theoretical rank of 4. Figure 3 summarizes performance in estimation of $\boldsymbol{\theta}$ and π. Specifically, in this and all subsequent simulation examples, we use the samples of the PARAFAC parameters to obtain samples of π and of $\boldsymbol{\theta}$—the latter by way of the Möbius transformation (see [36])—then use the ergodic average and median as point estimates for $\boldsymbol{\theta}$ and π, respectively. Normalized root mean squared error ($\mathrm{RMSE}(\hat{\boldsymbol{\theta}})/\operatorname{sd}(\boldsymbol{\theta})$) for estimation of $\boldsymbol{\theta}$, as well as the



$\mathrm{RMSE}(\hat{\boldsymbol{\theta}})/\mathrm{sd}(\boldsymbol{\theta})$

L_1 loss for π

FIG. 3. *Left figure*: *Boxplot of* $\mathrm{RMSE}(\hat{\boldsymbol{\theta}})/\,\mathrm{sd}(\boldsymbol{\theta})$ *for PARAFAC* (*P*), *lasso* (*L*) *and oracle MLE* (*O*) *estimated on data generated from ten replicate simulations from the sparse log-linear model in Example* 3.8. *The three subpanels of the figure show results for three different sample sizes* $N = 1000, 5000, 10{,}000$. *Right figure*: *identical arrangement*, *but here the plotted values are the* L_1 *loss for estimation of* π.

L_1 loss for estimation of π, are shown in Figure 3. Here, $\mathrm{sd}(\boldsymbol{\theta})$ is the true standard deviation of the entries of $\boldsymbol{\theta}$ in the simulations. Also shown for comparison are the identical quantities for the Lasso estimator and the oracle MLE. PARAFAC is seen to perform competitively with Lasso for estimating $\boldsymbol{\theta}$ and is superior for estimation of π, despite the fact that generating data from a sparse log-linear model seemingly favors Lasso, which also benefits from cross-validation. There are clear problems with identification for the oracle estimator in the smaller sample sizes resulting from sparsity of the sampled table.

4. Collapsed Tucker decompositions.

Corollaries 3.6 and 3.7 demonstrate the main ways in which exponential scaling of the PARAFAC rank in p can be avoided. However, these settings correspond to special cases of conditional independence mediated by a few variables or extensive marginal independence. More generally, Theorem 3.2 shows that low PARAFAC rank requires that all of the interactions can be accounted for by a small number of levels of the variables, as is the case in Example 3.8. Outside this relatively limited class, PARAFAC rank and, therefore, parameter complexity, scales unfavorably in the dimension of the contingency table. As such, statistical efficiency relative to the log-linear model is expected to degrade as dimensions increase. This is likely most evident in poor recovery of log-linear model parameters, as there may exist low rank expansions that well-approximate π but have quite different values of $\boldsymbol{\theta}$. We show several simulation examples in the sequel in which this degradation of statistical performance of the PARAFAC model occurs, particularly for estimation of $\boldsymbol{\theta}$.

As p grows, the number of classes in the PARAFAC model must grow rapidly to represent complex dependence among the variables. The Tucker decomposition, on the other hand, has p latent class variables, and thus the number of latent classes does not depend on p at all, as shown in the following corollary to Theorem 3.2.

COROLLARY 4.1. *If π is a probability tensor corresponding to a sparse log-linear model then the Tucker rank*

$$\operatorname{rnk}_T^+(\pi) \leq \bigwedge_{H \in \mathscr{H}} \bigvee_{j \in V} (|H_j| + 1),$$

where $\mathscr{H}$ is the collection defined in the statement of Theorem 3.2.

The parsimony gained in the Tucker model by requiring few latent classes is offset to varying degrees by the need to model the dependence between the p latent categorical variables through the $[\operatorname{rnk}_T^+(\pi)]^p$ core tensor—the parameter ϕ in (7). Clearly, unless $\operatorname{rnk}_T^+(\pi) \ll \max_j d_j$, the core is nearly as large as π. Therefore, while PARAFAC rank is an appropriate measure of parameter complexity in single latent class models, the Tucker rank is less meaningful unless d_j is large for most j. When p is even modest in size, parsimony and effective number of parameters in a Tucker model is mainly a function of *how the core is parametrized*. As a result, it becomes critical to count parameters in hierarchical models that induce Tucker decompositions of π rather than simply relying on the rank. For example, [3] used a hierarchical random effects model to borrow information across the entries in the core tensor, greatly reducing parameter complexity relative to having an unstructured prior on the entries of ϕ.

In what follows, we motivate and develop a meta-family of tensor decompositions obtained by allowing the dimension of the core tensor to be any value between 1 (the PARAFAC) and p (the Tucker). We refer to these as collapsed Tucker (c-Tucker) decompositions. These decompositions can be induced by hierarchical latent class models where the number of latent class variables is between 1 and p. To control parameter complexity, we choose to model the core through a latent PARAFAC decomposition. This is a modeling choice, and is not required to induce a c-Tucker decomposition. For example, one could instead choose an analogue of the random effects model of [3] to model the core. To illustrate the advantages of c-Tucker factorizations, we focus on data generated from sparse log-linear models with groups of variables in which there is arbitrary dependence for variables within a group but independence or structured dependence across groups.

4.1. *Independent PARAFACs.* To motivate the c-Tucker decomposition, we first show how a variation of the PARAFAC decomposition can eliminate the exponential factor of $\log(p)$ that appears in Corollary 3.7 in cases where there are multiple groups of variables that are marginally independent of all the other groups.

An example of a graphical model with this dependence structure is shown in Example 3 in Figure 1: two cliques with empty separators.

Divide $y_1, \dots, y_p$ into k groups, and let s_j indicate the group membership of variable j. For each $s \in \{1, \dots, k\}$ define a PARAFAC expansion for the marginal probability tensor corresponding to $\pi^{(s)} = \Pr(\{y_j : s_j = s\})$, as

$$\pi^{(s)} = \sum_{h=1}^{m_s} \nu_{sh} \bigotimes_{j:s_j=s} \lambda_h^{(j)}.$$

We define the joint distribution of $y_1, \dots, y_p$ as

$$\pi_{c_1,\dots,c_p} = \prod_{s=1}^{k} \prod_{j:s_j=s} \pi_{c_j}^{(s)}.$$

This model can be described succinctly as k independent PARAFACs. This is a generalization of the sparse PARAFAC (sp-PARAFAC) model of [44] to the case of more than two groups, and gives much stronger control over parameter growth than PARAFAC when the truth consists of marginally independent groups of variables. This is shown formally for the special case of graphical models with empty separators in Theorem 4.2.

THEOREM 4.2. *Consider a graphical log-linear model for binary data defined by parameters $\boldsymbol{\theta}$. Let $\mathscr{F}$ be the collection of all cliques, and suppose $|\mathscr{F}| = \mathcal{O}(k)$. Then if $\bigvee_{F\in\mathscr{F}} |F| = \mathcal{O}(\log_2(p))$ and all separators are empty, the tensor π can be expressed by k independent tensors $\pi^{(1)}, \dots, \pi^{(k)}$ with $\sum_{s=1}^{k} \operatorname{rnk}_P^+(\pi^{(s)}) = \mathcal{O}(kp)$.*

REMARK 4.1. In the special case where $\log_2(p)$ is an integer and all cliques have identical cardinality, we obtain $\sum_{s=1}^{k} \operatorname{rnk}_P^+(\pi^{(s)}) = \mathcal{O}\left(p^2/\log_2(p)\right)$.

REMARK 4.2. The result in Theorem 4.2 also holds for any weakly hierarchical log-linear model with the same dependence structure, since the graphical model has the maximum number of nonzero interaction terms for any set of dependence/independence relationships.

It follows that where marginally independent sets of variables exist, grouping variables and performing independent PARAFAC decompositions for each of the marginal probability tensors corresponding to the groups can reduce the effective number of parameters drastically. Although Theorem 4.2 is stated for the special case of binary outcomes, conceptually it applies for general d_j and the advantage is borne out empirically, as we show with the following example.

