Journal of Multivariate Analysis 100 (2009) 1282–1290

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Canonical representation of conditionally specified multivariate discrete distributions

Edward H. Ip [a,b,*], Yuchung J. Wang [c]

[a] Department of Biostatistical Sciences, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA

[b] Department of Social Sciences & Health Policy, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA

[c] Department of Mathematical Sciences, Rutgers University, Camden, NJ 08102, USA

ARTICLE INFO

Article history:
Received 27 September 2007
Available online 7 December 2008

AMS 2000 subject classifications:
primary 62E10
62E15
secondary 62E17
62H05

Keywords:
Canonical parameter
Characterizing set of interactions
Compatibility check
Exponential family
Near-compatible
Pseudo-Gibbs sampler

ABSTRACT

Most work on conditionally specified distributions has focused on approaches that operate on the probability space, and the constraints on the probability space often make the study of their properties challenging. We propose decomposing both the joint and conditional discrete distributions into characterizing sets of canonical interactions, and we prove that certain interactions of a joint distribution are shared with its conditional distributions. This invariance opens the door for checking the compatibility between conditional distributions involving the same set of variables. We formulate necessary and sufficient conditions for the existence and uniqueness of discrete conditional models, and we show how a joint distribution can be easily computed from the pool of interactions collected from the conditional distributions. Hence, the methods can be used to calculate the exact distribution of a Gibbs sampler. Furthermore, issues such as how near compatibility can be reconciled are also discussed. Using mixed parametrization, we show that the proposed approach is based on the canonical parameters, while the conventional approaches are based on the mean parameters. Our advantage is partly due to the invariance that holds only for the canonical parameters.

1. Introduction

Conditional statements are rather natural human perspectives. A physician, for example, would have no difficulty in citing the risk of stroke conditioned on combinations of risk factors such as smoking, obesity, diabetics etc. Nevertheless, a family of conditional distributions only contains partial information about the joint distribution. When multiple families of conditional distributions are presented, they may contain overlapping information about the joint, leading to possible conflicts. On the other hand, the entirety of the separate pieces of partial information may not be sufficient to fully specify a joint distribution. In this case, the set of conditionals is said to be incomplete. Note that in general a set of conditionals may be both overlapping and incomplete in terms of specifying a joint distribution.

The framework of conditionally specified distributions has intrigued many researchers [1–5], and a recent review of conditionally specified distributions was provided by Arnold et al. [6]. Three common issues arise from this body of research: (1) whether, given a set of conditional distributions, a joint distribution exists; (2) if the joint does exist, whether or not it is unique; and, (3) how a joint distribution can be efficiently computed. In this paper, we address all three issues for discrete distributions. Most work on conditionally specified discrete models in the literature has focused on the probability space (e.g., [7,6]); compatibility check in the probability space can be complicated and often requires solving constrained linear equations. Furthermore, many proposed methods may become "unwieldy" as the number of variables increases [8]. Here,

* Corresponding address: WC23 Medical Center Boulevard, Winston-Salem, NC 27157, USA.
E-mail addresses: eip@wfubmc.edu (E.H. Ip), yuwang@crab.rutgers.edu (Y.J. Wang).

0047-259X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.11.010

we adopt a different paradigm, in which, instead of operating on the probability space we operate on the space of canonical interactions. They are said to be canonical because these interactions are the canonical parameters in an exponential family. Moreover, interactions of various orders are defined by difference operators acting upon the logarithms of probabilities. To assess compatibility, every conditional distribution is first decomposed into interactions and then overlapping interactions are checked for possible conflicts. A joint distribution is built from the pool of interactions, if no conflict has been found.

The theoretical justifications of our approach are anchored on two important results. First, we prove that certain interactions of a joint distribution are shared with its conditional distribution. These shared interactions are said to be invariant. Second, we show how those invariant interactions can build a unique conditional distribution. We call such a collection of interactions the characterizing set of interactions (*CSOI*). The CSOI provides a useful platform for developing results that address the existence, uniqueness, and computation issues.

To motivate our approach and to familiarize readers with the necessary notation, consider the following example of adverse reaction to drug treatment and genotype in cancer patients. Treatment regimen that includes irinotecan has been found to be effective for colorectal cancer patients, but the individual reaction to drug toxicity is known to be highly variable [9]. Genotype in the regions of the gene UGT1A1 is known to be associated with adverse reaction, and the association can be explained through a link of the gene to an active irinotecan metabolite SN38 pharmacokinetics. The genotype, X_1, has three variations: TA7/TA7, TA7/TA6, and TA6/TA6, coded as 1-3, respectively, where $X_1 = 1$, $X_1 = 3$ are homozygous, and $X_1 = 2$ is heterozygous. Let X_2 represent the four levels of adverse reaction: severe, moderate, mild, and nil, coded as 1-4, respectively. Two conditional models are commonly used by clinicians: the diagnostic conditional model $X_1|X_2$, and the treatment conditional model $X_2|X_1$. Of special interest are the following two parameters: the diagnostic likelihood ratios $d_{ij} = P(X_1 = i|X_2 = j)/P(X_1 = i|X_2 = j+1)$, $1 \le i \le 3$, $1 \le j \le 3$, and the treatment risk ratios $t_{ij} = P(X_2 = j|X_1 = i)/P(X_2 = j|X_1 = i+1)$, $1 \le i \le 2$, $1 \le j \le 4$. Both parameters can be easily interpreted by clinicians. For example, the diagnostic d_{11} is the ratio of the likelihood that a positive test result (7/7) would be expected in a patient with severe adverse reaction to the likelihood that the same result would be expected in a patient with moderate adverse reaction. Post-test odds of adverse reaction given test result can be quickly obtained by multiplying the diagnostic likelihood ratio to the pre-test odds or through a device such as the nomogram.

To create a parsimonious diagnostic conditional model, suppose a clinician assumes that the diagnostic likelihood ratios follow the same multiplicative constant across adjacent levels of severity, i.e. $d_{ij}/d_{(i+1)j} = \delta_i$ and that a similar model can be specified for treatment conditional ratios, i.e., $t_{ij}/t_{i(j+1)} = \tau_j$. Two important questions of interest are immediate: (1) Are the two conditional models compatible under the designated model constraints? (2) If they are, then how would one compute the joint density?

As we shall see later from our compatibility results, the diagnostic and the treatment conditional models are only compatible when $\delta_i = \tau_j = c$ for all i, j and for some constant c. For the second question, we borrow a numerical example from Arnold et al. [7], p. 23) to show how the joint distribution can be efficiently computed from the CSOI of two compatible and complete conditional probabilities.

Example 1. Let $\mathbf{a}_{i|j}$ and $\mathbf{b}_{j|i}$ denote, respectively, the two families of conditional distributions $(X_1|X_2)$ and $(X_2|X_1)$, where

$$\mathbf{a}_{i|j} = \begin{pmatrix} 1/7 & 1/4 & 3/7 & 1/7 \\ 2/7 & 2/4 & 1/7 & 2/7 \\ 4/7 & 1/4 & 3/7 & 4/7 \end{pmatrix}, \qquad \mathbf{b}_{j|i} = \begin{pmatrix} 1/6 & 1/6 & 3/6 & 1/6 \\ 2/7 & 2/7 & 1/7 & 2/7 \\ 2/6 & 1/12 & 1/4 & 1/3 \end{pmatrix}.$$

Distribution $\mathbf{a}_{i|j}$ are characterized by (1) the last-column consecutive odds: $\alpha_i = P(X_1 = i|X_2 = 4)/P(X_1 = i+1|X_2 = 4)$, $1 \le i \le 2$, and hence $\alpha_1 = 1/2$ and $\alpha_2 = 1/2$; and (2) the cross-distribution odds ratios: $r_{ij} = P(X_1 = i|X_2 = j)P(X_1 = i+1|X_2 = j+1)/[P(X_1 = i+1|X_2 = j)P(X_1 = i|X_2 = j+1)]$, $1 \le i \le 2$, $1 \le j \le 3$. Similarly, distribution $\mathbf{b}_{j|i}$ are characterized by (1) the last-row consecutive odds: $\beta_j = P(X_2 = j|X_1 = 3)/P(X_2 = j+1|X_1 = 3)$, $1 \le j \le 3$, and hence $\beta_1 = 4$, $\beta_2 = 1/3$, and $\beta_3 = 3/4$; and (2) the cross-distribution odds ratios: $r'_{ij} = P(X_2 = j|X_1 = i)P(X_2 = j+1|X_1 = i+1)/[P(X_2 = j+1|X_1 = i)P(X_2 = j|X_1 = i+1)]$, $1 \le i \le 2$, $1 \le j \le 3$. The numerical values for the odds ratios are as follows:

$$(r_{ij}) = (r'_{ij}) = \begin{pmatrix} 1 & 1/6 & 6 \\ 1/4 & 6 & 2/3 \end{pmatrix}.$$

The two characterizing sets of odds and odds ratios overlap on odds ratios; moreover, their overlapping parts have no conflict, since $r_{ij} = r'_{ij}$ for all i, j. This implies that they are compatible and that a joint distribution exists. A formal proof will be given in Section 3.

