Statistics and Its Interface Volume 7 (2014) 251–261

Canonical ensembles for potentially incompatible dependency networks with applications to medical data

Shyh-Huei Chen*, Edward H. Ip†, and Yuchung J. Wang

A directed graph is either acyclic or cyclic. This paper focuses on the cyclic model, or dependency network, which represents a collection of univariate conditional distributions. The conditional approach allows a high level of flexibility in modeling because the dependency network is based on the notion that it is computationally convenient to estimate the local distribution of a variable given the remaining variables in a data set. However, the collection of conditional distributions individually estimated within a dependency network is generally not coherent with any joint distribution. The pseudo-Gibbs sampler (PGS) has often been used to estimate joint distributions for incompatible conditional models. We propose a new method for deriving a joint distribution from a given set of potentially incompatible univariate-conditional distributions such that the discrepancies between the given conditional distribution and those computed from the estimated joint distribution is minimized. The method is based on an ensemble of distributions, each of which can be derived from the canonical parameters of a set of given conditional distributions. Through simulation experiments and real data sets, we compare the performance of the ensemble method, the PGS, and a linear programming (LP)-based method. Our comparisons suggest that the ensemble method outperforms both the PGS and LP. The ensemble method is computationally efficient and scalable, and it therefore has the potential to open a new avenue for finding a nearly optimal solution for dependency networks of high dimensions.

Keywords and phrases: Characterizing set of interactions, Conditionally specified model, Dependency network, Ensemble method.

1. INTRODUCTION

Conditional reasoning is commonly used by physicians. An application of conditional reasoning is medical diagnosis in which the probability of a specific disease can be stated conditionally on results from various clinical tests and assessed risk factors. The advent of new graphical modeling tools facilitates the use of conditional reasoning in probabilistic terms by physicians. In directed acyclic-graph (DAG) modeling [25, 20], for example, symptoms and test results are often presented as parental nodes that have direct causal relationships with diseases, which are presented as child nodes. Diagnostic systems based on DAGs can become highly complex and computationally burdensome when a large number of parental nodes are involved in statistical learning and inference. Learning the structure of the DAG is also a challenging problem that is not completely solved [11]. Because DAG does not allow feedback from child nodes Heckerman *et al.* [16] argued for the use of dependence network (DN) in machine learning. Graphically a DN is represented by a directed cyclic-graph (DCG) which is in essence a collection of conditionally specified models (CSMs). Briefly, a DN can be built using a two-step approach: (1) creating conditional models for each individual variable given the remaining variables, and (2) "gluing" the conditional specified models together to form the joint density. As an example, consider three variables (X, Y, Z). In step (1), suppose three separate regression models—say using backward variable selection—are built: $p(X|Y, Z)$, $p(Y|Z)$, and $p(Z|X)$. The DN can now be represented by the graph in Figure 1. Unlike DAGs, DNs allow loops such as $X \to Z \to Y \to X$ in Figure 1. An important advantage of a DN over a DAG is its flexibility and the relative convenience in specifying conditional models for the variables one at a time. For step (2), pseudo-Gibbs sampling (PGS) has been suggested for computing the joint density [16]. In the above example, the PGS would iteratively sample from the following distributions: the first, y^1 from $p(Y|Z = z^0)$, x^1 from $p(X|z = z^0, Y = y^1)$, and then z^1 from $p(Z|X = x^1)$. The collection (x^1, y^1, z^1) forms a Gibbs sample for the DN. Some extensions and applications of DNs can be found in [19, 7, 12, 9].

One problem for DN or directed cyclic-graphical model in general is the possibility of incompatible conditional models. In the context within directed cyclic-graphs, Gelman and Raghuanthan [15] stated that "in general, reasonable-seeming conditional models will not be compatible with any single joint distribution." Heckerman *et al.* [16] argued that for large samples in a DN, the extent of incompatibility is negligible and can be ignored, and the PGS solution asymp-

*Corresponding author.
†Edward Ip's research was supported by NIH grants 1R21AG042761-01 and 1U01HL101066-01.

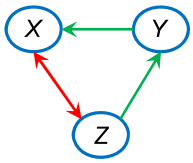

Figure 1. A Dependency network.

totically converges to a distribution that is consistent with the conditionally specified distributions (see also [28]). However, it is not clear how the procedure would perform when the sample size is small or moderate. There were alternative methods in the literature for handling potentially incompatible conditional distributions. For example, Arnold *et al.* [1, 2] proposed linear programming methods to find optimal solutions for the joint distribution. Indeed, the problem of making use of potentially incompatible CSMs has intrigued researchers across many areas of studies, including imputation of missing values [23, 28] and spatial statistics [4]. For a general review of CSMs, see [1, 2].

Several important questions seemed to pervade the discussion of potentially incompatible CSMs: (1) What are the conditions for a set of CSMs to be compatible? (2) If the CSMs are not compatible, then what is a reasonable solution for computing the joint distribution? and, (3) what price does one pay for using a "compromised" solution when no exact solution exists?

In this paper, we focus on the last two issues and propose an ensemble-based method for solving the joint density when the constituent CSMs of a DN are potentially incompatible. Ensemble methods are machine-learning approaches that are well studied especially in supervised learning [6]. The idea underlying the ensemble method is when a large number of weak classifiers are combined together, they can outperform a single strong classifier. In ensemble-based supervised learning, decisions from individual classifiers are generally combined through a weighted voting scheme, and the weight could be a function of some variance measure. Here we propose an ensemble of estimated joint distributions to reduce the risk of using a single estimate that might only do well in one part of the distributional space but not necessarily in other parts. Each estimated joint probability density function (pdf) within the ensemble is constructed by an algorithm that requires averaging over "overlapping" interaction terms, or the canonical parameters, of the CSMs. One of the important advantages of the proposed approach is its high efficiency: the canonical parameters can be extracted from the CSMs through simple arithmetic operations [17] and the computation of the joint density requires simple averaging procedures. Compared to other methods including linear programming and PGS, the canonical-ensemble method is easier to program as well as implement. More important, as we shall see later, the performance of this method, as measured by several divergence criteria, is superior to the other methods.

The remainder of the article is organized as follows. In §2 we present the necessary background to the approach. We first present an example to illustrate several key concepts—the canonical interaction terms, or the characterizing set of interaction (CSOI), for a collection of CSMs; an estimate of the joint distributions from the CSOI; and the ensemble of the joint distributions. We also define several divergence measures that are used as performance metrics for comparing different approaches. Results from two simulation studies follow in §3. Section §4 features two real examples of DNs to illustrate the canonical-ensemble approach—the first one regarding the adverse reactions to drug treatment in cancer patients of different genotypes, and the second regarding symptom structure. Finally in §5 we provide a brief discussion.

2. BACKGROUND FOR CANONICAL ENSEMBLE

Perhaps the most efficient way to understand the several key concepts underlying the canonical-ensemble (CE) method is through a few simple examples. In this paper we will focus on discrete distributions. The general theory of the concepts including CSOI and the creation of candidate ensembles in general is included in the Appendix.

Example 1a (Compatible conditionals)**.** Consider the following conditional distributions that are derived from a compatible joint distribution (x, y) in which both are discrete variables and take values from the set $\{1, 2\}$:

$$(1) \qquad f_{x|y} = \left[\begin{array}{c:c} \frac{1}{4} & \frac{1}{3} \\ \frac{3}{4} & \frac{2}{3} \end{array}\right], \; f_{y|x} = \left[\begin{array}{c:c} \frac{1}{3} & \frac{2}{3} \\ \frac{3}{7} & \frac{4}{7} \end{array}\right].$$

Thus, $P(x = 2|y = 1) = 3/4$.