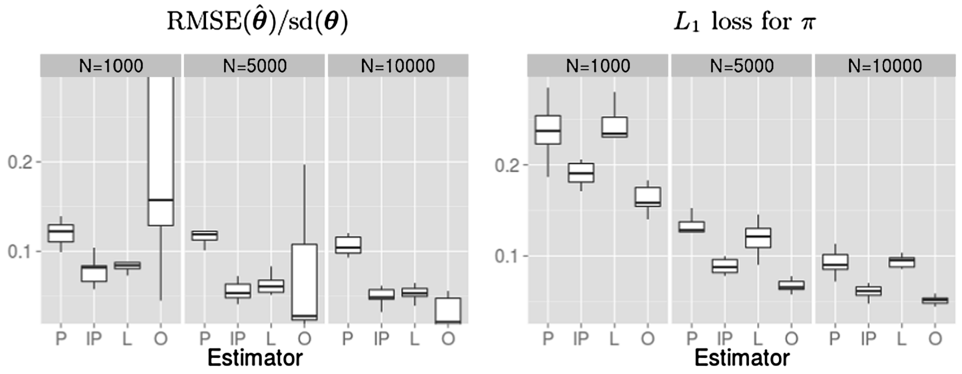


FIG. 4. *Left figure*: *Boxplot of* $\mathrm{RMSE}(\hat{\boldsymbol{\theta}})/\mathrm{sd}(\boldsymbol{\theta})$ *for PARAFAC* (*P*), *independent PARAFAC* (*IP*), *lasso* (*L*) *and oracle MLE* (*O*) *estimated on data generated from ten replicate simulations from the log-linear model in Example* 4.3. *The three subpanels of the figure show results for three different sample sizes* $N = 1000, 5000, 10{,}000$. *Right figure*: *identical arrangement*, *but here the plotted values are the* L_1 *loss for estimation of* π.

EXAMPLE 4.3. Let π be a d^5 probability tensor corresponding to a sparse log-linear model where all main effects are nonzero and in addition

$$\begin{array}{llll}
\theta_{\{1,2\}}(2, c_2) \neq 0 & \text{for } c_2 \geq 2, & \theta_{\{3,4\}}(c_3, 2) \neq 0 & \text{for } c_3 \geq 2, \\
\theta_{\{1,2\}}(c_1, 2) \neq 0 & \text{for } c_1 \geq 2, & \theta_{\{3,5\}}(c_3, 2) \neq 0 & \text{for } c_3 \geq 2, \\
\theta_{\{3,4\}}(2, c_4) \neq 0 & \text{for } c_4 \geq 2, & \theta_{\{4,5\}}(2, c_5) \neq 0 & \text{for } c_4 \geq 2, \\
\theta_{\{3,5\}}(2, c_5) \neq 0 & \text{for } c_5 \geq 2, & \theta_{\{4,5\}}(c_4, 2) \neq 0 & \text{for } c_4 \geq 2,
\end{array}$$

with all other interaction terms equal to zero. Letting $H = \{\{2\}, \{2\}, \{2\}, \{2\}, \{2\}\} \in \mathscr{H}$, we know $\mathrm{rnk}_P^+(\pi) \leq 2^5 = 32$, so the PARAFAC decomposition has approximately $31 + 32 \times 20 = 674$ parameters. The structure of sparsity guarantees that $y_1, y_2 \perp y_3, y_4, y_5$. As a result, the number of parameters in two independent PARAFAC decompositions is only $(3 + 4 \times 8) + (7 + 8 \times 12) = 138$.

We simulated data from the model in Example 4.3 with $d = 5$ using the same distribution for the nonzero log-linear model parameters as in the simulation study for Example 3.8. We performed computation by MCMC for the PARAFAC model as well as the independent PARAFAC model with two variable groups: y_1, y_2 and y_3, y_4, y_5. Figure 4 shows normalized RMSE for estimation of $\boldsymbol{\theta}$ and L_1 loss for estimation of π for PARAFAC, independent PARAFAC, Lasso and the oracle MLE. PARAFAC performs poorly relative to Lasso in estimation of $\boldsymbol{\theta}$ but is comparable to Lasso for estimation of π, suggesting that the posterior concentrates around a lower-rank tensor with entries that are very similar to π but for which the equivalent log-linear model has a rather different value of $\boldsymbol{\theta}$. This is probably a consequence of the fact that the true PARAFAC rank in Example 4.3 is much larger than the PARAFAC rank in Example 3.8, so that the exact expansion has high parameter complexity. In contrast, the independent PARAFAC performs slightly better than Lasso for estimation of $\boldsymbol{\theta}$ and substantially better for estimation of π, despite the fact that the data generating model is a sparse log-linear model.

The approach outlined above is limited to cases in which the variable groups are marginally independent, which in the special case of graphical models corresponds to empty separators. However, additional flexibility can be gained by introducing another set of parameters to control dependence between the groups. This is the essence of the collapsed Tucker model, where we project p dimensional $\mathbf{y}$ to $k \ll p$ dimensional $\mathbf{z}$ and model the joint p.m.f. of $\mathbf{z}$ via a PARAFAC.

4.2. *Latent class models inducing collapsed Tucker decompositions.* We now define c-Tucker decompositions. Specifically, let

$$\pi_{c_1,\dots,c_p} = \sum_{h_1=1}^{m} \cdots \sum_{h_k=1}^{m} \phi_{h_1,\dots,h_k} \prod_{j=1}^{p} \lambda^{(j)}_{h_j^* c_j}, \tag{19}$$

where $h_j^* = h_{s_j}$ with $s_j \in \{1,\dots,k\}$ for $j = 1,\dots,p$ and $k \ll p$ when p is moderate to large. The s_j's are group indices for $\{y_j : j \in V\}$, with $s_j = \rho$ denoting that y_j is allocated to group ρ. For a particular configuration of the s_j's, the p variables are assigned to k groups, and $s_j = s_{j'}$ indicates that y_j and $y_{j'}$ belong to the same group. We refer to (19) as a m-component collapsed Tucker (c-Tucker) factorization.

c-Tucker is a latent class model with k latent class indices. Letting $\mathbf{z} = (z_1,\dots,z_k)^{\mathrm{T}}$ denote a vector of group indices, the c-Tucker model in (19) has a hierarchical representation where given $\mathbf{z}$, $y_1,\dots,y_p$ are conditionally independent with $\Pr(y_j = c_j | \mathbf{z}, s_j) = \lambda^{(j)}_{z_{s_j} c_j}$. The parameter ϕ is a m^k nonnegative core tensor; it is a probability tensor that parametrizes the joint distribution of the latent categorical variables $z_1,\dots,z_k$. Clearly, for $k = 1$ we recover the PARAFAC decomposition and for $k = p$ we obtain the Tucker decomposition. Graphical representations of dependence between observed and latent variables in PARAFAC, Tucker, and c-Tucker models are shown in Figure 8 in the Appendix.

The number of parameters in the core ϕ grows exponentially in k, so superficially the problem of rapidly growing parameter complexity remains. To control this, we model ϕ using a PARAFAC decomposition

$$\phi_{h_1,\dots,h_k} = \sum_{l=1}^{r} \xi_l \prod_{s=1}^{k} \psi^{(s)}_{l h_s}, \tag{20}$$

where $\xi = \{\xi_l\}$ is a vector of probabilities, $\psi_l^{(s)} = \{\psi_{lh}^{(s)}\}$ are probability vectors of dimension m for $s = \{1, \dots, k\}$, and $1 < k < p$. If $r = 1$, we obtain a k-group independent PARAFAC model as in Section 4.1. Under (20), the number of free parameters[7] in a c-Tucker expansion scales as

$$r - 1 + r(m-1)k + m \sum_{j=1}^{p} (d_j - 1). \tag{21}$$

This effective parameter count depends not only on the number of components m but also on r and k, suggesting that unlike in the PARAFAC case the *rank is not useful by itself* as a measure of parsimony. Hence, we focus on parameter count (21) and rank of the core r instead of m in what follows. The first two terms in (21) are specific to the choice of a PARAFAC factorization for the core ϕ, while the term $m \sum_{j=1}^{p} (d_j - 1)$ appears in the parameter count for any c-Tucker factorization.

We can obtain insight into what types of log-linear models might be parsimonious in the c-Tucker representation but not the PARAFAC representation by considering the setup in Theorem 4.2: binary variables consisting of k independent groups each with at most $\log_2(p)$ members. In general, if k marginally independent groups of variables exist and all outcomes are binary, the PARAFAC rank will be of the order 2^{p-k}. The proof of this is straightforward and is omitted. Therefore, Theorem 4.2 gives conditions under which the ordinary PARAFAC rank is approximately 2^{p-k}, with parameter complexity $2^{p-k} - 1 + p2^{p-k}$. Under the same conditions, c-Tucker has parameter complexity of approximately $kp - k + p^2$, obtained from (21) with $r = 1$, $m = p$ and $d_j = 2$ for all j.[8] This is quadratic in p instead of exponential.

5. Estimation and applications for c-Tucker models.

We present an algorithm for inference and computation for c-Tucker models in the Bayesian paradigm. The model is illustrated in simulation studies and an application to the functional disability data from the national long term care survey (NLTCS).

5.1. *Bayesian inference for c-Tucker models.*

Bayesian inference for c-Tucker models requires priors on the parameters of the core, arms and the group memberships of the variables. We choose conjugate Dirichlet priors on the arms $\lambda^{(j)}_{h_j^* c_j}$. We specify truncated stick-breaking priors [27] on the latent class probabilities $\Pr(z_{is} = h)$ and fix the maximum number of latent classes. A similar approach is

[7] These are upper bounds rather than exact expressions. That the effective dimension of the parameter space for a PARAFAC model can in some cases be smaller than the nominal number of parameters in the expansion has been well documented; see, for example, [19] and [16].