If we pool the two characterizing sets, there are 6 r_{ij}s, 2 α_is, and 3 β_js, altogether 11 odds and odds ratios. The joint distribution also has $3 \times 4 - 1 = 11$ degrees of freedom. Hence, a unique joint distribution compatible with both $\mathbf{a}_{i|j}$ and $\mathbf{b}_{j|i}$ can be built from them. To compute the joint distribution, we start with a 3×4 positive matrix (q_{ij}) and assign $q_{34} = 1$. Next, select q_{ij} values such that the consecutive odds of the last row and last column of (q_{ij}) are identical to α_i and β_j, respectively. This leads to the following (q_{ij}) matrix:

$$(q_{ij}) = \begin{pmatrix} \cdot & \cdot & \cdot & 1/4 \\ \cdot & \cdot & \cdot & 1/2 \\ 1 & 1/4 & 3/4 & 1 \end{pmatrix}.$$

Then, starting from the lowest right corner, sequentially enter the remaining six q_{ij} such that the (r_{ij}) is preserved in (q_{ij})—that is, $q_{ij} = r_{ij}q_{i+1j}q_{ij+1}/q_{i+1j+1}$, $1 \le i \le 2$, $1 \le j \le 3$. We have

$$(q_{ij}) = \begin{pmatrix} 1/4 & 1/4 & 3/4 & 1/4 \\ 1/2 & 1/2 & 1/4 & 1/2 \\ 1 & 1/4 & 3/4 & 1 \end{pmatrix}.$$

Generally, the computation of (q_{ij}) can be implemented through an efficient algorithm. Let $\mathbf{B}_1$ and $\mathbf{B}_2$ be, respectively, 3×3 and 4×4 upper triangular matrices of 1's, and let $\mathbf{A} = \mathbf{B}_1 \otimes \mathbf{B}_2$, where the $\otimes$ is the Kronecker product. It will be shown later that (q_{ij}) are equivalent to $\exp(\mathbf{A}\boldsymbol{\theta})$, where $\boldsymbol{\theta}^{\mathrm{T}} = (\log(r_{11} = 1), \log(r_{21} = 1/4), \log(\beta_1 = 4), \log(r_{21} = 1/6), \log(r_{22} = 6), \log(\beta_2 = 1/3), \log(r_{31} = 6), \log(r_{23} = 2/3), \log(\beta_3 = 3/4), \log(\alpha_1 = 1/2), \log(\alpha_2 = 1/2), 0)$ consists of the logarithms of odds-ratios, last-row odds, and last-column odds in lexicographical order.

Finally, normalize (q_{ij}) to form the joint (π_{ij}):

$$(\boldsymbol{\pi}_{ij}) = (1/25)\begin{pmatrix} 1 & 1 & 3 & 1 \\ 2 & 2 & 1 & 2 \\ 4 & 1 & 3 & 4 \end{pmatrix}.$$

The same example may also be solved by the Gibbs sampler [2], which iteratively draws samples from $\mathbf{a}_{i|j}$ and $\mathbf{b}_{j|i}$.

We now provide a roadmap of the mathematical results we will present. In Section 2, we describe a decomposition of a multivariate discrete distribution and show that its canonical parameters are the interactions we will use. Afterward, in the same section, we show the invariance of the canonical interactions and demonstrate how they can be used to characterize conditional probabilities. We focus on conditional probabilities involving all J variables—that is, conditionals of the form $\boldsymbol{\pi}_{\bar{a}|a}$, where $a \subset \aleph$, $\bar{a} \cup a = \aleph$, and $\aleph$ represents the set of all J variables. Conditional distributions that satisfy this condition are called *J-conditionals*. Necessary and sufficient conditions for compatibility and uniqueness for J-conditionals are described in Section 3. If the interaction pool of a conditional model is not complete, the joint distribution is under-specified and hence can not be unique. The Posterior Slice Lemma, also described in Section 3, allows a special form of conditional information that supplements an incomplete conditional model.

In Section 4, we provide a theoretical justification of CSOI in which we show that CSOI is the canonical component of the mixed parametrization [10, p. 122] of the joint. As an application of the CSOI, we propose alternatives to the pseudo-Gibbs sampler [11] in finding near-compatible joint distributions.

2. Canonical representation of a discrete multivariate distribution and its conditional distributions

Assume that there are J discrete random variables and each $X_i = k_i$, $1 \le k_i \le K_i$ with positive probabilities. The random vector $\mathbf{X} = (X_1, \ldots, X_J)$ has joint distribution $\boldsymbol{\pi}_\aleph$ consisting of probabilities $\pi_{k_1,\ldots,k_J} = P[X_1 = k_1, \ldots, X_J = k_J]$. Let $\mathbf{B}_i$, $1 \le i \le J$ be an $K_i \times K_i$ upper triangular matrix of 1s, and let $\mathbf{A}$ be $\mathbf{B}_J \otimes \cdots \otimes \mathbf{B}_1$, where $\otimes$ is the Kronecker product. Also, let vector $\log \boldsymbol{\pi}_\aleph = (\log \pi_{1\ldots1}, \ldots, \log \pi_{K_1\ldots K_J})^{\mathrm{T}}$ be arranged in lexicographical order such that the first index changes the fastest and the last index the slowest. Clearly, $\log \boldsymbol{\pi}_\aleph$ is of length $K_1 \times \cdots \times K_J = K$, and $\mathbf{A}$ is a $K \times K$ invertible matrix. We can reparametrize $\boldsymbol{\pi}_\aleph$ as

$$\boldsymbol{\pi}_\aleph = \exp(\mathbf{A}\boldsymbol{\theta}_\aleph), \tag{1}$$

where $\boldsymbol{\theta}_\aleph = \mathbf{A}^{-1} \log \boldsymbol{\pi}_\aleph$. Eq. (1) is a log-linear decomposition of $\boldsymbol{\pi}_\aleph$, and $\boldsymbol{\theta}_\aleph$ is called the vector of interactions. To show the general structure of the matrices, we use an example. Consider a $3 \times 2 \times 2$ table with cell probabilities $\boldsymbol{\pi} = (\pi_{111}, \pi_{211}, \ldots, \pi_{322})^{\mathrm{T}}$. Let

$$\mathbf{B}_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{B}_1 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2}$$

Then the incidence matrix $\mathbf{A} = \mathbf{B}_2 \otimes \mathbf{B}_2 \otimes \mathbf{B}_1$. Because the inverse matrix of a $\otimes$-product of matrices is the $\otimes$-product of the inverse matrices, we have $\mathbf{A}^{-1} = \mathbf{B}_2^{-1} \otimes \mathbf{B}_2^{-1} \otimes \mathbf{B}_1^{-1}$, where

$$\mathbf{B}_2^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{B}_1^{-1} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

The components of $\boldsymbol{\theta}_\aleph$ are more conveniently represented by the following difference operator [12, p. 35]: for $1 \le k_i < K_i$, $1 \le i \le J$,

$$\bigtriangledown_i h(k_1, \ldots, k_i, \ldots, k_J) = h(k_1, \ldots, , k_{i-1}, k_i, k_{i+1}, \ldots, k_J) - h(k_1, \ldots, k_{i-1}, k_i + 1, k_{i+1}, \ldots, k_J),$$

and $\bigtriangledown_i h(k_1, \ldots, K_i, \ldots, k_J) = 0$. Here, the function h can be the logarithm of a joint distribution, or a J-conditional. Define $\boldsymbol{\nabla}^i$ as the vector $(\bigtriangledown_i, \ldots, \bigtriangledown_i, \vartheta)^{\mathrm{T}}$, where $\vartheta : \vartheta h(k_1, \ldots, k_J) = h(k_1, \ldots, k_J)$ is the identity operator. Furthermore, let $\boldsymbol{\nabla}$ denote $\boldsymbol{\nabla}^J \otimes \cdots \otimes \boldsymbol{\nabla}^1$, which is of length K. Define a Hamadan product $\cdot$ between two vectors as $(x_1, \ldots, x_K)^{\mathrm{T}} \cdot (y_1, \ldots, y_K)^{\mathrm{T}} = (x_1y_1, \ldots, x_Ky_K)^{\mathrm{T}}$.