Computation of the CSOI The CSOI is a generalized version of the canonical interaction parameters of a single distribution (cite earlier work). For a bivariate binary distribution (x, y), $x, y = 1, 2$, the canonical interaction parameter is the log odds $\log p(x = 1, y = 1) - \log p(x = 1, y = 2)$, $\log p(x = 2, y = 1) - \log p(x = 2, y = 2)$, and the log odds ratio $\log p(x = 1, y = 1) - \log p(x = 1, y = 2) - \log p(x = 2, y = 1) + \log p(x = 2, y = 2)$. In general, the canonical parameters can be formed by applying a difference operator to the log probability space. The CSOI of the collection of conditional distributions generalizes this concept and applies the difference operator to multiple probability space across conditional distributions.

To compute the CSOI for *Example 1a*, define a matrix B that indicates the dominance structure within x, and the matrix A that defines the ordered dominance relationship for (x, y):

$$B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \; A = B \otimes B = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $\otimes$ is the Kronecker product [17]. The B matrix represents the ordering among the two values of x (since y has two values also, so the ordering among its values is also B). For bivariate values, the ordering among (x_i, y_j) is represented by $B \otimes B$. If both x and y take values from the set $\{1, 2\}$, then the incidence matrix A indicates that, for example, the response $(x = 2, y = 1)$ dominates the response $(x = 1, y = 1)$ (entry $(1, 2) = 1$), but does not dominate the response $(x = 1, y = 2)$ (entry $(3, 2) = 0$). In general, the order in which the combination of variables is arranged follows the lexigraphical order in which the first index changes fastest and the last index changes slowest (e.g., in a 2×2 case, it is 11,21,12,22). The CSOI can be computed by first vectorizing the conditional distributions (in lexigraphical order) and then left multiply the log of the resulting vectors by A^{-1}. For example, the CSOI for $f_{x|y}$ is given by

$$\begin{aligned} & A^{-1} \log(f_{x|y}) \\ = & \begin{bmatrix} 1 & -1 & -1 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1.39 \\ -0.29 \\ -1.10 \\ -0.41 \end{bmatrix} = \begin{bmatrix} -0.41 \\ 0.12 \\ -0.69 \\ -0.41 \end{bmatrix}. \end{aligned}$$

Similarly, the CSOI for $f_{y|x}$ is given by $(-0.41, -0.29, 0.15, -0.56)$.

The matrix A^{-1} indeed acts as difference operator $\boldsymbol{\nabla} = (\nabla_{12}, \nabla_2, \nabla_1, \nabla_\vartheta)^T$ such that

$$\begin{aligned} & \nabla_1 \log p(x = 1|y = 1) \\ = \; & \log p(x = 1|y = 1) - \log p(x = 2|y = 1), \\ & \nabla_2 \log p(x = 1|y = 1) \\ = \; & \log p(x = 1|y = 1) - \log P(x = 1|y = 2), \quad \text{and} \\ & \nabla_{12} = \nabla_1 \nabla_2 \log p(x = 1|y = 1) \\ = \; & \nabla_1 [\log p(x = 1|y = 1) - \log p(x = 1|y = 2)] \\ = \; & \log p(x = 1|y = 1) - \log p(x = 2|y = 1) \\ & - \log p(x = 1|y = 2) + \log p(x = 2|y = 2), \end{aligned}$$

and ∇_ϑ is the identity operator. Thus, the value -0.41 in either CSOI represents a log odds ratio across conditional distributions. Ip and Wang [17] showed that $\nabla_{12} \log f_{x|y} = \nabla_{12} \log f_{y|x}$ is a necessary and sufficient condition for the two conditional distributions to be compatible, which is defined as the existence of a joint distribution that is capable of generating the given conditionals.

Recovery of the joint distribution Denote the CSOI for $x|y$ by $(\nabla_{12}^{x|y}, \nabla_2^{x|y}, \nabla_1^{x|y}, \nabla_\vartheta^{x|y})^T$ and similarly denote the CSOI for $y|x$ by $(\nabla_{12}^{y|x}, \nabla_2^{y|x}, \nabla_1^{y|x}, \nabla_\vartheta^{y|x})^T$. To "recover" the joint distribution, we follow these steps: (i) "Cross-select" the interaction terms $\theta_{12} = (\nabla_{12}^{x|y}, \nabla_2^{y|x}, \nabla_1^{x|y}, 0) = (-0.41, -0.29, -0.69, 0)$. The term derived from $\nabla_\vartheta^{\cdot|\cdot}$ is replaced by 0 in step (i), as there are only 3 degrees of freedom in the joint distribution. (ii) Left multiply the vector by A, which gives $A\theta_{12} = (-1.39, -0.29, -0.69, 0)$. (iii) Exponentiate the vector $A\theta_{12}$, and we have $\exp(A\theta_{12}) = (0.25, 0.75, 0.5, 1)$. (iv) Normalize the exponentiated vector to derive the vectorized joint pdf $(0.1, 0.3, 0.2, 0.4)$, and (v) unwind the vectorizing in lexigraphical order:

$$f_{xy} = \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix}. \tag{2}$$

Formation of a canonical ensemble One method to create an ensemble from the conditional distributions is to systemically interchange rows and columns. For example, one can create a variant of the given conditional distributions in (1) by interchanging their rows and columns at the same time:

$$f_{x|y} = \left[\begin{array}{c:c} \frac{1}{3} & \frac{3}{4} \\ \hdashline \frac{2}{3} & \frac{1}{4} \end{array}\right], \; f_{y|x} = \left[\begin{array}{c:c} \frac{4}{7} & \frac{3}{7} \\ \hdashline \frac{2}{3} & \frac{1}{3} \end{array}\right].$$

For compatible conditional distributions,the interchanging of rows and columns would not affect the recovering of the joint pdf, as long as a reverse interchange also takes place at the final step for computing f_{xy}. In this example, following the prescribed procedure,

$$f_{xy} = \begin{bmatrix} 0.4 & 0.3 \\ 0.2 & 0.1 \end{bmatrix}.$$

After reversing the interchanged rows and columns, we get back the joint pdf in (2).

The key concept for forming an ensemble for incompatible conditionals is that unlike compatible CSMs, the joint pdf computed using CSOI would be different with the row and/or column interchange. This creates an opportunity for generating different candidate solutions to form the required ensemble.