[8] The difference between this expression and that in Theorem 4.2 is simply a result of the latter being the sum of PARAFAC ranks and the former a count of free parameters.

used for the arms $\{\zeta_h^{(s)}\}$ in the PARAFAC expansion of the core. When group memberships are inferred, we use a Dirichlet$(1/k, \dots, 1/k)$ prior on variable group membership probabilities.

The Bayesian c-Tucker model can be expressed in hierarchical form as

$$\begin{array}{rl} y_{ij} \mid z_{i1}, \dots, z_{ik}, & \boldsymbol{\lambda}^{(j)} \sim \operatorname{Multi}\bigl(\{1, \dots, d_j\}, \lambda^{(j)}_{z_{ihs_j}1}, \dots, \lambda^{(j)}_{z_{ihs_j}d_j}\bigr), \\ & \boldsymbol{\lambda}_h^{(j)} \sim \operatorname{Diri}(a_{h1}, \dots, a_{hd_j}), \\ z_{is} \mid w_i, & \boldsymbol{\psi}^{(s)} \sim \operatorname{Multi}\bigl(\{1, \dots, m\}, \psi^{(s)}_{w_i 1}, \dots, \psi^{(s)}_{w_i m}\bigr), \\ & \operatorname{pr}(w_i = l) = \nu_l^* \prod_{t<l} (1 - \nu_t^*), \qquad \nu_l^* \sim \operatorname{beta}(1, \beta), \\ & \psi_{lh}^{(s)} = \zeta_{lh}^{(s)} \prod_{h'<h} (1 - \zeta_{lh'}^{(s)}), \qquad \zeta_{lh}^{(s)} \sim \operatorname{beta}(1, \delta_s), \\ & s_1, \dots, s_p \sim \operatorname{Multi}\bigl(\{1, \dots, k\}, \xi_1, \dots, \xi_k\bigr), \\ & \xi \sim \operatorname{Dirichlet}(1/k, \dots, 1/k), \end{array}$$

where the index $i = 1, \dots, n$ is a scalar subject index—not the multiindex $\mathbf{i}$ of a cell of the corresponding contingency table—and y_i is a p-vector of categorical observations for the ith subject. Bayesian computation for this model can be performed using a straightforward Gibbs sampler. Details of the computation are given in the supplement [28].

5.2. *Simulation studies and application for c-Tucker model.* We revisit Example 4.3 to illustrate the performance of the c-Tucker model in the case of marginally independent variable groups. Using data from the simulation procedure in Section 4.1, we performed computation for the c-Tucker model by MCMC using the algorithm in [28], first by fixing two variable groups (y_1, y_2 and y_3, y_4, y_5) and letting the algorithm learn the rank of the core, and then by setting the number of groups to be two and allowing the algorithm to learn both the groups and the rank of the core. In the latter case, the group membership was initialized by performing agglomerative clustering using one minus the pairwise Cramér's V statistic as a dissimilarity matrix for the variables.

Figure 5 shows boxplots of $\sum_{l \le h} \nu_l$ (for ν in descending magnitude order) for the PARAFAC and c-Tucker model with fixed groups. When $N = 1000$, the PARAFAC has approximate posterior rank five—judged by counting the minimal number of classes such that the cumulative class probability is at least 0.99—as does the c-Tucker core tensor, ϕ. However, as the sample size increases, the approximate PARAFAC rank grows, whereas the rank of the c-Tucker core ϕ decreases. With $N = 10{,}000$, the approximate rank of the c-Tucker core decreases to three, with most of the weight on the largest class, whereas the approximate

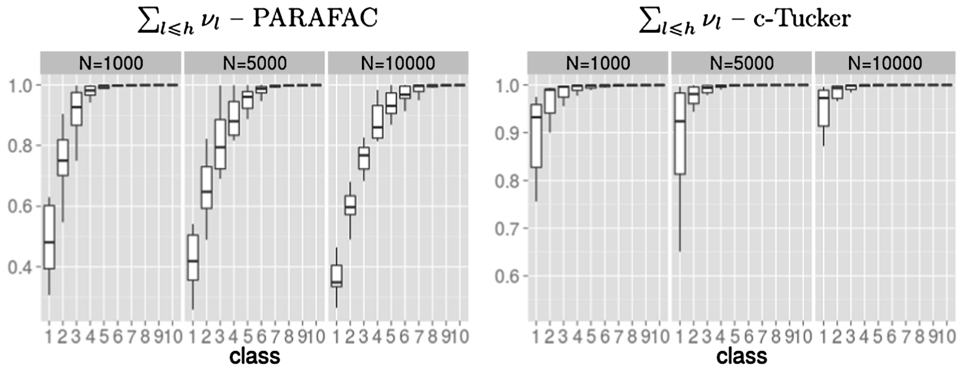


FIG. 5. *Left figure*: *Boxplots of posterior mean of* $\sum_{l\le h}\nu_l$ *for* $h = 1, \ldots, 10$ *for the PARAFAC model estimated on data from over ten replicate simulations from the log-linear model in Example* 4.3. *The boxes within each panel are the posterior means of* $\sum_{l\le h}\nu_l$ *for* $h = 1, \ldots, 10$, *and the three panels correspond to sample sizes* $N = 1000$ (*left*), $N = 5000$ (*center*) *and* $N = 10{,}000$ (*right*). *Right figure*: *the same posterior summary shown for the collapsed Tucker model with fixed groups*; *here*, ν_l *are the component weights in the PARAFAC expansion of the core tensor* ϕ.

PARAFAC rank increases to seven, and the weight on the largest class decreases. Recalling that the PARAFAC rank in this example is 32, while the rank of the c-Tucker core is one, this result is consistent with the ranks converging toward their true values as the sample size grows.

Figure 6 shows performance of PARAFAC, independent PARAFAC, c-Tucker with fixed groups, and c-Tucker with learned groups in estimation of $\boldsymbol{\theta}$ and π (the results for PARAFAC and independent PARAFAC are identical to those in Figure 4 but are shown for ease of comparison). The performance of c-Tucker is seen to be

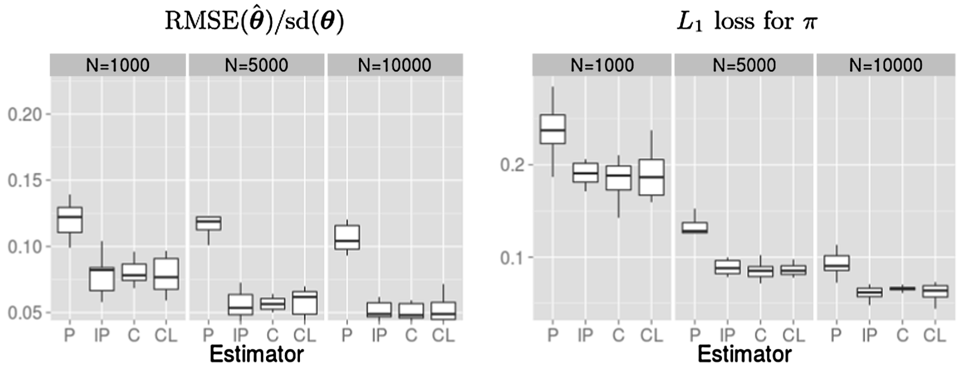


FIG. 6. *Left figure*: *Boxplot of* $\mathrm{RMSE}(\hat{\boldsymbol{\theta}})/\mathrm{sd}(\boldsymbol{\theta})$ *for PARAFAC* (*P*), *independent PARAFAC* (*IP*), *c-Tucker with fixed groups* (*C*) *and c-Tucker with learned groups* (*CL*) *estimated on data from ten replicate simulations from the log-linear model in Example* 4.3. *The three subpanels of the figure show results for three different sample sizes* $N = 1000, 5000, 10{,}000$. *Right figure*: *identical arrangement*, *but here the plotted values are the* L_1 *loss for estimation of* π.