Lemma 1. *The interaction vector $\boldsymbol{\theta}_{\aleph}$ can be expressed as $\boldsymbol{\nabla} \cdot \log \boldsymbol{\pi}_{\aleph}$.*

Proof. For an $K_i \times K_i$ upper triangular matrix of 1s, its inverse, $\mathbf{B}_i^{-1}$, consists of K_i row vectors of $\mathbf{R}_i = (0, \ldots, 1, -1, 0, \ldots, 0), i < K_i$, with 1 in the ith location and $\mathbf{R}_{K_i} = (0, \ldots, 0, 1)$ as the last row. Therefore, each $\mathbf{B}_i^{-1}$ is equivalent to $\boldsymbol{\nabla}^i$, and $\mathbf{A}^{-1}$ is equivalent to $\boldsymbol{\nabla}$.

See also Ip et al. [13] for a proof using mathematical induction. In the above example, matrix $\mathbf{B}_2^{-1} \otimes \mathbf{B}_2^{-1} \otimes \mathbf{B}_1^{-1}$ has the same effect as vector $(\bigtriangledown_3, \vartheta)^{\mathrm{T}} \otimes (\bigtriangledown_2, \vartheta)^{\mathrm{T}} \otimes (\bigtriangledown_1, \bigtriangledown_1, \vartheta)^{\mathrm{T}} = (\bigtriangledown_3 \bigtriangledown_2 \bigtriangledown_1, \bigtriangledown_3 \bigtriangledown_2 \bigtriangledown_1, \bigtriangledown_3 \bigtriangledown_2 \vartheta, \bigtriangledown_3 \vartheta \bigtriangledown_1, \bigtriangledown_3 \vartheta \bigtriangledown_1, \bigtriangledown_3 \vartheta \vartheta, \vartheta \bigtriangledown_2 \bigtriangledown_1, \vartheta \bigtriangledown_2 \bigtriangledown_1, \vartheta \bigtriangledown_2 \vartheta, \vartheta \vartheta \bigtriangledown_1, \vartheta \vartheta \bigtriangledown_1, \vartheta \vartheta \vartheta)^{\mathrm{T}}$. There is a clear need to simplify notation and we abbreviate $\bigtriangledown_1 \bigtriangledown_2 \bigtriangledown_3 \vartheta \cdots \vartheta$, which is of length J, as $\bigtriangledown_{123}$. For example, matrix $\mathbf{B}_2^{-1} \otimes \mathbf{B}_2^{-1} \otimes \mathbf{B}_1^{-1}$ above is expressed as $(\bigtriangledown_{123}, \bigtriangledown_{123}, \bigtriangledown_{23}, \bigtriangledown_{13}, \bigtriangledown_{13}, \bigtriangledown_3, \bigtriangledown_{12}, \bigtriangledown_{12}, \bigtriangledown_2, \bigtriangledown_1, \bigtriangledown_1, \vartheta^3)^{\mathrm{T}}$. In addition, if X_i, $i = 1, 2, 3$ are binary variables that can take value 1 or 2, then

$$\begin{aligned} \bigtriangledown_{12} f(1,1,2) &= \bigtriangledown_1 \bigtriangledown_2 \vartheta f(1,1,2) \\ &= \bigtriangledown_1 [f(1,1,2) - f(1,2,2)] \\ &= f(1,1,2) - f(2,1,2) - [f(1,2,2) - f(2,2,2)]. \end{aligned}$$

Also, we denote $\vartheta^J = \vartheta \cdots \vartheta$ by $\bigtriangledown_{\emptyset}$. Note that by lexicographical arrangement, for a distribution such as (X_1, X_2, X_3), $\bigtriangledown_{12}$ only applies to $\log \pi_{k_1 k_2 K_3}$—that is, only when X_3 takes the value of the last category K_3. Thus, $\bigtriangledown_{\emptyset}$ only applies to $\log \pi_{K_1 \cdots K_J}$.

The canonical interactions $\boldsymbol{\theta}_{\aleph}$, or simply $\boldsymbol{\theta}$, consist of logarithms of local adjacent-category logits, local log odds ratios, log ratios of odds ratios, and so on. The following example provides an illustration.

Example 2 (Example 1 *continued*). Let $\boldsymbol{\pi} = (\pi_{11}, \pi_{21}, \ldots, \pi_{34})^{\mathrm{T}}$; $\boldsymbol{\nabla} = (\bigtriangledown_2, \bigtriangledown_2, \bigtriangledown_2, \vartheta)^{\mathrm{T}} \otimes (\bigtriangledown_1, \bigtriangledown_1, \vartheta)^{\mathrm{T}} = (\bigtriangledown_{12}, \bigtriangledown_{12}, \bigtriangledown_2, \bigtriangledown_{12}, \bigtriangledown_{12}, \bigtriangledown_2, \bigtriangledown_{12}, \bigtriangledown_{12}, \bigtriangledown_2, \bigtriangledown_1, \bigtriangledown_1, \vartheta^2 = \bigtriangledown_{\emptyset})^{\mathrm{T}}$, with $\boldsymbol{\theta}$ equivalent to $(\bigtriangledown_{12} \log \pi_{11}, \cdots, \bigtriangledown_{12} \log \pi_{23}, \bigtriangledown_2 \log \pi_{31}, \bigtriangledown_2 \log \pi_{32}, \bigtriangledown_2 \log \pi_{33}, \bigtriangledown_1 \log \pi_{14}, \bigtriangledown_1 \log \pi_{24}, \log \pi_{34})^{\mathrm{T}}$. As an example, $\bigtriangledown_{12} \log \pi_{11}$ is the log odds ratio $\log[(\pi_{11}\pi_{22})/(\pi_{12}\pi_{21})]$, and $\bigtriangledown_1 \log \pi_{14}$ is the consecutive log odds $\log \pi_{14}/\pi_{24}$.

The advantage of $\boldsymbol{\nabla} \cdot \log \boldsymbol{\pi}_{\aleph}$ over $\mathbf{A}^{-1} \log \boldsymbol{\pi}_{\aleph}$ is that individual interactions can be easily identified using the former representation. With the difference operator, each interaction can be identified via products of operators with the appropriate subscript. Ip and Wang [14] use the difference operator $\bigtriangledown_a$ for marginal modeling of contingency tables. Other applications of interactions in measurement theory can be found in Ip et al. [13].

From the above discussion, it is also evident that each interaction represents one degree of freedom. Lemma 1 directly leads to the following likelihood function in terms of interactions, and the interaction vector in Lemma 1 becomes the canonical parameter in the likelihood.

Lemma 2. *Let $X_j = k_j$, $1 \le k_j \le K_j$, $1 \le j \le J$ be represented by indicator variables $Y_{k_1, \ldots, k_J} = 1$, when $(X_1, \ldots, X_J) = (k_1, \ldots, k_J)$, and 0 otherwise. The loglikelihood function for $\mathbf{X}$ can be expressed as:*

$$\ell(\boldsymbol{\theta}) = \mathbf{s}^{\mathrm{T}} \boldsymbol{\theta} - \kappa(\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\theta} = \boldsymbol{\nabla} \cdot \log \boldsymbol{\pi}_{\aleph}$, $\mathbf{s}^{\mathrm{T}} = \mathbf{y}^{\mathrm{T}} \mathbf{A}$ and $\kappa(\boldsymbol{\theta}) = \log(\sum_{k=1}^{K} \exp(\mathbf{R}_k \boldsymbol{\theta}))$, with $\mathbf{R}_k$ being the kth row vector of $\mathbf{A}$.