Example 1b (Incompatible conditionals)**.** We follow Arnold *et al.* [3] and consider the following incompatible conditional distributions:

$$f_{x|y} = \left[\begin{array}{c:c} \frac{1}{4} & \frac{1}{3} \\ \hdashline \frac{3}{4} & \frac{2}{3} \end{array}\right], \; f_{y|x} = \left[\begin{array}{c:c} \frac{1}{3} & \frac{2}{3} \\ \hdashline \frac{1}{10} & \frac{9}{10} \end{array}\right]. \tag{3}$$

The CSOIs $(\nabla_{12}, \nabla_2, \nabla_1, \nabla_\vartheta)$ for $f_{x|y}$ and $f_{y|x}$ are respectively $(-0.41, 0.12, -0.69, 0)$ and $(1.50, -2.20, -0.30, 0)$. Note that $\nabla_{12}^{x|y} \neq \nabla_{12}^{y|x}$. We propose to use the mean of ∇_{12} across $f_{x|y}$ and $f_{y|x}$ for computing the joint pdf. Thus, the interaction terms for the joint distribution is $\theta_{12} = (\frac{1}{2}(-0.41 + 1.50), -2.20, -0.69, 0)$, and the resulting joint distribution is given by:

$$f^{(1)} = \begin{bmatrix} 0.06 & 0.29 \\ 0.07 & 0.59 \end{bmatrix}.$$

Three candidate joint distributions can be obtained by (1) interchanging the rows, (2) interchanging the columns, and (3) interchanging both the rows and columns. Their respective pdfs are given by:

$$f^{(2)} = \begin{bmatrix} 0.12 & 0.25 \\ 0.14 & 0.49 \end{bmatrix}, f^{(3)} = \begin{bmatrix} 0.03 & 0.14 \\ 0.08 & 0.75 \end{bmatrix}, \text{ and}$$

$$f^{(4)} = \begin{bmatrix} 0.06 & 0.12 \\ 0.18 & 0.63 \end{bmatrix}.$$

In order to create an ensemble estimate based on $\mathcal{E} = \{f^{(1)}, f^{(2)}, f^{(3)}, f^{(4)}\}$, we derive the respective conditional distributions $f^{(i)}_{x|y}$ and $f^{(i)}_{y|x}$, $i = 1, \ldots, 4$, and compare each to the given conditionals in (3). Here, for the purpose of illustration, we use the divergence measure G^2 (see the subsection on divergence measures) for indicating the difference between the derived and the given conditional distributions. For the above example, the G^2 measure for $f^{(1)}, \ldots, f^{(4)}$ are respectively 0.37, 0.30, 0.36, and 0.28.

The final step for creating a CE estimate of the joint distribution can be summarized as follows: (1) derive the weight for each candidate $f^{(i)}$ as the reciprocal of the divergence measure, and (2) form the weighted combination of the candidate joint distributions from the ensemble.

The CE estimate of the joint distribution for *Example 1b* is given by

$$f^*_{xy} = \begin{bmatrix} 0.07 & 0.20 \\ 0.12 & 0.61 \end{bmatrix},$$

which has a divergence measure G^2 of 0.16, smaller than the G^2 of every $f^{(i)}$.

2.1 Discrepancy and weighted averages

To evaluate the quality of an ensemble of pdfs that is "fitted" to a given DN, we consider the following discrepancy measures:

(1) Freeman-Turkey [14]: $F^2(\hat{p}; p) = 4\sum_{i=1}^n (\sqrt{\widehat{p}_i} - \sqrt{p_i})^2$;
(2) The divergence measure of [18, p. 33-34]: $G^2(\hat{p}; p) = 2\sum_{i=1}^n p_i \log(p_i/\widehat{p}_i)$;
(3) Neyman's chi-square [5, p. 348]: $N^2(\hat{p}; p) = \sum_{i=1}^n (\widehat{p}_i - p_i)^2/p_i$.

Here, $\widehat{p}_i$ represents the conditional probability of a pdf of the ensemble and p_i represents the given conditional probability. The G^2 measures goodness-of-fit and is often used for model selection in statistics literature; N^2 measures the closeness of the DN to the computed $\hat{p}_i$ and reflects the predictive power of $\hat{p}_i$; and F^2 is a bona fide distance and has been used in the study of robustness.

3. SIMULATION EXPERIMENTS

We conducted two simulation experiments to compare the performances of canonical ensemble (CE) with the methods of PGS and LP. In the first experiment, we used two ways to generate low-dimensional conditional distributions: (1) with pre-specified entries of conditional probability, and (2) with conditional probability entries randomly generated and then normalized. To further study the performance of the various approaches in more practical settings, in the second experiment we simulated 10-dimensional DNs for comparison. For randomly generated conditional distributions in both experiments, we simulated 100 sets of random DNs for evaluating the competitive performance of CE.

3.1 Experiment 1: Bivariate conditional models

The following DN (Example 3.1) is taken from [17]. The conditional models are specified as

$$f_{x|y}(c,d) = \begin{bmatrix} \frac{1}{7} & \frac{1}{4} & \frac{3}{7}+c & \frac{1}{7}+d \\ \frac{2}{7} & \frac{2}{4} & \frac{1}{7} & \frac{2}{7} \\ \frac{4}{7} & \frac{1}{4} & \frac{3}{7}-c & \frac{4}{7}-d \end{bmatrix} \text{ and}$$

$$f_{y|x}(a,b) = \begin{bmatrix} \frac{1}{6}-a & \frac{1}{6}+a & \frac{3}{6} & \frac{1}{6} \\ \frac{2}{7} & \frac{2}{7} & \frac{1}{7} & \frac{2}{7} \\ \frac{2}{6}+b & \frac{1}{12}+b & \frac{1}{4}-2b & \frac{1}{3} \end{bmatrix}.$$

When $a = b = c = d = 0$, $f_{x|y}$ and $f_{y|x}$ are compatible. We designated the following 4 sets of values for the perturbation parameters a, b, c, and d to create increasing magnitude of incompatibility accordingly: Case (1): $a = -1/12, b = c = d = 0$; Case (2): $c = -1/7, d = 1/7, a = b = 0$; Case (3): $c = -1/7, d = 1/7, a = -1/12, b = 0$; and Case (4): $c = -1/7, d = 1/7, a = -1/12, b = 1/12$. The degree of incompatibility increases from Case (1) to Case (4).

The comparison methods for CE are the LP [3] and PGS [16]. LP computes a joint $\widehat{p}_{ij} > 0$ with minimum ϵ_{ij} such that $|\widehat{p}_{ij} - f_{x|y}\widehat{p}_{+j}| \le \epsilon_{ij}$, $|\widehat{p}_{ij} - f_{y|x}\widehat{p}_{i+}| \le \epsilon_{ij}$ and $\sum_{i,j} \widehat{p}_{ij} = 1$. The LP formulation frames the derivation of the joint distribution as a a multi-objective optimization. By imposing restrictions $\epsilon_{ij} = \epsilon$, the problem can be transformed into a single-objective linear programming problem [3]. For the example in this experiment, there are 48 inequalities, 1 equality, and 13 unknowns.

Table 1 lists the three divergence measures for PGS (using 100,000 samples), LP, and CE. From Case (1) through Case (4)—i.e., when incompatibility increases—the advantage of LP over PGS diminishes from 15.5% ($G^2 = 0.056$ vs. 0.0663) to 2.7% ($G^2 = 0.6248$ vs. 0.6418). Furthermore, from Table 1, the CE solutions consistently outperform the computationally intensive LP. The G^2 errors for CE are 26.6%, 47.9%, 43.4%, and 31.5% smaller than the respective G^2 for LP across Cases (1) to (4).