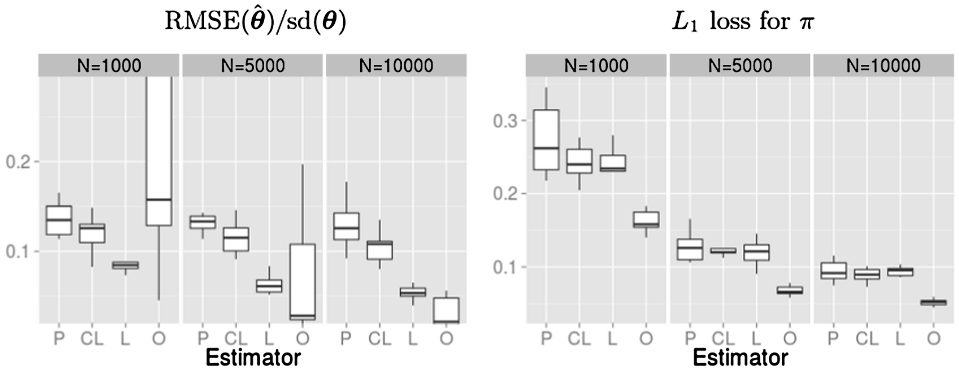


FIG. 7. *Left figure*: *Boxplot of* $\mathrm{RMSE}(\hat{\boldsymbol{\theta}})/\mathrm{sd}(\boldsymbol{\theta})$ *for PARAFAC* (*P*), *c-Tucker with learned groups* (*CL*), *Lasso* (*L*) *and oracle MLE* (*O*) *estimated on data from ten replicate simulations from the log-linear model in Example* 3.4. *The three subpanels of the figure show results for three different sample sizes* $N = 1000, 5000, 10{,}000$. *Right figure*: *identical arrangement*, *but here the plotted values are the* L_1 *loss for estimation of* π.

virtually identical to that of independent PARAFAC at each sample size, regardless of whether groups are fixed or learned, showing that the enhanced flexibility of c-Tucker need not result in loss of performance when the truth is exactly an independent PARAFAC. Recalling that independent PARAFAC is superior to Lasso in this simulation on these loss functions, this indicates better performance for c-Tucker as well. PARAFAC performs poorly relative to methods that incorporate variable grouping, which is as expected for the reasons described in Section 4.1. The superior performance of c-Tucker is consistent with the theoretical results in Sections 3 and 4. In this example, the effective posterior parameter complexity in the PARAFAC and c-Tucker models—computed using (21)—is roughly equivalent, as shown in Figure S.2 in [28]. Thus, c-Tucker provides lower estimation error with similar parameter complexity.

A final simulation illustrates the c-Tucker model in the more challenging case when there are no marginally independent groups of variables, based on Example 3.4. The nonzero entries of $\boldsymbol{\theta}$ were sampled as described in Section 4.1, and ten replicates of each simulation were performed for sample sizes $N = 1000$, 5000 and 10,000. We perform computation by MCMC for both PARAFAC and c-Tucker, and in the latter, we set the number of variable groups to three, allowing learning of the group memberships. Normalized RMSE for estimation of $\boldsymbol{\theta}$ and L_1 loss for estimation of π are shown in Figure 7. c-Tucker outperforms PARAFAC with respect to MSE for estimating $\boldsymbol{\theta}$, while showing similar performance for estimation of π. Lasso is superior for estimation of $\boldsymbol{\theta}$, but similar to PARAFAC and c-Tucker for estimation of π. That PARAFAC performs similarly to Lasso on either metric is surprising given that π corresponds to a sparse log-linear model (only 57 nonzero parameters of 3125), whereas the PARAFAC rank is relatively high (a rank of 32, corresponding to 671 free parameters).[9] This is consistent with the result for Example 3.8 and merits a similar interpretation.

We apply the c-Tucker model with learned groups to analysis of functional disability data from the national long term care survey (NLTCS). The data take the form of a 2^{16} contingency table, and are extensively described in [13], who applied a novel copula Gaussian graphical model. Their model is extremely flexible while favoring parsimony, but has the primary disadvantage of being highly computationally intensive, lacking scalability beyond relatively small tables. Our interest here is in assessing whether the much more computationally efficient c-Tucker model can perform comparably to the [13] approach for these data. We performed posterior computation using the MCMC algorithm described in [28]. Table S.1 in [28] shows the posterior means of pairwise Cramér's V and $\Pr(H_{1,\rho}|\mathbf{y})$, where $H_{1,\rho} = \mathbb{1}(\rho > 0.1)$ and ρ is the pairwise Cramér's V. For comparison, we reproduce the same results based on posterior samples for the copula Gaussian graphical model from [13] in Table S.2 in [28]. Our results demonstrate close agreement with [13].

6. Conclusion. The relationship between the sparsity of a log-linear model and the rank of the associated probability tensor derived here makes clear that a large class of very sparse log-linear models nonetheless has high PARAFAC tensor rank. The statistical consequence of this result is that estimation performance for single latent class models for the joint distribution of multivariate categorical data will tend to degrade as the number of variables grows large, unless dependence in the true model can be accounted for by a small number of levels of the variables, as is the case when marginal independence or highly structured conditional independence exists.

This motivates development of more flexible tensor factorizations that can parsimoniously characterize a broader class of interactions in multivariate categorical data. Tucker factorizations are promising in this regard, and we obtain theory on parameter complexity of Tucker factorizations of sparse log-linear models. These results lead naturally to a novel meta-class of tensor decompositions we refer to as collapsed Tucker. These decompositions are considerably more flexible than either Tucker or PARAFAC, and are highly promising in broad applications. We illustrate some of this promise in simulation examples and an application to real data showing similar results to those obtained with sophisticated graphical modeling methods, which are much more computationally intensive. In fact, computational algorithms for estimation in tensor factorization models in the classes we consider are vastly more scalable to high-dimensional data than algorithms for estimation of sparse log-linear models, so the theoretical results and methods developed here are of substantial practical consequence for high-dimensional statistics with discrete data.

[9] The effective dimension may be smaller than this—see [19].

TABLE 1
Notation reference

Symbol	**Definition**
V	Set of variables, usually $\{1, \ldots, p\}$
$\mathcal{I}_j$	Levels of jth variable, by default $\{1, \ldots, d_j\}$
$\mathbf{i}$	$(i_1, \ldots, i_p)$ with $i_j \in \mathcal{I}_j$, generic notation to denote a cell
$\mathcal{I}_V$	$\times_{j \in V} \mathcal{I}_j$, the collection of all cells
$\mathbf{y}$	$(y_1, \ldots, y_p)$, collection of p variables with $y_j \in \mathcal{I}_j$
π	Joint p.m.f. of $\mathbf{y}$, $\pi_{\mathbf{i}} = \pi_{i_1,\ldots,i_p} = \Pr(y_1 = i_1, \ldots, y_p = i_p)$ for $\mathbf{i} \in \mathcal{I}_V$
$\mathcal{I}_E$	Marginal E-table, $\times_{j \in E} \mathcal{I}_j$
$\mathbf{i}_E$	A cell in the marginal E-table
$\theta_E(\mathbf{i}_E)$	Interactions among variables in E corresponding to the levels in $\mathbf{i}_E$
$\boldsymbol{\theta}$	Free log-linear model parameters in corner parameterization
$\operatorname{spt}(z)$	Support of a discrete random variable, that is, $\{h : \Pr(z = h) > 0\}$
S_θ	Collection of nonzero parameters in $\boldsymbol{\theta}$
$C_\theta^{(j)}$	Collection of levels $c_j \in \mathcal{I}_j$ with a nonzero two-way interaction
C_θ	Collection of tuples $(E, \mathbf{i}_E)$ such that $\theta_E(\mathbf{i}_E) \neq 0$
$C_{\theta,2}$	Collection of tuples $(E, \mathbf{i}_E)$ with $\lvert E\rvert = 2$ such that $\theta_E(\mathbf{i}_E) \neq 0$
H	Collection of sets of indices $\{H_1, \ldots, H_p\}$ with $H_j \subset \mathcal{I}_j$
$T_{C_\theta,H}$	Set of $(E, \mathbf{i}_E) \in C_\theta$ such that $i_j \in H_j$ for some $j \in E$
$\mathscr{H}$	$\{H : T_{(C_\theta,H)} = C_\theta\}$
$\mathcal{P}_{H,j}$	Partition of $\mathcal{I}_j$ consisting of singletons $\{c_j\}$ for $c_j \in H_j$ and the set $\bar{H}_j = \mathcal{I}_j \setminus H_j$
$\mathcal{P}_H^0$	The product partition $\times_{j \in V} \mathcal{P}_{H,j}$
$\operatorname{rnk}_P^+(\pi)$	The nonnegative PARAFAC rank of a nonnegative tensor π
$\operatorname{rnk}_T^+(\pi)$	The nonnegative Tucker rank of a nonnegative tensor π

APPENDIX: PROOFS AND AUXILIARY RESULTS

A.1. Notation. Table 1 provides a summary of notation used throughout the paper.

A.2. Auxiliary results. We state and prove Lemma A.1 which is used to prove Theorem 3.1.

LEMMA A.1. *Let π and ψ be two nonnegative d^p tensors. Then* $\operatorname{rnk}_P^+(\pi \circ \psi) \leq \operatorname{rnk}_P^+(\pi)\operatorname{rnk}_P^+(\psi)$, *where $\circ$ denotes a Hadamard product, and* $\operatorname{rnk}_P^+(\pi + \psi) \leq \operatorname{rnk}_P^+(\pi) + \operatorname{rnk}_P^+(\psi)$.