Example 3. Consider four binary variables (X_1, X_2, X_3, X_4) with their interactions specified as follows for easy identification:

$$\begin{aligned} \boldsymbol{\theta}^{\mathrm{T}} &= (\bigtriangledown_{1234}, \bigtriangledown_{234}, \bigtriangledown_{134}, \bigtriangledown_{34}, \bigtriangledown_{124}, \bigtriangledown_{24}, \bigtriangledown_{14}, \bigtriangledown_4, \bigtriangledown_{123}, \bigtriangledown_{23}, \bigtriangledown_{13}, \bigtriangledown_3, \bigtriangledown_{12}, \bigtriangledown_2, \bigtriangledown_1, \bigtriangledown_{\emptyset}) \\ &= (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0). \end{aligned}$$

The **A** matrix of Lemma 2 is

$$\mathbf{A} = \begin{pmatrix} \mathbf{C} & \mathbf{C} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}, \text{ where}$$

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

After normalization of $\exp(\mathbf{A}\boldsymbol{\theta})$, the joint probability in lexicographical order is

$$\begin{aligned} \boldsymbol{\pi}_{1234}^{1} = (&.1432, .0697, .0725, .0486, .0786, .0506, .0516, .0406, \\ &.0999, .0571, .0582, .043, .0606, .044, .0445, .0375). \end{aligned}$$

Notice that we use 0 for $\nabla_{\emptyset}$ in $\boldsymbol{\theta}$ so that its right value, log(.0375), emerges after normalization. Individual probability π_k is computed via $\exp(\mathbf{R}_k\boldsymbol{\theta})/\sum_{k=1}^{K}\exp(\mathbf{R}_k\boldsymbol{\theta})$ so that the sum of π_k, $1 \le k \le K$ is always 1. Therefore, there is no restriction of values of canonical parameters of $\boldsymbol{\theta}$.

The following lemma identifies the shared interactions between a J-dimensional joint and its J-conditional distributions. First, denote the power set of $\aleph$ by $2^{\aleph}$.

Lemma 3 (*Invariance*). *For $a, b \in 2^{\aleph}$ such that $b \not\subset a$, $\nabla_b \log \boldsymbol{\pi}_{\aleph} = \nabla_b \log \boldsymbol{\pi}_{\bar{a}|a}$.*

Proof. Because $b \not\subset a$, there is a variable X_j such that $j \in b$, but $j \notin a$. Hence, we have $\nabla_j \log(\boldsymbol{\pi}_{\cdots j,\cdots}/\boldsymbol{\pi}_a) = \log(\boldsymbol{\pi}_{\cdots j,\cdots}/\boldsymbol{\pi}_a) - \log(\boldsymbol{\pi}_{\cdots j+1,\cdots}/\boldsymbol{\pi}_a) = \nabla_j \log \boldsymbol{\pi}_{\cdots j,\cdots}$, which implies that $\nabla_b \log \boldsymbol{\pi}_{\aleph} = \nabla_{b\setminus\{j\}}\nabla_j \log \boldsymbol{\pi}_{\aleph} = \nabla_{b\setminus\{j\}}\nabla_j \log \boldsymbol{\pi}_{\bar{a}|a}$.

Building upon this result, the next lemma shows that the set of invariant interactions, $\{\nabla_b \log \boldsymbol{\pi}_{\aleph}, b \not\subset a\}$, is equivalent to $\boldsymbol{\pi}_{\bar{a}|a}$. Two sets are said to be equivalent if all elements from any one set can be derived from the other set.

Lemma 4 (*Characterization*). *Let $a \subset \{1, \ldots, J\}$ be fixed. Probability distributions $\boldsymbol{\pi}_{\aleph}$, and $\boldsymbol{\pi}_{\bar{a}|a}$ are, respectively, characterized by the sets of interactions $\jmath = \{\nabla_b \log \boldsymbol{\pi}_{\aleph}, b \neq \emptyset\}$, and $\imath = \{\nabla_b \log \boldsymbol{\pi}_{\aleph}, b \not\subset a\}$.*

The proof is given in Appendix B. The sets $\jmath$ and $\imath$ are, respectively, called the CSOI of $\boldsymbol{\pi}_{\aleph}$ and $\boldsymbol{\pi}_{\bar{a}|a}$. We illustrate the Characterization Lemma with an example (a continuation of Example 1) that shows how $\mathbf{a}_{i|j} = \boldsymbol{\pi}_{1|2}$ can be derived from its CSOI:$\{\nabla_{12}, \nabla_1\}$.

Example 4 (Example 1 *continued*). The CSOI for $\boldsymbol{\pi}_{1|2}$ consists of $\nabla_1 \log a_{24} = \log(1/2)$, $\nabla_1 \log a_{14} = \log(1/2)$, and $\nabla_{12} \log a_{ij} = \log r_{ij}$, $1 \le i \le 2, 1 \le j \le 3$; whence, set the remaining interactions to zero. The vector of interactions is therefore $\boldsymbol{\lambda}' = (\log 1, \log(1/4), 0, \log(1/6), \log 6, 0, \log 6, \log(2/3), 0, \log(1/2), \log(1/2), 0)^{\mathrm{T}}$. Let (q'_{ij}) be $\exp(\mathbf{A}\boldsymbol{\lambda}')$ arranged in a matrix, which gives:

$$(q'_{ij}) = \begin{pmatrix} 1/4 & 1 & 1 & 1/4 \\ 1/2 & 2 & 1/3 & 1/2 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Normalizing each column of (q'_{ij}) gives the conditional distribution $\boldsymbol{\pi}_{1|2}$.

3. Compatibility and uniqueness conditions

The Invariance Lemma and Characterization Lemma facilitate checking the compatibility and uniqueness of conditional models. Compatible conditional distributions are those that can be derived from the same joint distribution:

Definition 1. Suppose a and c are both non-empty subsets of $\aleph$. A joint $\boldsymbol{\pi}_{\aleph}$ is said to be *capable of generating* $\boldsymbol{\pi}^1_{\bar{a}|a}$ if $\boldsymbol{\pi}^1_{\bar{a}|a} = \boldsymbol{\pi}_{\aleph}/\boldsymbol{\pi}_a$. Two families of J-conditionals, $\boldsymbol{\pi}^1_{\bar{a}|a}$ and $\boldsymbol{\pi}^2_{\bar{c}|c}$, are said to be *compatible* if there exists a joint $\boldsymbol{\pi}_{\aleph}$ that is capable of generating both families of J-conditionals. And $\boldsymbol{\pi}_{\aleph}$ is called a compatible joint distribution.

Compatibility becomes an issue only when some interactions are redundantly specified. Naturally, compatibility is affirmed when the redundant or overlapping interactions are identical.

Theorem 1 (*Compatibility*). *Let $\boldsymbol{\pi}^1_{\bar{a}|a}$ and $\boldsymbol{\pi}^2_{\bar{c}|c}$ be two conditional distributions having the same rectangular J-dimensional support, and let $\mathcal{A}$ and $\mathcal{C}$ denote the CSOI of $\boldsymbol{\pi}^1_{\bar{a}|a}$ and $\boldsymbol{\pi}^2_{\bar{c}|c}$, respectively. In other words, $\mathcal{A} = \{\nabla^1_b = \nabla_b \log \boldsymbol{\pi}^1_{\bar{a}|a}, b \not\subset a\}$ and $\mathcal{C} = \{\nabla^2_d = \nabla_d \log \boldsymbol{\pi}^2_{\bar{c}|c}, d \not\subset c\}$. Then, the two conditional distributions are compatible if and only if $\nabla^1_b = \nabla^2_d$ holds for $b = d$.*

Proof. If the given conditionals are compatible with some $\boldsymbol{\pi}^*_{\aleph}$, then, by the Invariance Lemma, $\nabla^1_b = \nabla_b \log \boldsymbol{\pi}^*_{\aleph}$ and $\nabla^2_d = \nabla_d \log \boldsymbol{\pi}^*_{\aleph}$. Hence, $\nabla^1_b = \nabla^2_d$ holds for $b = d$.