To further examine the distribution of the errors, Figure 2 shows the heat map of G^2 error, computed cell-wise for Case (4). Recall that when $a = b = c = d = e = 0$, $f_{x|y}(c,d)$ and $f_{y|x}(a,b)$ are compatible and determine a unique joint pdf f^*_{xy} such that there is zero discrepancies between the estimated and the given conditional distributions. The heat maps A_1 and B_1 show, respectively, the cell-level difference (error) between $f_{x|y}(-1/7, 1/7)$ and $f^*_{x|y}$, and between

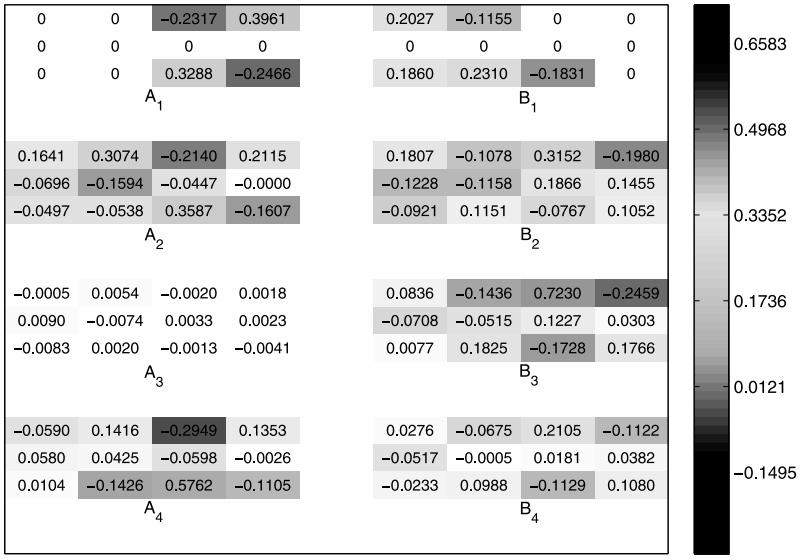

Figure 2. Heat maps based on G^2 error, computed at the cell level, for Example 3.1 Case (4). A and B respectively represent $f_{x|y}$ and $f_{y|x}$, and 1, 2, 3, 4 respectively represent the given conditional distributions, the LP solution, the PGS solution, and the CE solution.

Table 1. Comparison of the Performance Among LP, PGS, and CE for Example 3.1

		G^2	N^2	F^2
Compatible	LP	0.0000	0.0000	0.0000
	PGS	0.0006	0.0006	0.0006
	CE	0.0000	0.0000	0.0000
Case (1)	LP	0.0560	0.0551	0.0557
	PGS	0.0663	0.0768	0.0680
	CE	0.0411	0.0410	0.0409
Case (2)	LP	0.2186	0.2242	0.2191
	PGS	0.2231	0.2720	0.2304
	CE	0.1138	0.1166	0.1141
Case (3)	LP	0.2712	0.2905	0.2741
	PGS	0.2891	0.3518	0.2987
	CE	0.1536	0.1619	0.1551
Case (4)	LP	0.6248	0.6258	0.6163
	PGS	0.6419	0.8523	0.6637
	CE	0.4278	0.4696	0.4321

$f_{y|x}(-1/12, 1/12)$ and $f^*_{y|x}$—i.e., all of the non-white (non-zero) areas in A_1 and B_1 arise because of the perturbations in a, b, c, and d. Similarly, the heat maps A_2 and B_2, A_3 and B_3, and A_4 and B_4 respectively show the cell-level differences between the derived and given conditionals using LP, PGS, and CE. From A_2 and B_2, it can be seen that the LP method tends to distribute cell-wise discrepancies rather evenly over the entire support. In other words, the LP method tends to "disregard" the original pattern of incompatibility as suggested in A_1 and B_1. In contrast, there exists a strong imbalance in the distribution of errors in A_3 and B_3, suggesting that PGS tends to concentrate the errors to the $(y|x)$ pdf, while the entries of A_3, representing the error in the $(x|y)$ pdf, are all less than 0.01.

Another interesting observation of the CE solution is that the conditional distributions not only have the smallest local errors, they also exhibit a distributional pattern similar to that in A_1 and B_1. The similarities between A_1 and A_4 (and between B_1 and B_4) are not surprising because the canonical parameters in the CSOI are measures of local dependency and as a result they also reflect local incompatibility. In other words, the CE solution creates a joint density of which the derived conditionals mimic the source of incompatibility in the given conditionals. This kind of information may be useful for fine tuning the DN toward a more compatible model.

Randomly generated bivariate conditional distributions This simulation study follows a sequence of steps, as follows: (1) Generated 100 pairs of 3×4 matrices with random positive integers between 1 and 100. (2) Convert one matrix to $f_{x|y}$ and another to $f_{y|x}$ by column normalization and row normalization, respectively. (3) For each pair, compute the LP, the PGS (using 100,000 sample), and the CE solutions. (4) Calculate the respective divergence measures in G^2, N^2, and F^2. (5) Compute the percentage of reduction relative to PGS. A positive percentage implies a reduction in error relative to PGS. (6) Average the percentages of reduction relative to PGS over the 100 simulations.

The average percentages of reduction for LP relative to the PGS were 9.5% in G^2, -19.2% in N^2, and -2.6% in F^2; while the percentages of reduction for the CE were 41.6% in G^2, 41.5% in N^2, and 37.3% in F^2. Thus, the CE method outperforms both the LP and PGS methods by significant

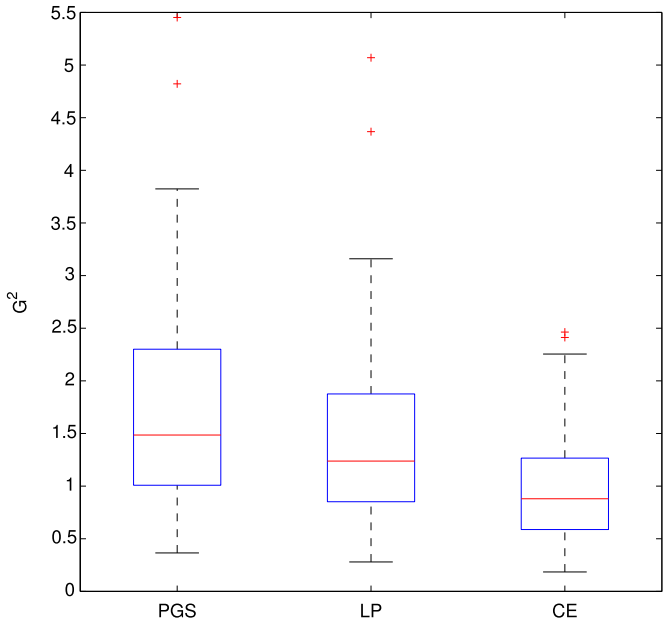

Figure 3. Box plots for G^2 among PGS, LP, and CE based on 100 randomly generated 3×4 conditional models.

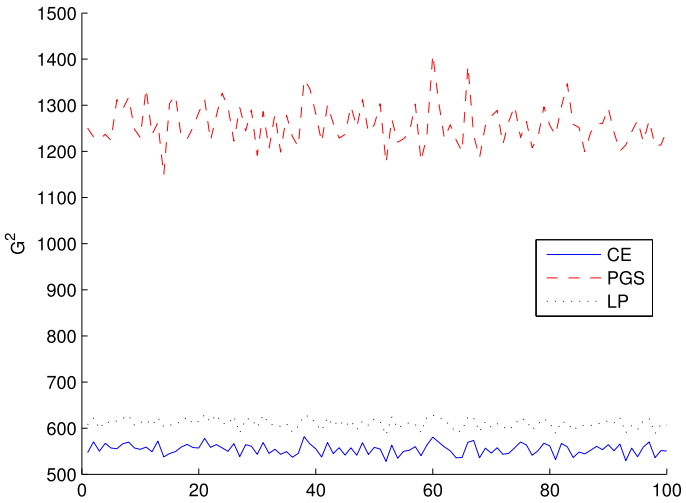

Figure 4. G^2 among PGS, LP, and CE based on 100 randomly generated 10-dimension conditional models.

margins. Figure 3 shows the box plots of the distributions of the G^2 for PGS, LP, and CE across the 100 simulations. The graphs demonstrate that CE not only has a smaller overall error, it also has less dispersion in the distribution, which is suggestive of its potential robustness in dealing with incompatibility.