PROOF. Let $\operatorname{rnk}_P^+(\pi) = m$, $\operatorname{rnk}_P^+(\psi) = k$ and $\phi = \pi \circ \psi$. For $1 \leq j \leq p$, there exist nonnegative vectors $\lambda_h^{(j)} \in \mathbb{R}_+^d$, $h = 1, \ldots, m$ and $\zeta_l^{(j)} \in \mathbb{R}_+^d$, $l = 1, \ldots, k$, such that $\pi = \sum_{h=1}^m \lambda_h^{(1)} \otimes \cdots \otimes \lambda_h^{(p)}$ and $\psi = \sum_{l=1}^k \zeta_h^{(1)} \otimes \cdots \otimes \zeta_h^{(p)}$. Then it is easy to see that

$$\phi = \sum_{h=1}^{m} \sum_{l=1}^{k} \gamma_{hl}^{(1)} \otimes \cdots \otimes \gamma_{hl}^{(p)},$$

where $\gamma_{hl}^{(j)} = \lambda_h^{(j)} \circ \zeta_l^{(j)}$ for $1 \leq j \leq p$. Clearly, for any j, $\gamma_{hl}^{(j)} \in \mathbb{R}_+^d$ for $h = 1, \dots, m; l = 1, \dots, k$. Thus, $\operatorname{rnk}_P^+(\phi) \leq mk$.

In particular, if $\operatorname{rnk}_P^+(\psi) = 1$, we have $\operatorname{rnk}_P^+(\phi) \leq m$. This bound cannot be globally improved, or in other words, the upper bound can be achieved. Take for example, $\psi = \zeta^{(1)} \otimes \cdots \otimes \zeta^{(p)}$, with $\zeta^{(j)} = (1, \dots, 1)^{\mathrm{T}}$ for all j.

Finally, we note that if

$$\pi = \sum_{h=1}^{m_1} \bigotimes_{j=1}^{p} \tilde{\lambda}_h^{(j)} \quad \text{and} \quad \psi = \sum_{h=1}^{m_2} \bigotimes_{j=1}^{p} \tilde{\zeta}_h^{(j)}$$

then

$$\pi + \psi = \sum_{h=1}^{m_1} \bigotimes_{j=1}^{p} \tilde{\lambda}_h^{(j)} + \sum_{h=1}^{m_2} \bigotimes_{j=1}^{p} \tilde{\zeta}_h^{(j)}$$

so $\operatorname{rnk}_P^+(\pi + \psi) = m_1 + m_2 = \operatorname{rnk}_P^+(\pi) + \operatorname{rnk}_P^+(\psi)$. $\square$

Proof of Theorem 3.1. Without loss of generality, we assume σ is the identity permutation and drop the corresponding subscripts. Let $\mathcal{P}^{(1)}$ be the partition of $\mathcal{I}_1$ consisting of the singleton sets $\{c\}$ for $c \in B_1$ and the set $(B_1)^c$. Weak hierarchicality ensures that $y_1 \mathbb{1}_{(y_1 \in A)} \perp\!\!\!\perp y_{[-1]}$ for any $A \in \mathcal{P}^{(1)}$. Using the fact that for any two random variables Z_1, Z_2 and any measurable set A, $Z_1 \mathbb{1}_{(Z_1 \in A)} \perp\!\!\!\perp Z_2 \Leftrightarrow Z_1 \perp\!\!\!\perp Z_2 | A$, we have $y_1 \perp\!\!\!\perp y_{[-1]} | A$ for any $A \in \mathcal{P}^{(1)}$. Enumerating the sets in $\mathcal{P}^{(1)}$ as $A_1, \dots, A_{m_1}$, with $m_1 = |\mathcal{P}^{(1)}| = |B_1| + 1$, we can write π as

$$\pi_{c_1,\dots,c_p} = \sum_{h=1}^{m_1} \nu_h \lambda_{hc_1} \psi_{hc_2,\dots,c_p}, \tag{23}$$

where for each $1 \leq h \leq m_1$, $\nu_h = \Pr(A_h)$, $\lambda_h \in \Delta^{(d-1)}$ with $\lambda_{hc} = \Pr(y_1 = c | A_h)$ and ψ_h is a d^{p-1} nonnegative tensor representing the joint probability of $y_{[-1]} | A_h$, that is,

$$\psi_{hc_2,\dots,c_p} = \Pr(y_2 = c_2, \dots, y_p = c_p | A_h).$$

Define d^p tensors $\{\pi_h^{(1)}\}$ and $\{\pi_h^{(2)}\}$ by

$$\pi_h^{(1)} = \lambda_h \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1},$$

$$(\pi_h^{(2)})_{c_1,\dots,c_p} = \nu_h \psi_{hc_2,\dots,c_p}.$$

The expansion of π in (23) can now be written in tensor notation as $\pi = \sum_{h=1}^{m_1} \pi_h^{(1)} \circ \pi_h^{(2)}$. Clearly, $\operatorname{rnk}_P^+(\pi_h^{(1)}) = 1$ and it is easily verified that $\operatorname{rnk}_P^+(\pi_h^{(2)}) \le \operatorname{rnk}_P^+(\psi_h)$ for all h. Therefore, using Lemma A.1 we have that $\operatorname{rnk}_P^+(\pi) \le m_1 r$, where $r = \operatorname{rnk}_P^+(\psi_h)$.

Recursively applying this process for the variables $y_2, \dots, y_p$, we can show that $r \le \prod_{j=2}^{p} m_j = \prod_{j=2}^{p}(|B_j| + 1)$, so that

$$\operatorname{rnk}_P^+(\pi) \le \prod_{j=1}^{p}(|B_j| + 1).$$

For any permutation σ, we can obtain a result as in the above display by scanning through the variables in the sequence $\sigma(1), \dots, \sigma(p)$. Taking the minimum over all permutations σ, we obtain the desired result.

Proof of (14) in Theorem 3.2. Fix $H \in \mathscr{H}$. Let $\bar{H}_j = \mathcal{I}_j \setminus H_j$ and let $\mathcal{P}_{H,j}$ denote the partition of $\mathcal{I}_j$ consisting of the singleton sets $\{i_j\}$ for $i_j \in H_j$ and the set $\bar{H}_j$. Define a partition $\mathcal{P}_H^0$ of $\mathcal{I}_V$ as the Cartesian product of the partitions $\mathcal{P}_{H,j}$ as in (11). We show that for any set $A \in \mathcal{P}_H^0$, (12) is satisfied, that is,

$$\Pr(y_1 = i_1, \dots, y_p = i_p | A) = \prod_{j=1}^{p} \Pr(y_j = i_j | A), \tag{24}$$

for any $\mathbf{i} \in \mathcal{I}_V$. Based on the discussion in Section 3.1, the random variable $z = z_H^0$ corresponding to the partition $\mathcal{P}_H^0$ defined via (10) will then satisfy (9), implying

$$\operatorname{rnk}_P^+(\pi) \le |\mathcal{P}_H^0| = \prod_{j=1}^{p} |\mathcal{P}_{H,j}| = \prod_{j=1}^{p}(|H_j| + 1).$$

We now proceed to establish (24). Fix $A \in \mathcal{P}_H^0$. By construction,

$$A = \bigtimes_{k \in \bar{J}} \{c_k\} \times \bigtimes_{j \in J} \bar{H}_j \tag{25}$$

for some $J \subset V$, $\bar{J} = V \setminus J$ and $c_k \in H_k$ for all $k \in \bar{J}$. Without loss of generality, we assume $J = \{q, \dots, p\}$ for some integer $q \ge 1$.

Let $\tilde{\mathcal{I}}_V$ denote the subset of $\mathcal{I}_V$ consisting of cells $\mathbf{i}$ such that $i_k = c_k$ for all $k \in \bar{J}$ and $i_j \in \bar{H}_j$ for all $j \in J$. It is easy to see that for any $\mathbf{i} \notin \tilde{\mathcal{I}}_V$, (24) is satisfied trivially since both sides are reduced to zero or one simultaneously. Hence, it suffices to show that (24) holds for any $\mathbf{i} \in \tilde{\mathcal{I}}_V$.