To prove the converse, consider the union $\mathcal{E} = \mathcal{A} \cup \mathcal{C}$. The condition ensures that every $\nabla_e \in \mathcal{E}$ is uniquely determined from either $\boldsymbol{\pi}^1$ or $\boldsymbol{\pi}^2$. If $\mathcal{E}$ does not have all the interactions (when $a \cap c \neq \emptyset$), we assign $\nabla_f = 0, f \subset a \cap c$ and let $\mathcal{F} = \mathcal{E} \cup \{\nabla_f\}$. By the Characterization Lemma, a joint distribution can be constructed from $\mathcal{F}$, and this joint is capable of generating both $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$ because its CSOI has both $\mathcal{A}$ and $\mathcal{C}$ as its subsets.

From Theorem 1, it is straightforward to derive the following corollary due to [2]:

Corollary 1. *The two-dimensional conditionals $\mathbf{a}_{i|j}$ of $(X_1|X_2)$ and $\mathbf{b}_{j|i}$ of $(X_2|X_1)$ are compatible if and only if $\nabla_{12} \log \mathbf{a}_{i|j} = \nabla_{12} \log \mathbf{b}_{j|i}$.*

Corollary 1 implies that, in the genotype example in Section 1, the constrained diagnostic conditional model and the treatment conditional model are only compatible when their cross-distribution odds ratios are consistent.

Definition 2. A given set of J-conditional distributions: $\{\boldsymbol{\pi}_{\bar{a}_i|a_i}, a_i \subset \aleph, 1 \le i \le l\}$ is said to be *complete for* $\aleph$ if every interaction ∇_b, $b \in 2^{\aleph}$, $b \neq \emptyset$ can be derived from one of the given J-conditionals. Otherwise, the given set of J-conditionals is said to be incomplete.

Theorem 2 (*Uniqueness*). *A given set of J-conditional distributions* $\{\boldsymbol{\pi}_{\bar{a}_i|a_i}, 1 \le i \le l\}$ *specifies a unique joint distribution* $\boldsymbol{\pi}_{\aleph}$ *if and only if* (1) *the set is complete for* $\aleph$, *and* (2) *every pair of conditional distributions are compatible.*

Proof. The proof is a direct application of the Characterization Lemma to $\mathcal{A} = \cup_{i=1}^{i=l}\{\nabla_b \log \boldsymbol{\pi}_{\bar{a}_i|a_i}, b \not\subset a_i\}$.

The following examples illustrate the use of Theorems 1 and 2.

Example 5. For the $\boldsymbol{\pi}^1_{1|2}$ and $\boldsymbol{\pi}^2_{2|1}$ with $\bar{a} = \{1\}$ and $\bar{c} = \{2\}$, their CSOIs are $\mathcal{A} = \{\nabla^1_{12}, \nabla^1_2\}$ and $\mathcal{C} = \{\nabla^2_{12}, \nabla^2_1\}$, respectively. From Theorem 1, $\boldsymbol{\pi}^1_{1|2}$ and $\boldsymbol{\pi}^2_{2|1}$ are compatible if and only if $\nabla^1_{12} = \nabla^2_{12}$.

Example 6. Consider checking the compatibility between $\boldsymbol{\pi}^1_{12|34}$ and $\boldsymbol{\pi}^2_{13|24}$ for $\aleph = \{1234\}$ with $\bar{a} = \{12\}$ and $\bar{c} = \{13\}$. Their CSOIs are $\mathcal{A} = \{\nabla^1_{1234}, \nabla^1_{123}, \nabla^1_{124}, \nabla^1_{134}, \nabla^1_{234}, \nabla^1_{12}, \nabla^1_{13}, \nabla^1_{14}, \nabla^1_{23}, \nabla^1_{24}, \nabla^1_1, \nabla^1_2\}$ and $\mathcal{C} = \{\nabla^2_{1234}, \nabla^2_{123}, \nabla^2_{124}, \nabla^2_{134}, \nabla^2_{234}, \nabla^2_{12}, \nabla^2_{13}, \nabla^2_{14}, \nabla^2_{23}, \nabla^2_{34}, \nabla^2_1, \nabla^2_3\}$, respectively. All of the interactions, except $\nabla^1_{24}, \nabla^1_2, \nabla^2_{34}, \nabla^2_3$, must be equal in order for $\boldsymbol{\pi}^1_{12|34}$ and $\boldsymbol{\pi}^2_{13|24}$ to be compatible.

In Example 6, the collection $\{\boldsymbol{\pi}_{12|34}, \boldsymbol{\pi}_{13|24}\}$ is not complete because interactions $\nabla_4 \log \pi_{K_1K_2K_3l}$, $1 \le l \le (K_4 - 1)$ are absent. When a given collection of conditional distributions are compatible but not complete, a unique family of conditional distributions, instead of the joint distribution, can be constructed. For example, a unique family of conditional distributions $\boldsymbol{\pi}_{123|4}$ can be derived from $\{\boldsymbol{\pi}_{12|34}, \boldsymbol{\pi}_{13|24}\}$ using the union of $\mathcal{A}$ and $\mathcal{C}$ in Example 6.

Theorem 3 (*Intersection*). *An incomplete set of conditional distributions* $\{\boldsymbol{\pi}_{\bar{a}_i|a_i}, 1 \le i \le l\}$ *specifies a unique conditional distribution* $\boldsymbol{\pi}_{\bar{\mathcal{B}}|\mathcal{B}}$, *where* $\mathcal{B} = \cap_{i=1}^{l} a_i$, *if and only if every pair of conditional distributions is compatible.*

Proof. The union of the CSOI of $\{\boldsymbol{\pi}_{\bar{a}_i|a_i}, 1 \le i \le l\}$ is $\mathcal{A} = \cup_{i=1}^{l}\{\nabla_b, b \not\subset a_i\} = \{\nabla_b, b \not\subset \cap_{i=1}^{l} a_i = \mathcal{B}\}$. Compatibility implies that every ∇_b for $b \not\subset \mathcal{B}$ is uniquely determined; whence, by Characterization Lemma 4, $\mathcal{A}$ determines a unique $\boldsymbol{\pi}_{\bar{\mathcal{B}}|\mathcal{B}}$.

There are several ways to supplement an incomplete set of conditionals so that a joint distribution can be obtained. For example, to supplement $(X_1, X_2|X_3)$, options include finding: (1) a marginal distribution of X_3, (2) the "missing" interaction ∇_3, or (3) a so-called posterior slice of X_3 in the form $P[X_3 = j|X_1 = K_1, X_2 = K_2]$, $1 \le j \le K_3$. From the Characterization Lemma, the last two options are clearly equivalent because ∇_3 is indeed the CSOI of $\boldsymbol{\pi}_{X_3|X_1=K_1,X_2=K_2}$.

Lemma 5 (*Posterior Slice*). *Let* $a = \{1, \ldots, m\}$ *be fixed. Conditional distributions* $\boldsymbol{\pi}^1_{\bar{a}|a}$ *and* $\boldsymbol{\pi}^2_{a|x_{m+1}=K_{m+1},\ldots,x_J=K_J}$ *determine a unique joint distribution.*

Proof. By Lemma 4, the CSOI of $\boldsymbol{\pi}^1_{\bar{a}|a}$ is $\mathcal{A} = \{\nabla_b \log \boldsymbol{\pi}^1_{\aleph}, b \not\subset a\}$. The CSOI of $\boldsymbol{\pi}^2_{a|x_{m+1}=K_{m+1},\ldots,x_J=K_J}$ is $\mathcal{C} = \{\nabla_c \log \boldsymbol{\pi}^2_{a|x_{m+1}=K_{m+1},\ldots,x_J=K_J}, c \subset a\}$. Thus, $\mathcal{A} \cup (\mathcal{C} \setminus \nabla_{\emptyset})$ is complete for $\aleph$ and hence it determines a unique joint distribution. Moreover, since $\mathcal{A}$ and $\mathcal{C}$ are indexed by disjoint sets, compatibility is automatic.