3.2 Experiment 2: Randomly generated 10-dimensional DNs

We conducted similar comparisons for 100 randomly generated DNs with d, the dimension of the joint distribution, set at 10, and each variable is assumed to be binary. Thus a joint pdf has $2^{10} = 1{,}024$ cells. In this experiment, we randomly selected 100 pdfs from the possible candidate canonical solutions to form the ensemble. For the PGS, we used 1,000,000 PGS samples for every DN. For CE, the G^2 errors were used for computing ensemble weights. Figure 4 shows box plots of G^2 errors for the 100 simulated 10-dimensional DNs for the three methods. LP has a substantial smaller G^2 than PGS, and CE consistently outperforms PGS and LP in the G^2 error. The same ranking order in performance also holds for the N^2 and F^2 error measures. Table 2 summarizes the means and standard deviations of the 100 G^2, N^2, and F^2. Compared with PGS, the reductions in error for CE are respectively 56%, 49%, and 49%.

In terms of a computational overhead, the PGS was the largest and CE was the least (all approaches were implemented using Matlab). For each simulated DN, the PGS required 15 minutes of CPU time, and the LP used 50 seconds. In contrast, the CE method used only 2.5 seconds. In summary, both in terms of computational efficiency and error reduction, the CE solution outperformed LP and PGS.

Table 2. Comparison of the Average Performance, Mean (Standard Deviation) Among LP, PGS, and the CE on 100 Randomly Specified 10-Dimension Conditional Models

	G^2	N^2	F^2
PGS	1257.7 (45.6)	5036.0 (268.7)	2365.6 (69.2)
LP	610.4 (10.7)	3375.9 (152.4)	1368.3 (26.9)
CE	554.4 (12.0)	2567.0 (115.4)	1208.3 (27.5)

4. REAL EXAMPLES

4.1 Example I: Hematologic toxicity

The first real data example is a study of a relatively simple two-variable DN of risk factor and treatment response. The DN thus involves both a diagnostic model and a treatment model. This example was used in illustrating the PGS in [10] and we use it here to compare the performance of several different approaches.

Table 3 is taken from a study reported in [26], which then comprised one of the largest prospective studies for investigating the relationship between polymorphism in the gene region UGT1A1*28 and the response to irinotecan for metastatic colorectal cancer patients. It was observed that a significant increased risk of developing severe hematologic toxicity exists among patients carrying the TA_7 allele [26]. The hypothesis is that genetic testing for the UGT1A1*28 polymorphism may have utility as a predictor of the response to irinotecan. In Table 3, the row variable x represents polymorphism in gene region UGT1A*28 with three genotypes, TA_6/TA_6, TA_6/TA_7, and TA_7/TA_7. These genotypes are known to be associated with the response to treat-

Table 3. Cross-Tabulation by UGT1A1*28 Polymorphism and Response to Treatment [26]

Polymorphism (x)	Response to Chemotherapy Treatment (y)				Total
	Complete Response	Partial Response	Stable Disease	Progressive Disease	
TA_6/TA_6	10	34	29	36	109
TA_6/TA_7	5	40	32	31	108
TA_7/TA_7	3	11	5	2	21

Table 4. Comparisons of LP, PGS, and CE for the Conditional Model of Table 3

	G^2	N^2	F^2
LP	0.0245	0.0360	0.0265
PGS	0.0240	0.0291	0.0245
CE	0.0138	0.0182	0.0145

ment of a combination of irinotecan fluorouracil and leucovorin, which is represented by the column variable y. The four categories of y are *complete response*, *partial response*, *stable disease*, and *progressive disease*, respectively coded as 1–4.

Logistic regression was used to build the conditional model. Specifically, $f_{x|y}$ was estimated by applying multinomial logistic regression of x on y, and $f_{y|x}$ was estimated by applying ordinal logistic regression of y on x:

$$f_{x|y} = \begin{bmatrix} 0.5052 & 0.3877 & 0.4476 & 0.5388 \\ 0.4226 & 0.5008 & 0.4631 & 0.3975 \\ 0.0722 & 0.1115 & 0.0893 & 0.0637 \end{bmatrix} \text{ and}$$

$$f_{y|x} = \begin{bmatrix} 0.0648 & 0.3379 & 0.2861 & 0.3112 \\ 0.0679 & 0.3469 & 0.2847 & 0.3005 \\ 0.1677 & 0.4944 & 0.2034 & 0.1345 \end{bmatrix}.$$

Clinicians commonly use two conditional models for such data: the diagnostic model is $f_{x|y}(x|y)$ and the treatment model is $f_{y|x}(y|x)$. Of practical interest are the following sets of parameters: the diagnostic odds $d_{ij} = Pr(x = i|y = j)/Pr(x = i|y = j+1), 1 \le i \le 3, 1 \le j \le 3$, and the response odds $t_{ij} = Pr(y = j|x = i)/Pr(y = j|x = i+1), 1 \le i \le 2, 1 \le j \le 4$. We computed the joint distributions using, respectively, LP, PGS, and CE. Table 4 compares the various divergences between the observed joint distribution (Table 3) and the estimated joint distributions. Compared to PGS, the G^2 of LP is 2.08% larger, and the N^2 of LP is 23.7% larger. It was a surprise to see that the computationally intensive LP has slightly larger errors than PGS given the results in our simulation experiments. On the other hand, the performance of the CE method (weights based on G^2)is consistent with the results in the simulation experiments, and it outperforms PGS across G^2, N^2, and F^2 with 47.5% reduction in G^2, 37.5% reduction in N^2, and 40.8% reduction in F^2. CE also achieves similar percentages of error reduction relative to LP.

4.2 Example II: DN for symptom structure

The second data set contained information from $N = 100$ patients with either a primary brain tumor or brain metastases. Specific details of the data set were reported elsewhere [22]. Here we only focused on a specific domain of self-reported quality of life measures: emotion. Response data were extracted from the functional assessment of cancer therapy (FACT) and the brain-tumor-specific subscale (FACT-B), both of which have been validated in other studies [8, 30]. For illustrating the different approaches of the DN, we used data collected from the following four emotion items: "I feel sad" (Sad, x_2); "I am losing hope in the fight against my illness" (Losing Hope, x_4); "I feel nervous" (Nervous, x_1); and "I worry that my condition will get worse" (Condition, x_3). The item responses were dichotomized as 1 = presence of symptom and 0 = absence of the concern or symptom.

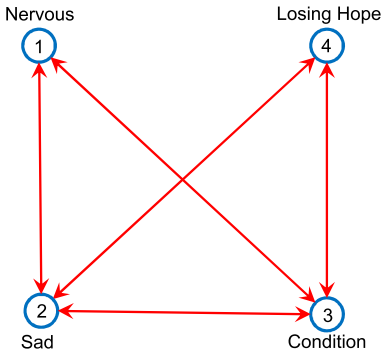

Figure 5. Dependency network for concerns and symptoms in cancer patients.

Logistic regressions were applied to the data set such that each individual variable was treated as a dependent variable and other variables were treated as predictors. Backward selection was used to select the relevant predictors. The resulting DN is shown in Figure 5.