Fix $\mathbf{i} \in \tilde{\mathcal{I}}_V$. Let $A_{\mathbf{i}}$ denote the subset of $\mathcal{I}_V$ corresponding to the event $\{y_j = i_j, j \in V\}$ in $\mathcal{Y}$, so that

$$A_{\mathbf{i}} = \bigtimes_{j \in V} \{i_j\}, \qquad \Pr(A_{\mathbf{i}}) = \pi_{\mathbf{i}}.$$

Clearly, $A_{\mathbf{i}} \subset A$, which implies $\Pr(A_{\mathbf{i}}|A) = \pi_{\mathbf{i}}/\Pr(A)$. Further, $\Pr(y_k = i_k|A) = 1$ for any $k \in \bar{J}$, since $i_k = c_k$ for $k \in \bar{J}$. Therefore, (24) reduces to showing

$$\frac{\pi_{\mathbf{i}}}{\Pr(A)} = \prod_{l \in J} \Pr(y_l = i_l|A). \tag{26}$$

For $E \subset V$, we introduce the notation

$$\bar{\mathbf{H}}_E = \prod_{j \in E} \bar{H}_j.$$

We shall use $\boldsymbol{\alpha}$ to generically denote an element of $\bar{\mathbf{H}}_J$, that is, $\boldsymbol{\alpha}$ is a $|J|$-vector of indices with α_j the entry in $\boldsymbol{\alpha}$ corresponding to variable $j \in J$. For $l \in J$, $J^{(-l)}$ shall denote the set $J \setminus \{l\}$. We use $\boldsymbol{\alpha}^{(l)}$ to generically denote an element of $\bar{\mathbf{H}}_{J^{(-l)}}$, with $\alpha_j^{(l)}$ the entry in $\boldsymbol{\alpha}^{(l)}$ corresponding to variable $j \in J^{(-l)}$.

Finally, for a partition of V into J_1, J_2, J_3, denote[10]

$$\pi_{f_j g_k h_l}^{(J_1,J_2,J_3)} := \Pr\Big[\bigtimes_{j \in J_1} \{f_j\} \times \bigtimes_{k \in J_2} \{g_k\} \times \bigtimes_{l \in J_3} \{h_l\}\Big]. \tag{27}$$

For any $l \in J$,

$$\begin{aligned} \Pr(y_l = i_l|A) &= \frac{\Pr\Big[\bigtimes_{k \in \bar{J}} \{c_k\} \times \{i_l\} \times \bigtimes_{j \in J^{(-l)}} \bar{H}_j\Big]}{\Pr(A)} \\ &= \frac{\pi_{\mathbf{i}}}{\Pr(A)} \sum_{\boldsymbol{\alpha}^{(l)} \in \bar{\mathbf{H}}_{J^{(-l)}}} \frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha_j^{(l)}}}{\pi_{\mathbf{i}}}. \end{aligned} \tag{28}$$

In the above display, we adopt the notation in (27), with V partitioned into $(\bar{J}, \{l\}, J^{(-l)})$ and

$$\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha_j^{(l)}} = \Pr\Big[\bigtimes_{k \in \bar{J}} \{c_k\} \times \{i_l\} \times \bigtimes_{j \in J^{(-l)}} \{\alpha_j^{(l)}\}\Big].$$

From (28), we have

$$\prod_{l \in J} \Pr(y_l = i_l|A) = \left[\frac{\pi_{\mathbf{i}}}{\Pr(A)}\right]^{|J|} \sum_{\boldsymbol{\alpha}^{(q)} \in \bar{\mathbf{H}}_{J^{(-q)}}} \cdots \sum_{\boldsymbol{\alpha}^{(p)} \in \bar{\mathbf{H}}_{J^{(-p)}}} \prod_{l \in J} \frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha_j^{(l)}}}{\pi_{\mathbf{i}}}.$$

[10] Recall our convention, noted in Section 3.1, of identifying the event $\{y_1 \in B_1, \ldots, y_p \in B_p\}$ with the event $\bigtimes_{j=1}^{p} B_j$ in the discrete σ-algebra generated by $\mathcal{I}_V$.

Substituting this in (26), we have (26) is equivalent to showing

$$\begin{aligned}\left[\frac{\Pr(A)}{\pi_{\mathbf{i}}}\right]^{|J|-1} &= \sum_{\boldsymbol{\alpha}^{(q)}\in\bar{\mathbf{H}}_{J(-q)}} \cdots \\ &\quad \times \sum_{\boldsymbol{\alpha}^{(p)}\in\bar{\mathbf{H}}_{J(-p)}} \prod_{l\in J} \frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha^{(l)}_j}}{\pi_{\mathbf{i}}}.\end{aligned} \tag{29}$$

Recalling the set A from (25), we have

$$\frac{\Pr(A)}{\pi_{\mathbf{i}}} = \sum_{\boldsymbol{\alpha}\in\bar{\mathbf{H}}_J} \frac{\pi^{(\bar{J},J)}_{c_k\alpha_j}}{\pi_{\mathbf{i}}},$$

implying

$$\left[\frac{\Pr(A)}{\pi_{\mathbf{i}}}\right]^{|J|-1} = \sum_{\boldsymbol{\alpha}_q\in\bar{\mathbf{H}}_J} \cdots \sum_{\boldsymbol{\alpha}_{p-1}\in\bar{\mathbf{H}}_J} \prod_{l\in J^{(-p)}} \frac{\pi^{(\bar{J},J)}_{c_k\alpha_{lj}}}{\pi_{\mathbf{i}}}, \tag{30}$$

where $\boldsymbol{\alpha}_q,\dots,\boldsymbol{\alpha}_{p-1}$ denote $|J|-1$ independent copies of the running index $\boldsymbol{\alpha}$, and α_{lj} is the entry in $\boldsymbol{\alpha}_l$ corresponding to variable j.

It now amounts to show that the expressions in the RHS of (29) and (30) are the same. We first argue that both expressions contain the same number of terms. To see this, let $|\bar{H}_j| = m_j$. The expression of $\Pr(y_l = i_l|A)$ in (28) is a sum over $\prod_{j\neq l} m_j$ terms, and so $\prod_{l\in J}\Pr(y_l = i_l|A)$ has $\prod_{l\in J}\prod_{j\neq l} m_j = \prod_{l\in J} m_l^{(|J|-1)}$ terms. Accordingly, the RHS in (29) has $\prod_{l\in J} m_l^{(|J|-1)}$ many terms. On the other hand, $\Pr(A)/\pi_{\mathbf{i}}$ is a sum over $\prod_{j\in J} m_j$ terms, and hence $\{\Pr(A)/\pi_{\mathbf{i}}\}^{(|J|-1)}$ in (30) also has $\prod_{j\in J} m_j^{(|J|-1)}$ terms.

Therefore, it now amounts to show that each term inside the summation in the RHS of (29) has a one-to-one correspondence with a term in the RHS of (30). We establish this by showing

$$\prod_{l\in J} \frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha^{(l)}_j}}{\pi_{\mathbf{i}}} = \prod_{l\in J^{(-p)}} \frac{\pi^{(\bar{J},J)}_{c_k\alpha_{lj}}}{\pi_{\mathbf{i}}}, \tag{31}$$

when for each l, $\alpha^{(l)}_j = \alpha_{lj}$ for all $j \neq l$. Introducing additional notation, let $\mathcal{E} = \{E = E_1 \cup \{j\} : E_1 \subset \bar{J}, j \in J_2\}$, $\mathcal{E}^{(-l)} = \{E = E_1 \cup \{j\} : E_1 \subset \bar{J}, j \in J^{(-l)}\}$ and $\mathcal{E}^{(l)} = \{E = E_1 \cup \{l\} : E_1 \subset \bar{J}\}$. For any l, clearly $\mathcal{E}$ is a disjoint union of $\mathcal{E}^{(-l)}$ and $\mathcal{E}^{(l)}$. Let $\mathbf{i}^{(l)}$ denote the cell such that $i^{(l)}_k = c_k$ for $k \in \bar{J}$ and $i^{(l)}_j = \alpha_{lj}$ for $j \in J$.

First, consider the expression in the RHS of (31). We have

$$\begin{aligned}\frac{\pi^{(\bar{J},J)}_{c_k\alpha_{lj}}}{\pi_{\mathbf{i}}} &= \exp\Bigl[\sum_{E\subset V}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr]\\ &= \exp\Bigl[\sum_{E\subset \mathcal{E}}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr] \qquad (32)\\ &= \exp\Bigl[\sum_{E\subset \mathcal{E}^{(-l)}}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr]\exp\Bigl[\sum_{E\subset \mathcal{E}^{(l)}}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr].\end{aligned}$$

The second equality in the above display simply follows from the expression of the cell probabilities for log-linear models in (2). The third equality is the key one which uses (i) since $i_k^{(l)} = i_k = c_k$ for all $k \in \bar{J}$, all interaction terms corresponding to $E \subset \bar{J}$ cancel out; and (ii) any $E \subset V$ such that $|E \cap J| \geq 2, \theta_E(\mathbf{i}_E^{(l)}) = \theta_E(\mathbf{i}_E) = 0$, given weak hierarchically and the condition $C_\theta = T_{C_\theta,H}$. To see this, suppose that there exists $E \subset V$ with $|E \cap J| \geq 2$ such that $\theta_E(\mathbf{i}_E) \neq 0$ for some $\mathbf{i} \in A$. By weak hierarchicality, there must be $j, j^* \in J$ such that $\theta_{\{j,j^*\}}(\alpha_j, \alpha_{j^*}) \neq 0$ for some $(\alpha_j, \alpha_{j^*}) \in \bar{\mathbf{H}}_j \times \bar{\mathbf{H}}_{j^*}$. Then $\theta_{\{j,j^*\}}(\alpha_j, \alpha_{j^*}) \notin T_{C_\theta,H}$, contradicting $C_\theta = T_{C_\theta,H}$.