Example 7 (Example 3 *continued*). Suppose $(X_3, X_4|X_1, X_2)$ is known. (According to the Characterization lemma, it is uniquely determined by Example 3's $\boldsymbol{\theta}$ except $\nabla_a = 0$, $a \subset \{1, 2\}$ and its computation is outlined in Appendix). Its complementary posterior slice should be $(X_1, X_2|X_3 = 2, X_4 = 2)$. To verify this, we take $\nabla_{12}, \nabla_2, \nabla_1$ from $\boldsymbol{\theta}$ and add $\nabla_{\emptyset} = 0$ to form the CSOI of the posterior slice. Then, the interaction vector is $\boldsymbol{\lambda} = (\nabla_{12}, \nabla_2, \nabla_1, \nabla_{\emptyset}) = (0.15, 0.16, 0.17, 0)$. It follows that

$$\exp(\mathbf{B}_2\boldsymbol{\lambda}) = \exp(0.48, 0.16, 0.17, 0) = (1.616, 1.1735, 1.1853, 1),$$

where $\mathbf{B}_2$ is taken from Eq. (2). After normalization, the vector takes the value $(0.3248, 0.2358, 0.2383, 0.2010)$. This has exactly the same value as the distribution of $(X_1, X_2|X_3 = 2, X_4 = 2)$ derived from the joint probabilities: $(\pi^1_{1122}, \pi^1_{2122}, \pi^1_{1222}, \pi^1_{2222}) = (0.0606, 0.044, 0.0445, 0.0375)$.

The Posterior Slice Lemma can be useful in applications when marginal data are hard to get but data are available for a specific combination of values of the conditioning variables. Consider an example in which X indicates the presence or absence of one or more gene markers, and Z is disease statuses of a newly identified or rare disease. Suppose heredity theory suggests that $Z|X$ follows a specific multinomial distribution. In addition, only data for $Z = z_0$, and hence $X|Z = z_0$ – the conditional distribution of gene markers given a specific disease status – are available. This is not an uncommon occurrence in practice. Then, the two pieces of information, $Z|X$ and $X|Z = z_0$, can be combined to create a unique joint distribution for (X, Z).

4. Discussions and applications

Using mixed parameters, we offer a heuristic derivation for the Characterization Lemma. Afterward, we discuss two applications of the interaction-based approach.

4.1. Mixed parameters

Even though Eq. (1) is mathematically equivalent to the saturated loglinear model, there is a major difference: interactions $\boldsymbol{\theta}_{\aleph}$ by themselves are minimally sufficient parameters in the sense of [15], while loglinear interactions are not identifiable without additional constraints. To our knowledge, no log-linear model has been proposed for conditional probabilities. We consider $\boldsymbol{\theta}_{a|\bar{a}} = \boldsymbol{\nabla} \cdot \log \boldsymbol{\pi}_{a|\bar{a}}$ a log-linear decomposition for the J-conditional because $\boldsymbol{\pi}_{a|\bar{a}} = \exp(A\boldsymbol{\theta}_{a|\bar{a}})$.

We now use the mixed parametrizations of [10, p. 122], to justify COSI. Let $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$ and $(\mathbf{s}^{(1)}, \mathbf{s}^{(2)})$, respectively, denote matching partitions of the canonical interactions $\boldsymbol{\theta}$ and the canonical statistic $\mathbf{s}$ in Eq. (2). The mean parameters are $\boldsymbol{\mu}^{(i)} = E_{\boldsymbol{\theta}}\mathbf{s}^{(i)}$, $i = 1, 2$. A complementary mixed parameter $(\boldsymbol{\mu}^{(2)}, \boldsymbol{\theta}^{(1)})$ can be formed by "cross-breeding" $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$. Barndorff–Nielsen [10, p. 122] proved that the mixed parameter $(\boldsymbol{\mu}^{(2)}, \boldsymbol{\theta}^{(1)})$ uniquely specifies the joint distribution. The motivation of the Characterization Lemma is as follows: if we can make $\boldsymbol{\mu}^{(2)}$ equivalent to the marginal distribution $\boldsymbol{\pi}_a$, then $\boldsymbol{\theta}^{(1)}$ should characterize $\boldsymbol{\pi}_{\bar{a}|a}$, and hence be the CSOI. For illustration, consider the mixed parameter of a bivariate distribution $\boldsymbol{\pi}_{ij}$. When the canonical interactions are partitioned into $\boldsymbol{\theta}^{(1)} = \{\nabla_{12}, \nabla_1\}$, and $\boldsymbol{\theta}^{(2)} = \{\nabla_2, \vartheta\}$, the mean parameter corresponding to $\boldsymbol{\theta}^{(2)}$ is $\sum_{j=1}^{k} \boldsymbol{\pi}_{+j}$, $1 \le k \le J_2$—i.e., the cumulative marginal distribution X_2. Thus, $\boldsymbol{\theta}^{(1)}$ must characterize $\boldsymbol{\pi}_{1|2}$. A further partition of $\boldsymbol{\theta}^{(1)}$ into $\{\nabla_{12}\}$ and $\{\nabla_1\}$, and a second application of mixed parametrization, leads to the well-known specification of $\boldsymbol{\pi}_{ij}$ in terms of $\boldsymbol{\pi}_{i+}$, $\boldsymbol{\pi}_{+j}$, and ∇_{12}.

In the literature, almost all of the theoretical developments of conditional models have focused on the mean parameters (probabilities). For example, to find a joint distribution from $\boldsymbol{\pi}_{\bar{a}|a}$ and $\boldsymbol{\pi}_{\bar{c}|c}$, Arnold et al. [16] propose finding marginal distributions $\boldsymbol{\pi}_a^1$ and $\boldsymbol{\pi}_c^2$ by solving the constrained linear equations $\boldsymbol{\pi}_{\bar{a}|a}\boldsymbol{\pi}_a^1 = \boldsymbol{\pi}_{\bar{c}|c}\boldsymbol{\pi}_c^2$, which have a total of K equations, where $K - 1$ is the degree of freedom of the joint. Because the combined number of cells in $\boldsymbol{\pi}_a^1$ and $\boldsymbol{\pi}_c^2$ is often much fewer than K, the system of equations is over-specified. Furthermore, probabilities must be non-negative. All of these restrictions often make the linear-equation approach a computationally challenging endeavor. The advantage of the proposed approach is partly due to its invariance property (Lemma 3).

4.2. Besag's compatibility check

There is a connection between the compatibility check of [17, explained below] and Theorem 1. Consider the case of three binary variables; check compatibility for $\{\boldsymbol{\pi}_{1|23}^{(1)}, \boldsymbol{\pi}_{2|13}^{(2)}, \boldsymbol{\pi}_{3|12}^{(3)}\}$. According to [17], there are $3! = 6$ different paths directed from $(1, 1, 1)$ to $(2, 2, 2)$. For example, $R_1 : (1, 1, 1) \to (2, 1, 1) \to (2, 2, 1) \to (2, 2, 2)$. Besag defined the following ratio for this specific path: $\mathrm{BR}_1 = [\boldsymbol{\pi}^{(1)}(1|11)\boldsymbol{\pi}^{(2)}(1|21)\boldsymbol{\pi}^{(3)}(1|22)]/[\boldsymbol{\pi}^{(1)}(2|11)\boldsymbol{\pi}^{(2)}(2|21)\boldsymbol{\pi}^{(3)}(2|22)]$. Without a formal proof, he argued that if the all six ratios are the same, then the three conditional distributions are compatible. From the CSOI perspective, the ratio BR_1 is equal to $\nabla_{123}^{(1)} + \nabla_{13}^{(1)} + \nabla_{12}^{(1)} + \nabla_1^{(1)} + \nabla_{23}^{(2)} + \nabla_2^{(2)} + \nabla_3^{(3)}$. The superscript indicates the conditional distribution from which the ∇ is derived. For another path, $R_2 : (1, 1, 1) \to (2, 1, 1) \to (2, 1, 2) \to (2, 2, 2)$, the Besag's ratio is $\mathrm{BR}_2 = [\boldsymbol{\pi}^{(1)}(1|11)\boldsymbol{\pi}^{(3)}(1|21)\boldsymbol{\pi}^{(2)}(1|22)]/[\boldsymbol{\pi}^{(1)}(2|11)\boldsymbol{\pi}^{(3)}(2|21)\boldsymbol{\pi}^{(2)}(2|22)]$, which is equal to $\nabla_{123}^{(1)} + \nabla_{13}^{(1)} + \nabla_{12}^{(1)} + \nabla_1^{(1)} + \nabla_{23}^{(3)} + \nabla_3^{(3)} + \nabla_2^{(2)}$. Therefore, $\mathrm{BR}_1 = \mathrm{BR}_2$ if and only if ∇_{23} of $\boldsymbol{\pi}^{(2)}$ is equal to the ∇_{23} of $\boldsymbol{\pi}^{(3)}$. When all six of Besag's ratios are equal, the overlapping interactions are identical, and thus compatible. It can be shown that the number of checks of Theorem 1 is also $J!$. Again, Besag [17] operated with the mean parameters, while Theorem 1 relies on the canonical parameters.