We compare the divergences of three methods: LP, PGS, and CE. For PGS, we implemented four different scan patterns (e.g., 1234 represents the scan pattern $x_1 \to x_2 \to x_3 \to x_4$), each with 1,000 burn-in cycles and 1,000,000 sample points. Table 5 shows the divergences for LP, PGS, and CE. Note that different scan patterns from the PGS produce different results, an observation that is consistent with previous findings [10]. From Table 5, the CE method tends to have smaller divergences compared to the LP and the PGS methods. For example, using N^2, the canonical-ensemble method has a divergence of 0.016, which is 57% less than the divergence of the PGS (average divergence across four

Table 5. Comparisons of LP, PGS, and the CE for the DN in Figure 5

	G^2	N^2	F^2
LP	NaN[1]	0.4146	0.8186
PGS (1234)	0.0079	0.0349	0.0360
PGS (2341)	0.0145	0.0601	0.0649
PGS (3421)	0.0057	0.0281	0.0268
PGS (4123)	0.0051	0.0250	0.0239
CE	0.0035	0.0161	0.0162

[1]LP produces a joint distribution with structure zeros.

scan patterns = 0.037), and 92% less than the divergence of LP (0.41).

5. CONCLUSIONS

The DN represents a flexible and powerful modeling tool to capture complex relationships between a large number of variables. One can think of the joint distribution derived from the DN as an approximation to a causal structure represented by a Markov network. It is also a generalization of DAG models. Thus, while the DN is often used as a "black box" for optimizing prediction, it can also be used to serve the purpose of an explanatory model in health and medical sciences. In this paper, we propose an efficient ensemble method for computing the joint distributions for incompatible CSMs. The method may be viewed as two-level model averaging. At the first level, we take averages of "copies" of the characterizing interactions generated from incompatible conditional distributions. A joint distribution is formed by "gluing" averaged interaction terms from the conditionals. At the second level, we take the (weighted) average of ensemble members to form a final solution. The ensemble is created by changing the anchor (i.e., switching rows and/or columns) from which the canonical interactions are calculated.

Based on two simulation experiments and two real data sets, our main finding is as follows: compared with two existing procedures, the PGS and LP, the two-level averaging CE procedure tends to produce both smaller overall error and less localized error distributions. The percentages of reduction in error, as measured by a variety of divergence measures, are substantial. An added bonus for the CE approach is that its computations, which are mostly direct arithmetic operations, are rather straightforward and also highly scalable. The procedure can also be easily parallelized. For an LP-based analysis, the computational overhead is high. For the 10-dimensional simulation Example 3.2, LP needs to solve an optimization problem with a total of 10,240 inequalities and one equality in 1,024 unknowns. Despite its computational overhead, the performance of LP optimization is moderate compared to PGS and CE, as evidenced by our results in the simulation experiments and the examples with real data. In addition, because the LP method minimizes the cell-wise deviations, the computational overhead grows quickly with the dimension of the problem, which implies that LP may be limited only to problems of low dimensionality. The PGS has been commonly used in DNs and will probably continue to be a benchmark procedure to which other new procedures will have to compare. We demonstrated in this study that CE performs remarkably well compared to PGS.

The problem of incompatible CSMs is not limited to DNs. Another potential application of methods for potentially incompatible CSMs is the multiple imputation of missing data [27, 28]. Recent developments of multiple imputation by chained equations (MICE), which makes use of a Gibbs sampler or other Markov chain Monte Carlo-based methods that operate on a set of conditionally specified pdfs, have drawn significant interest from researchers studying missing values in complex data sets. For each variable with a missing value, an imputed value is created under an individual conditional-regression model. This kind of procedure bears a strong resemblance to a DN. Rubin [23] argued that MICE combined the best features of many currently available multiple-imputation approaches. Due to its flexibility over compatible multivariate-imputation models [24] and ability to handle different variable types (continuous, binary, and categorical) MICE has gained acceptance for its practical treatment of missing data, especially in high-dimensional data sets [21]. However, MICE has the limitation of potentially encountering incompatible conditional-regression models, and it has been shown that an incompatible imputation model can lead to biased estimates from imputed data [13]. This is an area in which the current approach could make a contribution.

There are limitations to this study. First, we focus on discrete distributions. Not unlike many early graphical model applications, continuous variables need to be discretized for the CE method described in this article. It is possible to create canonical parameters for CSMs for continuous variables [29] and current work is underway. Second, our results were derived from CSMs of low-modest dimensions. Further studies will be required to examine the performance of CE.

In conclusion, the ensemble approach based on canonical parameters represents a viable method for reconciling incompatible (or nearly compatible) conditional distributions. We showed that the CE method performs well in low- to moderate-dimensional problems, compared to two existing approaches—linear programming and the Gibbs sampler. Thus, the canonical-ensemble method is a serious competitor for computing models based on the DN, which is increasingly used in a large, complex netwrok of variables in medicine and health-related sciences.

APPENDIX A. THE CHARACTERIZING SET OF INTERACTIONS

Suppose that $X = (x_1, \ldots, x_d)$, where each x_j is a discrete random variable with support $\Omega_j = \{1, \ldots, K_j\}$. Set

$\aleph = \{1, \ldots, d\}$ and $\Omega_\aleph = \prod_{j=1}^{d} \Omega_j$. The joint distribution of X is a row vector $f_\aleph = (f_\aleph(1, \ldots, 1), \ldots, f_\aleph(K_1, \ldots, K_d))$, which is arranged in lexicographical order. Also, define $\log f_\aleph = (\log f(1, \ldots, 1), \ldots, \log f(K_1, \ldots, K_d))$. The $(x_j|x_{-j})$ conditional distribution of $f_\aleph$ is $f_{j|-j}(x) = f_\aleph(x)/f_{-j}(x_{-j})$, where $f_{-j}(x_{-j})$ is the marginal distribution of $x_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d) \in \Omega_1 \times \cdots \times \Omega_{j-1} \times \Omega_{j+1} \times \cdots \times \Omega_d$.

Let $\mathbf{B}_i, i = 1, \ldots, d$, be a $K_i \times K_i$ upper triangular matrix of 1's and $\mathbf{A} = \mathbf{B}_d \otimes \cdots \otimes \mathbf{B}_1$ where $\otimes$ is the Kronecker product. Then, $\boldsymbol{\theta}_\aleph = \mathbf{A}^{-1}(\log f_\aleph)^{\mathrm{T}}$ and $\boldsymbol{\theta}_{i|-i} = \mathbf{A}^{-1}(\log f_{i|-i})^{\mathrm{T}}$ are called the canonical interaction vectors of $f_\aleph$ and $f_{i|-i}$, respectively. The orders of interactions can be more conveniently represented by the following difference operator ∇_i and identity operator ϑ: for $1 \le k_i < K_i, i = 1, \ldots, d$,

$$\begin{aligned} & \nabla_i h(k_1, \ldots, k_i, \ldots, k_d) \\ = \; & h(k_1, \ldots, k_{i-1}, k_i, k_{i+1}, \ldots, k_d) - \\ & h(k_1, \ldots, k_{i-1}, k_i + 1, k_{i+1}, \ldots, k_d), \end{aligned}$$