Using the same argument and additionally the fact that $\alpha_j^{(l)} = \alpha_{lj}$ for all $j \neq l$, we can simplify the expression in LHS of (31) as

$$\frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha_j^{(l)}}}{\pi_{\mathbf{i}}} = \exp\Bigl[\sum_{E\subset \mathcal{E}^{(-l)}}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr]. \tag{33}$$

Therefore,

$$\begin{aligned}&\prod_{l\in J^{(-p)}}\frac{\pi^{(\bar{J},J)}_{c_k\alpha_{lj}}}{\pi_{\mathbf{i}}}\\ &\quad= \prod_{l\in J^{(-p)}}\exp\Bigl[\sum_{E\subset \mathcal{E}^{(-l)}}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr]\\ &\qquad\times\prod_{l\in J^{(-p)}}\exp\Bigl[\sum_{E\subset \mathcal{E}^{(l)}}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr]\\ &\quad= \prod_{l\in J}\exp\Bigl[\sum_{E\subset \mathcal{E}^{(-l)}}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr] = \prod_{l\in J}\frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha_j^{(l)}}}{\pi_{\mathbf{i}}},\end{aligned}$$

establishing (31). The second inequality in the above display used

$$\prod_{l\in J^{(-p)}}\exp\Bigl[\sum_{E\subset \mathcal{E}^{(l)}}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr] = \exp\Bigl[\sum_{E\subset \mathcal{E}^{(-p)}}\bigl\{\theta_E\bigl(\mathbf{i}_E^{(l)}\bigr)-\theta_E(\mathbf{i}_E)\bigr\}\Bigr],$$

since $\mathcal{E}^{(-p)} = \bigcup_{l \neq p} \mathcal{E}^{(l)}$ is a disjoint union.

Proof of (15) in Theorem 3.2. The main idea in this part of the proof is that we can merge certain sets in $\mathcal{P}_H^0$ to create a coarser partition without sacrificing the conditional independence.

For a set $A = \bigtimes_{j \in V} A_j$ in $\mathcal{P}_H^0$ and $J \subset V$, let $\Pi_J(A)$ denote

$$\Pi_J(A) = \prod_{j \in J} A_j.$$

With a slight abuse of notation, we shall use $\Pi_l(A)$ to denote the lth coordinate projection, that is, $\Pi_l(A) = A_l$.

Fix $l \in V$ and let $V^{(-l)} = V \setminus \{l\}$. In this proof, we shall use $\boldsymbol{\alpha}$ to denote a $V^{(-l)}$-cell suppressing the dependence on l. Given $\boldsymbol{\alpha}$, let

$$\mathcal{P}_{H,l}^{\boldsymbol{\alpha}} = \left\{ A \in \mathcal{P}_H^0 : \Pi_{V^{(-l)}}(A) = \bigtimes_{j \neq l} \{\alpha_j\} \right\}. \tag{34}$$

Let $\mathcal{A}$ denote the collection of all $V^{(-l)}$-cells $\boldsymbol{\alpha}$ such that $\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}$ is nonempty. For $\boldsymbol{\alpha} \in \mathcal{A}$, let

$$B^{\boldsymbol{\alpha}} = \bigcup_{A \in \mathcal{P}_{H,l}^{\boldsymbol{\alpha}}} A. \tag{35}$$

Note that for any $\boldsymbol{\alpha} \in \mathcal{A}$, $|\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}| = |H_l| + 1$, since $\Pi_l(A)$ ranges over the elements of $\mathcal{P}_{H,l}$, that is, $\{i_l\}$ for $i_l \in H_l$ and $\bar{H}_l$. It is also evident that $B^{\boldsymbol{\alpha}} = \bigtimes_{j \neq l} \{\alpha_j\} \times \mathcal{I}_l$.

We now create a coarser partition $\mathcal{P}_H^{(l)}$ out of $\mathcal{P}_H^0$ by replacing the collection of sets $\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}$ by the single set $B^{\boldsymbol{\alpha}}$ for every $\boldsymbol{\alpha} \in \mathcal{A}$, so that

$$\mathcal{P}_{H,l} = \bigcup_{\boldsymbol{\alpha} \in \mathcal{A}} [(\mathcal{P}_H^0 \setminus \mathcal{P}_{H,l}^{\boldsymbol{\alpha}}) \cup \{B^{\boldsymbol{\alpha}}\}]. \tag{36}$$

The main idea is that if $(|V| - 1)$ coordinate projections $\Pi_j(A)$ are singletons $\{\alpha_j\}$, we can simply set the lth coordinate projection of A to be $\mathcal{I}_l$ and achieve conditional independence (24). This follows immediately from the expression in the display after (11). However, our construction of $\mathcal{P}_H^0$ clearly contains sets of the form $\bigtimes_{j \neq l} \{\alpha_j\} \times \{i_l\}$ for $i_l \in H_l$ and $\bigtimes_{j \neq l} \{\alpha_j\} \times \bar{H}_l$ which are redundant. To avoid this redundancy, we merge these sets in $\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}$ to form $B^{\boldsymbol{\alpha}} = \bigtimes_{j \neq l} \{\alpha_j\} \times \mathcal{I}_l$ for every $\boldsymbol{\alpha} \in \mathcal{A}$.

It only remains to calculate the cardinality of $\mathcal{P}_{H,l}$ now. As pointed out in the previous paragraph, $|\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}| = |H_l| + 1$ for all $\boldsymbol{\alpha} \in \mathcal{A}$, and hence the net reduction in the number of elements from $\mathcal{P}_H^0$ to $\mathcal{P}_{H,l}$ is

$$\mathcal{P}_H^0 - \mathcal{P}_{H,l} = |\mathcal{A}||H_l|.$$

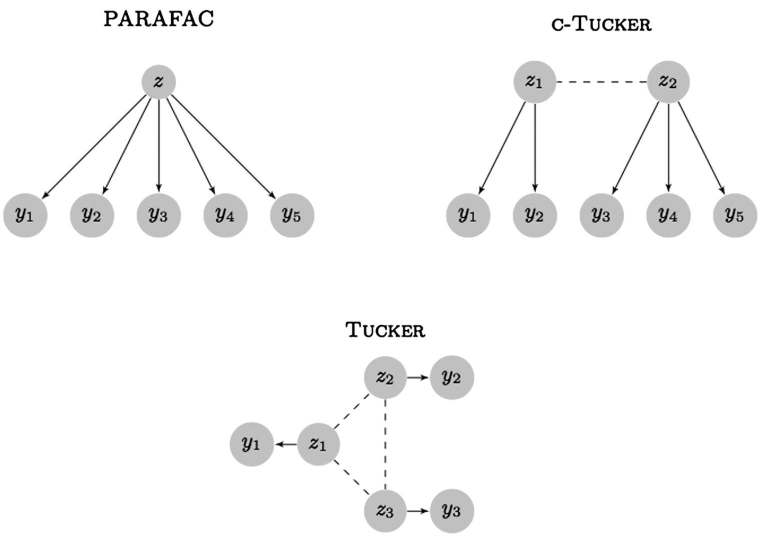
It thus remains to calculate $|\mathcal{A}|$. We need to count the number of distinct $\boldsymbol{\alpha}$ such that (34) is satisfied. Recall that for any $A \in \mathcal{P}_H^0$ and any $j \in V$, $\Pi_j(A)$ ranges over the elements of the partition $\mathcal{P}_{H,j}$. The number of singleton sets in $\mathcal{P}_{H,j}$ is $|H_j|$ as long as $|H_j| < (d-1)$ (the sets $\{i_j\}$ for $i_j \in H_j$). However, when $|H_j| = (d-1)$, $\bar{H}_j$ is also a singleton set, and hence the number of singleton sets in $\mathcal{P}_{H,j}$ in that case becomes $|H_j| + 1$. Therefore, we conclude

$$\mathcal{A} = \Bigg[\prod_{j\neq l:|H_j|=d-1} (|H_j|+1)\Bigg]\Bigg[\prod_{j\neq l:|H_j|<d-1} |H_j|\Bigg].$$

The proof is completed by noting $|\mathcal{A}||H_l| = \prod_{j\in W_l}(|H_j|+1)\prod_{j\in\bar{W}_l}|H_j|$ and taking minimum over $l \in V$ and $H \in \mathscr{H}$.