4.3. Near-compatible conditional models

When the conditional models are incompatible, the existence of a joint distribution is jeopardized. Under such a circumstance, one alternative is to find a so-called near-compatible joint distribution. There has been growing interest in finding joint distributions that are close to the specified (incompatible) conditional distributions [6,18,11]. Two related areas of active research regarding incompatible conditional distributions include Markov Chain Monte Carlo and Gibbs sampling [21], and multiple imputation [19,20].

There are three areas that CSOIs are uniquely suitable to address in the realm of near-compatible conditional models. The first is to devise a measure to quantify the nearness of a collection of conditionals to a compatible joint model. The second is to identify the source of incompatibility in terms of their CSOIs. The third is to search for a solution by modifying the source of incompatibility.

To illustrate this approach, we modify the conditional probabilities in Example 1 so that they become incompatible. Specifically, $(\mathbf{b}_{j|i}(1, 1), \mathbf{b}_{j|i}(2, 1))$ are changed from the original $(1/6, 1/6)$ to $(1/4, 1/12)$ (both italicized):

$$\mathbf{a}_{i|j} = \begin{pmatrix} 1/7 & 1/4 & 3/7 & 1/7 \\ 2/7 & 2/4 & 1/7 & 2/7 \\ 4/7 & 1/4 & 3/7 & 4/7 \end{pmatrix}, \qquad \mathbf{b}_{j|i} = \begin{pmatrix} \textit{1/4} & \textit{1/12} & 3/6 & 1/6 \\ 2/7 & 2/7 & 1/7 & 2/7 \\ 2/6 & 1/12 & 1/4 & 1/3 \end{pmatrix}.$$

The corresponding cross-distribution odds ratios (see Example 1) for $\mathbf{a}_{i|j}$ and $\mathbf{b}_{j|i}$ are, respectively,

$$(r_{ij}^{\mathbf{a}}) = \begin{pmatrix} 1 & 1/6 & 6 \\ 1/4 & 6 & 2/3 \end{pmatrix}, \qquad (r_{ij}^{\mathbf{b}}) = \begin{pmatrix} 3 & \textit{1/12} & 6 \\ 1/4 & 6 & 2/3 \end{pmatrix}.$$

Table 1
Errors from using different combinations of interactions for incompatible joint distributions. The L_2 error is $\sum(\boldsymbol{\pi}_{i|j} - \mathbf{a}_{i|j})^2 + \sum(\boldsymbol{\pi}_{j|i} - \mathbf{b}_{j|i})^2$. Pseudo-Gibbs1 started the Gibbs sampling chain with $\mathbf{b}_{j|i}$, and pseudo-Gibbs2 started with $\mathbf{a}_{i|j}$.

	r^{π}_{11}	r^{π}_{12}	L_2 error
$(r^{\mathbf{a}}_{11}, r^{\mathbf{a}}_{12})$	1	1/6	0.03472
$(r^{\mathbf{a}}_{11}, r^{\mathbf{b}}_{12})$	1	1/12	0.09114
$(r^{\mathbf{b}}_{11}, r^{\mathbf{a}}_{12})$	3	1/6	0.05851
$(r^{\mathbf{b}}_{11}, r^{\mathbf{b}}_{12})$	3	1/12	0.02294
Arithmetic mean	2	1/8	0.01888
Geometric mean	1.732	0.1179	0.00981
Pseudo-Gibbs1	1.009	0.285	0.0115
Pseudo-Gibbs2	3.96	0.203	0.0333

A conventional approach for finding a joint that is least at variance with the given conditional probabilities is to minimize an objective function such as L_2 error or Kullback–Leibler pseudo-distance [7, p. 30–36]. Any probability-based minimization will average out the discrepancies over the entire support, and there is no way to preserve the interactions that are already in agreement. An interaction-based alternative is to minimize the same objective function, while retaining the consistent interactions during minimization. For the above incompatible example, one alternative is to specify the joint probabilities, π_{ij}, such that they satisfy $r^{\pi}_{ij} = (\pi_{ij}\pi_{i+1j+1})/(\pi_{i+1j}\pi_{ij+1}) = \alpha r^{\mathbf{a}}_{ij} + (1-\alpha) r^{\mathbf{b}}_{ij}$. Therefore, $r^{\pi}_{11} = \alpha + 3(1-\alpha)$, $r^{\pi}_{12} = \alpha'/6 + (1-\alpha')/12$, and $r^{\pi}_{ij} = r^{\mathbf{a}}_{ij}$, otherwise. This approach not only reduces the dimension of minimization from 12 to 2 (i.e., α and α') but also alleviates the constraints placed on the minimization, because the ranges of r^{π}_{ij} are basically unconstrained.

Table 1 shows the L_2 errors for different combinations of r^{π}_{11} and r^{π}_{12} of $\boldsymbol{\pi}_{ij}$, in which the other necessary interaction terms remain unchanged, using their values from either $\mathbf{a}_{i|j}$ or $\mathbf{b}_{j|i}$. The first four entries in Table 1 represent the four combinations of the $r^{\mathbf{a}}_{11}$, $r^{\mathbf{a}}_{12}$, $r^{\mathbf{b}}_{11}$ and $r^{\mathbf{b}}_{12}$, which are considered as the four corners of the range for $(r^{\pi}_{11}, r^{\pi}_{12})$. Hence, their errors will tend to be large. The next two entries employ the component-wise arithmetic means, $((1+3)/2, (1/6+1/12)/2)$, and geometric means, $(\sqrt{3}, \sqrt{1/72})$, each of which is considered to be a compromise between the two incompatible distributions. The last two entries are directly derived from the pseudo-Gibbs sampler, which will be discussed next.

Incompatible conditional specifications are common phenomena in dependence networks; Heckerman et al. [11] coined the term pseudo-Gibbs sampling for conditional probabilities "without [necessarily] respecting the consistency constraints". Also, see Arnold et al. [18]. To make comparisons between the pseudo-Gibbs and the CSOI approach, we generate joint distributions $\boldsymbol{\pi}^*$ by applying pseudo-Gibbs samplers between the incompatible $\mathbf{a}_{i|j}$ and $\mathbf{b}_{j|i}$. The numerical results of the Gibbs sampler are summarized as follows. (1) $\boldsymbol{\pi}^*$ depends upon the starting value—that is, whether $\mathbf{a}_{i|j}$ or $\mathbf{b}_{j|i}$ (correspondingly pseudo-Gibbs1 and Gibbs2) was used to start the chain. (2) Their L_2 errors are, respectively, 0.0115 and 0.03325. (3) Their respective CSOIs are log(1.009, 0.285, 0.15, 5.98, 6.06, 0.64, 0.52, 0.48, 4.42, 0.31, 0.76) and log(3.96, 0.203, 0.067, 6.81, 6.33, 0.62, 0.49, 0.51, 4.16, 0.32, 0.76). The pseudo-Gibbs sampler changes the entire set of interactions, so none of the originally consistent interactions ($r_{ij}, i \neq 1$ and $j \neq 1, 2$) are preserved. Moreover, the $\bigtriangledown_1$ of $\mathbf{a}_{i|j}$ and $\bigtriangledown_2$ of $\mathbf{b}_{j|i}$ are altered by the Gibbs samplers even though they do not cause incompatibility. (4) None of the pseudo-Gibbs can outperform the geometric mean—that is, $(\sqrt{3}, \sqrt{1/72})$ gives the smallest error. Geometric means are equivalent to averages of the overlapping interactions. The results from Table 1, while limited, suggest that estimating the joint distribution from some average of the incompatible interactions while keeping the consistent portion intact may be a reasonable alternative to the brute force (pseudo-) Gibbs sampler. Further study in this direction may offer fruitful opportunities.

4.4. Conclusion

Gelman and Raghuanthan [21] noted that "the study of conditional distributions is an area where theory has not caught up with practice." It is our hope that the theoretical work reported here will contribute to some of the central issues such as invariance, characterization, and solving the incompatibility of conditionally specified models.

Acknowledgments

The authors would like to express their gratitude toward the three referees of this article and Dr. Beverly Snively for their helpful comments and suggestions.