and $\vartheta h(k_1, \ldots, k_d) = h(k_1, \ldots, k_d)$, with $\nabla_i h(k_1, \ldots, K_i, \ldots, k_d) = 0$. Define $\boldsymbol{\nabla^i}$ be the vector $(\nabla_i, \ldots, \nabla_i, \vartheta)^T$. Furthermore, let $\boldsymbol{\nabla}$ be $\boldsymbol{\nabla^d} \otimes \cdots \otimes \boldsymbol{\nabla^1}$. For example, with $K_j = 3$ and $1 \le j \le 2$, the $\boldsymbol{\nabla}$ vector is $(\nabla_2\nabla_1, \nabla_2, \nabla_2, \nabla_2\nabla_1, \nabla_2, \nabla_2, \nabla_1, \nabla_1, \vartheta)^T$, and for $K_j = 2$ and $1 \le j \le 3$, the $\boldsymbol{\nabla}$ vector is $(\nabla_3\nabla_2\nabla_1, \nabla_2\nabla_3, \nabla_1\nabla_3, \nabla_3, \nabla_2\nabla_1, \nabla_2, \nabla_1, \vartheta)^T$. Thus, there are orders within $\boldsymbol{\nabla}$, and we will use the shorthand notation $\nabla_a, a \subset \aleph$ for $\prod_{j \in a} \nabla_j$. Define Hamadan product $\cdot$ between two vectors as $(x_1, \ldots, x_K)^{\mathrm{T}} \cdot (y_1, \ldots, y_K)^{\mathrm{T}} = (x_1 y_1, \ldots, x_K y_K)^{\mathrm{T}}$. It can be proved that $\boldsymbol{\theta}_\aleph = \boldsymbol{\nabla} \cdot (\log f_\aleph)^{\mathrm{T}}$ and $\boldsymbol{\theta}_{i|-i} = \boldsymbol{\nabla} \cdot \log f_{i|-i}$ [17, Theorem 1]. We call $\nabla_a \log f_\aleph$ ($\nabla_a \log f_{i|-i}$) the a-interaction of $f_\aleph$ ($f_{i|-i}$). For example, $\nabla_{12} \log f_{1|-1} = \nabla_1\nabla_2 \log f_{1|-1}$ is the $(1,2)$-interaction, and $\nabla_{123} \log f_{2|-2} = \nabla_1\nabla_2\nabla_3 \log f_{2|-2}$ is the $(1,2,3)$-interaction. The following two results [17] relate the a-interaction of $f_{i|-i}$ with the a-interaction of $f_\aleph$, and states the compatibility condition in terms of interactions.

(1) (Invariance) The a-interaction of $f_{i|-i} = f_\aleph/f_{-i}$ is identical to the a-interaction of $f_\aleph$ provided that a contains i—i.e., for $i \in a$, $\nabla_a \log f_\aleph = \nabla_a \log f_{i|-i}$—which is termed an *invariant* interaction. The totality of all the invariant interactions of a DN is called the *characterizing set of interactions* (CSOI). For $d = 2$, the CSOI is the union of $\{\nabla_{12}, \nabla_1\}$ of $f_{1|2}$ and $\{\nabla_{12}, \nabla_2\}$ of $f_{2|1}$.

(2) (Compatibility) Two conditional $f_{i|-i}$ and $f_{j|-j}$ are compatible if and only if for $(i,j) \in a$, $\nabla_a \log f_{i|-i} = \nabla_a \log f_{j|-j}$. When every pair of conditionals of $\{f_{i|-i}, 1 \le i \le d\}$ are compatible, there exists a unique joint probability density function (pdf), $f^*_\aleph$ from which every $f_{i|-i}$ can be exactly derived. Otherwise, the DN is incompatible.

APPENDIX B. THE ENSEMBLE APPROACH

We first describe the bivariate case and then the general case.

B.1 The bivariate case

We first use the bivariate case to set up the notation and motivate the approach. Let $X = (x_1, x_2)$, and $f_{1|2}$ and $f_{2|1}$ be the two conditional pdfs, from which a joint pdf f_{12} is to be constructed. Assume that x_1 takes values $1, \ldots, K_1$ and x_2 takes $1, \ldots, K_2$. The invariant interactions of $f_{1|2}$ and $f_{2|1}$ are, respectively, $\mathcal{A}_1 = \{\nabla_1 \log f_{1|2}(i|K_2), \nabla_{12} \log f_{1|2}(i|j)\}$ and $\mathcal{A}_2 = \{\nabla_2 \log f_{2|1}(j|K_1), \nabla_{12} \log f_{2|1}(j|i)\}$, for $1 \le i < K_1$ and $1 \le j < K_2$. Their degrees of freedom are $(K_1 - 1) + (K_1 - 1)(K_2 - 1)$ and $(K_2 - 1) + (K_1 - 1)(K_2 - 1)$, respectively. Then, we have the following equivalent conditions:

(I) $\nabla_{12} \log f_{1|2} = \nabla_{12} \log f_{2|1}$; that is $f_{1|2}$ and $f_{2|1}$ are compatible.

(II) There exists a unique f^*_{12} such that $f^*_{12}/f^*_{+2} = f_{1|2}$ and $f^*_{12}/f^*_{1+} = f_{2|1}$, where "+" represents summation over replaced subscript. For example, f^*_{1+} is the x_1-marginal of f^*_{12}.

(III) The interactions uniquely determining f^*_{12} are $\{\nabla_1 \log f_{1|2}, \nabla_2 \log f_{2|1}, \nabla_{12} \log f_{1|2}\}$.

The compatibility condition (I) reduces the total number of interactions to $K_1K_2 - 1$ interactions of (III) above. When $\nabla_{12} \log f_{1|2}(i|j) \neq \nabla_{12} \log f_{2|1}(j|i)$ for some i or j, there does not exist a unique f^*_{12} satisfying condition (2). Our goal here is to compute a f^*_{12} such that the corresponding $f^*_{1|2}$ and $f^*_{2|1}$ are least-deviated from $f_{1|2}$ and $f_{2|1}$ collectively. The ensemble of the joint distributions is created as follows:

1. Take the vector operators $\boldsymbol{\nabla^1} = (\nabla_1, \ldots, \nabla_1, \vartheta)^{\mathrm{T}}$ and $\boldsymbol{\nabla^2} = (\nabla_2, \ldots, \nabla_2, \vartheta)^{\mathrm{T}}$. Both have the identical operator in the last category $x_1 = K_1$ and $x_2 = K_2$, respectively. For operators $\boldsymbol{\nabla^2} \otimes \boldsymbol{\nabla^1}$, the cell (K_1, K_2) is called an anchor.
2. Interchange the ∇_1 at location $x_1 = i$ with the ϑ at location $x_1 = K_1$ within ∇_1.
3. Interchange the ∇_2 at location $x_2 = j$ with the ϑ at location $x_2 = K_2$ within ∇_2. After both interchanges, position (i,j) becomes the new anchor.
4. Use the vectors of rearranged difference operators to generate the interaction vector. Denote the interaction vector generated from using (i,j) as the anchor by $\nabla_a \log f^{(i,j)}_{1|2}$ and $\nabla_a \log f^{(i,j)}_{2|1}$.
5. Select from the above (i,j)-anchored interactions to form the interaction terms for the joint distribution. Specifically, the canonical parameters of the joint $\boldsymbol{\theta}^{(i,j)}$ is given by: $\boldsymbol{\theta}^{(i,j)} \triangleq \{\frac{1}{2}(\nabla_{12} \log f^{(i,j)}_{1|2} + \nabla_{12} \log f^{(i,j)}_{2|1}), \quad \nabla_1 \log f^{(i,j)}_{1|2}, \nabla_2 \log f^{(i,j)}_{2|1}\}$ and denote the corresponding joint pdf by $g^{(i,j)}_{12}$.