Proof of Theorem 4.2. The condition that $\bigvee_{F\in\mathscr{F}} |F| = \mathcal{O}(\log_2(p))$ gives that for each clique F the number of terms in the PARAFAC expansion corresponding to that clique is linear in p. This follows because the maximum PARAFAC rank corresponding to the joint distribution of the variables in each clique is bounded by $2^{\lceil \log_2(p)-1\rceil} = \mathcal{O}(p)$. So the joint distribution can be represented by the Hadamard product of k probability tensors $\pi^{(1)}, \ldots, \pi^{(k)}$, with $\operatorname{rnk}_P^+(\pi^{(l)}) = \mathcal{O}(p)$ for every $l = 1, \ldots, k$. Thus, $\sum_{s=1}^{k} \operatorname{rnk}_P^+(\pi^{(s)}) = \mathcal{O}(kp)$. Note that in the special case where $\log_2(p)$ is an integer and all cliques have identical size, this will give $\sum_{s=1}^{k} \operatorname{rnk}_P^+(\pi^{(s)}) = p^2/\log_2(p)$.

Dependence graphs associated with PARAFAC and c-Tucker.



PARAFAC

c-TUCKER

TUCKER

FIG. 8. *Graphical representations of hierarchical models inducing PARAFAC, c-Tucker and Tucker decompositions of π. Dashed edges indicate that there may or may not be an edge between nodes.*

Acknowledgments. The authors wish to thank Kristian Lum for helpful conversations during the preparation of this manuscript and comments on drafts. They also thank the reviewers for helpful comments that led to substantial improvements in the paper.

SUPPLEMENTARY MATERIAL

Supplement to: "Tensor decompositions and sparse log-linear models" (DOI: 10.1214/15-AOS1414SUPP; .pdf). We provide a supplement with three parts. In the first part, we provide a proof of Remark 3.4 and a constructive proof of a bound on nonnegative rank for d^2 tensors corresponding to sparse log-linear models. The second part provides an MCMC algorithm for posterior computation in c-Tucker models and the third part provides supplementary figures and tables for Section 5.

REFERENCES

[1] AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. Wiley, New York. MR1914507

[2] ANDERSON, T. W. (1954). On estimation of parameters in latent structure analysis. *Psychometrika* **19** 1–10. MR0075492

[3] BHATTACHARYA, A. and DUNSON, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *J. Amer. Statist. Assoc.* **107** 362–377. MR2949366

[4] BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (2007). *Discrete Multivariate Analysis*: *Theory and Practice*. Springer, New York. MR2344876

[5] COHEN, J. E. and ROTHBLUM, U. G. (1993). Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra Appl.* **190** 149–168. MR1230356

[6] DAHINDEN, C., KALISCH, M. and BÜHLMANN, P. (2010). Decomposition and model selection for large contingency tables. *Biom. J.* **52** 233–252. MR2756875

[7] DARROCH, J. N., LAURITZEN, S. L. and SPEED, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8** 522–539. MR0568718

[8] DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267

[9] DELLAPORTAS, P. and FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86** 615–633. MR1723782

[10] DE LATHAUWER, L., DE MOOR, B. and VANDEWALLE, J. (2000). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21** 1253–1278 (electronic). MR1780272

[11] DE LATHAUWER, L. DE MOOR, B. and VANDEWALLE, J. (2000). On the best rank-1 and rank-$(r_1, r_2, \dots, r_n)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.* **21** 1324–1342.

[12] DOBRA, A., HANS, C., JONES, B., NEVINS, J. R., YAO, G. and WEST, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212. MR2064941

[13] DOBRA, A. and LENKOSKI, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. MR2840183

[14] DOBRA, A. and MASSAM, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Stat. Methodol.* **7** 240–253. MR2643600

[15] DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004

[16] FIENBERG, S. E., HERSH, P., RINALDO, A. and ZHOU, Y. (2010). Maximum likelihood estimation in latent class models for contingency table data. In *Algebraic and Geometric Methods in Statistics* 27–62. Cambridge Univ. Press, Cambridge. MR2642657

[17] FIENBERG, S. E. and RINALDO, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137** 3430–3445. MR2363267

[18] GARCIA, L. D., STILLMAN, M. and STURMFELS, B. (2005). Algebraic geometry of Bayesian networks. *J. Symbolic Comput.* **39** 331–355. MR2168286

[19] GEIGER, D., HECKERMAN, D., KING, H. and MEEK, C. (2001). Stratified exponential families: Graphical models and model selection. *Ann. Statist.* **29** 505–529. MR1863967

[20] GIBSON, W. A. (1955). An extension of Anderson's solution for the latent structure equations. *Psychometrika* **20** 69–73. MR0075493

[21] GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936

[22] GREGORY, D. A. and PULLMAN, N. J. (1983). Semiring rank: Boolean rank and nonnegative rank factorizations. *J. Comb. Inf. Syst. Sci.* **8** 223–233. MR0783759

[23] HABERMAN, S. J. (1974). Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations. *Ann. Statist.* **2** 911–924. MR0458687

[24] HARSHMAN, R. A. (1970). Foundations of the parafac procedure: Models and conditions for an "explanatory" multimodal factor analysis. UCLA Working Papers in Phonetics **16** 1–84.

[25] HU, J., JOSHI, A. and JOHNSON, V. E. (2009). Log-linear models for gene association. *J. Amer. Statist. Assoc.* **104** 597–607. MR2751441

[26] HUMPHREYS, K. and TITTERINGTON, D. M. (2003). Variational approximations for categorical causal modeling with latent variables. *Psychometrika* **68** 391–412. MR2272386

[27] ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729

[28] JOHNDROW, J. E., BATTACHARYA, A. and DUNSON, D. B. (2016). Supplement to "Tensor decompositions and sparse log-linear models." DOI:10.1214/15-AOS1414SUPP.

[29] KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056

[30] KUNIHAMA, T. and DUNSON, D. B. (2013). Bayesian modeling of temporal dependence in large sparse contingency tables. *J. Amer. Statist. Assoc.* **108** 1324–1338. MR3174711

[31] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

[32] LAZARSFELD, P. F. and HENRY, N. W. (1968). *Latent Structure Analysis*. Houghton, Mifflin, New York.

[33] LETAC, G. and MASSAM, H. (2012). Bayes factors and the geometry of discrete hierarchical loglinear models. *Ann. Statist.* **40** 861–890. MR2985936

[34] LIM, L.-H. and COMON, P. (2009). Nonnegative approximations of nonnegative tensors. *J. Chemom.* **23** 432–441.

[35] MADANSKY, A. (1960). Determinantal methods in latent class analysis. *Psychometrika* **25** 183–197. MR0112763

[36] MASSAM, H., LIU, J. and DOBRA, A. (2009). A conjugate prior for discrete hierarchical log-linear models. *Ann. Statist.* **37** 3431–3467. MR2549565

[37] NARDI, Y. and RINALDO, A. (2012). The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18** 945–974. MR2948908

[38] ROTH, V. and FISCHER, B. (2008). The group-lasso for generalized linear models: Uniqueness of solutions and efficient algorithms. In *Proceedings of the* 25*th International Conference on Machine Learning* 848–855. ACM, New York.

[39] RUSAKOV, D. and GEIGER, D. (2002). Asymptotic model selection for naive Bayesian networks. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* 438–455. Morgan Kaufmann, San Francisco, CA.

[40] SETTIMI, R. and SMITH, J. Q. (1998). On the geometry of Bayesian graphical models with hidden variables. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* 472–479. Morgan Kaufmann, San Francisco, CA.

[41] SMITH, J. Q. and CROFT, J. (2003). Bayesian networks for discrete multivariate data: An algebraic approach to inference. *J. Multivariate Anal.* **84** 387–402. MR1965229

[42] STOUFFER, S. A., GUTTMAN, L., SUCHMAN, E. A., LAZARSFELD, P. F., STAR, S. A. and CLAUSEN, J. A. (1950). Measurement and prediction. Princeton Univ. Press, Princeton, NJ.

[43] TUCKER, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31** 279–311. MR0205395

[44] ZHOU, J., BHATTACHARYA, A., HERRING, A. H. and DUNSON, D. B. (2015). Bayesian factorizations of big sparse tensors. *J. Amer. Statist. Assoc.* **110** 1562–1576. MR3449055

J. E. JOHNDROW
D. B. DUNSON
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
114 OLD CHEMISTRY BUILDING
DURHAM, NORTH CAROLINA 27708
USA
E-MAIL: james.johndrow@duke.edu
dunson@duke.edu
URL: http://www.isds.duke.edu/~dunson/

A. BHATTACHARYA
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
155 IRELAND STREET
COLLEGE STATION, TEXAS 77843
USA
E-MAIL: anirbanb@stat.tamu.edu