Appendix. Proof of the characterization lemma

(I) From probabilities to interactions. By Lemma 1, the case for $\jmath$ is trivial. For $\imath$, arrange $\log \boldsymbol{\pi}_{\bar{a}|a}$ in the same lexicographical order as $\boldsymbol{\pi}_{\aleph}$, and consider $\bigtriangledown \cdot \log \boldsymbol{\pi}_{\bar{a}|a}$. Deleting $\bigtriangledown_b \log \boldsymbol{\pi}_{\bar{a}|a}$, $b \subset a$ from $\bigtriangledown \cdot \log \boldsymbol{\pi}_{\bar{a}|a}$ gives the set $\imath$.

(II) From interactions to probabilities. We first describe how to construct the joint distribution from $\jmath$. Add $\bigtriangledown_{\emptyset} = 0$ to $\jmath$ to provide a complete set of interactions, then arrange the interactions of $\jmath \cup \{\bigtriangledown_{\emptyset}\}$ into a vector lexicographically and call it

$\boldsymbol{\lambda}$. Let $\mathbf{Q} = \exp(\mathbf{A}\boldsymbol{\lambda}) = (q_{k_1,\cdots,k_J})$. Normalizing the vector gives the following joint probabilities:

$$\boldsymbol{\pi}_{\aleph}(k_1,\ldots,k_J) = q_{k_1,\ldots,k_J} \Bigg/ \left(\sum_{i_1=1}^{K_1} \cdots \sum_{i_J=1}^{K_J} q_{i_1,\ldots,i_J} \right).$$

Because $\boldsymbol{\pi}_{\aleph}$ and $\mathbf{Q}$ only differ by a constant multiplier, they have the same set of interactions. In other words, $\jmath = (\triangledown \cdot \log \mathbf{Q}) \setminus \triangledown_{\emptyset}$, and $\boldsymbol{\pi}$ is the required joint.

Analogously, the family of conditional probabilities $\boldsymbol{\pi}_{\bar{a}|a}$ can be constructed from $\imath$. Assume without loss of generality that $a = \{1,\ldots,m\}$. For each value of $(X_1,\ldots,X_m) = (i_1,\ldots,i_m)$, add $\triangledown_e = 0$ for all $e \subset a$ to $\imath$ to form a complete set of interactions. Arrange each set of interactions lexicographically into a $K \times 1$ vector $\boldsymbol{\lambda}'$. Compute $\mathbf{Q}' = \exp(\mathbf{A}\boldsymbol{\lambda}') = (q'_{k_1,\ldots,k_J})$. Normalize $(q'_{k_1,\ldots,k_J})$ for each $(i_1,\ldots,i_m)$ to obtain the following family of conditional densities:

$$\boldsymbol{\pi}_{\bar{a}|(x_1,\ldots,x_m)=(i_1,\ldots,i_m)}(i_{m+1},\ldots,i_J) = q'_{i_1,\ldots,i_J} \Bigg/ \left(\sum_{l_{m+1}=1}^{K_{m+1}} \cdots \sum_{l_J=1}^{K_J} q'_{i_1,\ldots,i_m,l_{m+1},\cdots,l_J} \right).$$

Because all the supplemented interactions, $\triangledown_e$, are constant, they will not affect the values of the interaction terms in $\imath$. We can then follow the same proof of the Invariance Lemma to show that $\triangledown_b \log \boldsymbol{\pi}_{\bar{a}|a} = \triangledown_b \log \mathbf{Q}' = \triangledown_b \log \boldsymbol{\pi}_{\aleph}$, for every $b \not\subset a$.

For example, the conditional distribution of $(X_3, X_4|X_1, X_2)$ of Example 7 is computed using the same $\mathbf{A}$ matrix of Example 3 with the interaction vector being

$$\begin{aligned}\boldsymbol{\lambda}'^{\mathrm{T}} &= (\triangledown_{1234}, \triangledown_{234}, \triangledown_{134}, \triangledown_{34}, \triangledown_{124}, \triangledown_{24}, \triangledown_{14}, \triangledown_{4}, \triangledown_{123}, \triangledown_{23}, \triangledown_{13}, \triangledown_{3}, \triangledown_{12}, \triangledown_{2}, \triangledown_{1}, \triangledown_{\emptyset}) \\ &= (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.11, 0.12, 0.13, 0.14, 0, 0, 0, 0).\end{aligned}$$

The four 0s in $\boldsymbol{\lambda}$ represent the four normalizing steps, one for each combination of $(X_1 = i, X_2 = j)$.

References

[1] J.E. Besag, Spatial interaction and the statistical analysis of lattice systems (with discussions), J. Roy. Statist. Soc. B 36 (1974) 192–236.
[2] B.C. Arnold, S. Press, Compatible conditional distributions, J. Amer. Statist. Assoc. 84 (1989) 152–156.
[3] J.P. Hobert, G. Casella, Functional compatibility, Markov chains, and Gibbs sampling with improper posteriors, J. Comp. Graphical Statist. 7 (1998) 42–60.
[4] A. Gelman, T.P. Speed, Characterizing a joint probability distribution by conditionals, J. Roy. Statist. Soc. B 55 (1993) 185–188.
[5] A. Gelman, T.P. Speed, Corrigendum: Characterizing a joint probability distribution by conditionals, J. Roy. Statist. Soc. B 61 (1999) 483.
[6] B.C. Arnold, E. Castillo, J.M. Sarabia, Conditionally specified distributions: An introduction (with discussions), Statist. Sci. 16 (2001) 249–274.
[7] B.C. Arnold, E. Castillo, J.M. Sarabia, Conditional Specification of Statistical Models, Springer, New York, 1999.
[8] B.C. Arnold, E. Castillo, J.M. Sarabia, Compatibility of partial or complete conditional probabilities specifications, J. Statist. Plann. Inference 123 (2004) 133–159.
[9] G. Toffoli, E. Cecchin, G. Corona, A. Russo, A. Buonadonna, M. D'Andrea, L.M. Pasetto, S. Pessa, D. Errante, V. De Pangher, M. Giusto, M. Medici, F. Gaion, P. Sandri, B. Galligioni, S. Bonura, M. Boccalon, P. Biason, S. Frustaci, The role of UGT1A1*28 polymorphism in the pharmacodynamics and pharmacokinetics of irinotecan in patients with metastatic colorectal cancer, J. Clin. Oncol. 24 (2006) 3061–3068.
[10] O.E. Barndorff-Nielsen, Information and Exponential Families in Statistical Theory, John Wiley, New York, 1978.
[11] D. Heckerman, D.M. Chickering, C. Meek, R. Rounthwaite, C. Kadie, Dependence networks for inference, collaborative filtering and data visualization, J. Machine Learning 1 (2000) 49–57.
[12] J. Whitaker, Graphical Models in Applied Mathematical Multivariate Statistics, Wiley, West Sussex, U.K., 1990.
[13] E.H. Ip, Y. Wang, P. De Boeck, M. Meulders, Locally dependent latent trait models for polytomous responses with application to inventory of hostility, Psychometrika 69 (2004) 191–216.
[14] E.H. Ip, Y. Wang, A strategy for designing telescoping models for analyzing multiway contingency tables using mixed parameters, Sociol. Methods Res. 31 (2003) 291–324.
[15] E.W. Barankin, Sufficient parameters: solution of minimal dimensionality problem, Ann. Inst. Statist. Math. 12 (1961) 91–118.
[16] B.C. Arnold, E. Castillo, J.M. Sarabia, Specification of distributions by combinations of marginal and conditional distributions, Statist. Probab. Lett. 26 (1996) 153–157.
[17] J.E. Besag, Discussion of Markov chains for exploring posterior distributions, Ann. Statist. 22 (1994) 1734–1741.
[18] B.C. Arnold, E. Castillo, J.M. Sarabia, Exact and near compatibility of discrete conditional distributions, Comput. Statist. Data Anal. 40 (2002) 231–252.
[19] D.B. Rubin, Nested multiple imputation of NMES via partially incompatible MCMC, Statistica Neerlandica 57 (2003) 3–18.
[20] S. Van Buuren, Multiple imputation of discrete and continuous data by fully conditional specification, Statist. Methods Med. Res. 16 (2007) 219–242.
[21] A. Gelman, T.E. Raghuanthan, Comment of Conditionally specified distributions: An introduction (with discussions), Statist. Sci. 16 (2001) 268–269.