6. Perform the reverse interchanges for $g_{12}^{(i,j)}$ to obtain the joint pdf $h_{12}^{(i,j)}$. In the process, $x_1 = i, x_1 = K_1, x_2 = j$ and $x_2 = K_2$ are returned to their original positions.
7. Repeat steps 1–6 using a different anchor. As a result of using different anchors (i, j), $1 \le i \le K_1, 1 \le j \le K_2$, a total of $K_1 \times K_2$ different $h_{12}^{(i,j)}$, are generated.

To illustrate how the anchored-interchange procedure works, we use an incompatible 2×2 conditional model [3]. The original DN before any interchange (anchored at $(2,2)$) is:

$$f_{1|2}^{(2,2)} = \begin{bmatrix} \frac{1}{4} & \frac{1}{3} \\ \frac{3}{4} & \frac{2}{3} \end{bmatrix}, \; f_{2|1}^{(2,2)} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{10} & \frac{9}{10} \end{bmatrix}.$$

The other conditional distributions that are anchored at a different cell are:

(1): $1 \rightleftarrows 2$ for x_1 and no interchange for x_2

$$f_{1|2}^{(1,2)} = \begin{bmatrix} \frac{3}{4} & \frac{2}{3} \\ \frac{1}{4} & \frac{1}{3} \end{bmatrix}, \; f_{2|1}^{(1,2)} = \begin{bmatrix} \frac{1}{10} & \frac{9}{10} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix};$$

(2): $1 \rightleftarrows 2$ for x_2 and no interchange for x_1

$$f_{1|2}^{(2,1)} = \begin{bmatrix} \frac{1}{3} & \frac{1}{4} \\ \frac{2}{3} & \frac{3}{4} \end{bmatrix}, \; f_{2|1}^{(2,1)} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{9}{10} & \frac{1}{10} \end{bmatrix};$$

(3): $1 \rightleftarrows 2$ for x_1 and $1 \rightleftarrows 2$ for x_2

$$f_{1|2}^{(1,1)} = \begin{bmatrix} \frac{2}{3} & \frac{3}{4} \\ \frac{1}{3} & \frac{1}{4} \end{bmatrix}, \; f_{2|1}^{(1,1)} = \begin{bmatrix} \frac{9}{10} & \frac{1}{10} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}.$$

B.2 The d-component case

The ensemble procedure described in the previous section can be generalized to higher dimensions. Consider the following CSMs: $\mathcal{F} = \{f_{j|-j}(x_j|x_{-j}), 1 \le j \le d\}$. When all of the invariant interactions from $\mathcal{F}$ are pooled together, there are d versions of the form ${}^{j}\nabla_{1\ldots d}$, one from each $f_{j|-j}, 1 \le j \le d$, and $d-1$ versions of ${}^{j}\nabla_{1\ldots(d-1)}$, one from each $f_{j|-j}, 1 \le j \le d-1$, and so on. The notation $i_j \rightleftarrows K_j$ indicates the relocation of the cell originally residing at $(i_1, \ldots, i_d)$ to $(K_1, \ldots, K_d)$ and forming the new anchor for calculating interactions. As an illustration, the following table lists the invariant interactions for a trivariate DN.

Conditional pdf	Invariant Interactions
$f_{1\|23}(x_1\|x_2, x_3)$	${}^{1}\nabla_{123}, {}^{1}\nabla_{12}, {}^{1}\nabla_{13}, {}^{1}\nabla_{1}$
$f_{2\|13}(x_2\|x_1, x_3)$	${}^{2}\nabla_{123}, {}^{2}\nabla_{12}, {}^{2}\nabla_{23}, {}^{2}\nabla_{2}$
$f_{3\|12}(x_3\|x_1, x_2)$	${}^{3}\nabla_{123}, {}^{3}\nabla_{13}, {}^{3}\nabla_{23}, {}^{3}\nabla_{3}$

The ensemble uses the average of the invariant interactions to formulate a joint pdf. The set of canonical parameters consists of the following averaged interaction terms:

$$\begin{aligned} \nabla^*_{1\ldots d} &= \frac{1}{d}\sum_{j=1}^{d} {}^{j}\nabla_{1\ldots d}; \\ \nabla^*_{1\ldots(d-1)} &= \frac{1}{d-1}\sum_{j=1}^{d-1} {}^{j}\nabla_{1\ldots(d-1)}; \\ &\vdots \\ \nabla^*_{ij} &= \frac{1}{2}({}^{i}\nabla_{ij} + {}^{j}\nabla_{ij}); \text{ and} \\ \nabla^*_{j} &= {}^{j}\nabla_{j}. \end{aligned}$$

Let the corresponding joint pdf be $f_{\aleph}^{(i_1,\ldots,i_d)}$, where the superscript indicates the anchor. The number of pdfs in the ensemble $\{f_{\aleph}^{(i_1,\ldots,i_d)}, i_j = 1, \ldots, K_j; j = 1, \ldots, d.\}$ equals the number of interchanges, which is $N = \prod_{j=1}^{d} K_j$. We recommend that a random sample (say of size 100) be selected to form the ensemble if N is too large.

Received 25 July 2013

REFERENCES

[1] Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999). *Conditional Specification of Statistical Models.* New York, NY, Springer-Verlag. MR1716531

[2] Arnold, B. C., Castillo, E., and Sarabia, J. M. (2001). Conditionally specified distributions: An introduction (with discussion). *Statistical Science* **16** 249–274. MR1874154

[3] Arnold, B. C., Castillo, E., and Sarabia, J. M. (2002). Exact and near compatibility of discrete conditional distributions. *Computational Statistics and Data Analysis* **16** 231–252. MR1924008

[4] Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal Royal Statistical Society: Series B* **36** 192–236. MR0373208

[5] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press, Cambridge, MA. MR0381130

[6] Bühlmann, P. (2004). Bagging, boosting and ensemble methods. In: *Handbook of Computational Statistics: Concepts and Methods*, J. Gentle, W. Härdle, and Y. Mori, eds, Springer-Verlag, Berlin, Heidelberg, pp. 877–907. MR2090165

[7] Carlson, J. M., Brumme, Z. L., Rousseau, C. M., Brumme, C. J., Matthews, P., Kadie, C., Mullins, J. I., Walker, B. D., Harrigan, P. R., Goulder, P. Jr., and Heckerman, D. (2008). Phylogenetic dependency networks: Inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Computational Biology* **4** e1000225. MR2470157

[8] Cella, D. F., Tulsky, D. S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., Siberman, M., Yellen, S. B., Winicour, P., Brannon, J., Eckberg, K., Lloyd, S., Purl, S., Blendowski, C., Goodman, M., Barnicle, M., Stewart, I., McHale, M., Bonomi, P., Kaplan, E., Taylor IV, S., Thomas, C. R., and Harris, J. (1993). The functional assessment of cancer therapy scale: Development and validation of the general measure. *Journal of Clinical Oncology* **11** 570–579.

[9] Chen, F., Cheng, Q., Liu, H., Xu, W., and Wang, S. (2013). A new inference framework for dependency networks. *Communications in Statistics – Theory and Methods* **42** 56–75. MR3004645

[10] Chen, S.-H., Ip, E. H., and Wang, Y. J. (2011). Gibbs ensembles for nearly compatible and incompatible conditional models. *Computational Statistics and Data Analysis* **55** 1760–1769. MR2748677

[11] Daly, R., Shen, Q., and Aitken, S. (2011). Review: Learning bayesian networks: Approaches and issues. *The Knowledge Engineering Review* **26** 99–157.
[12] Dobra, A. (2009). Variable selection and dependency networks for genomewide data. *Biostatistics* **10** 621–639.
[13] Drechsler, J. and Rässler, S. (2008). Does convergence really matter? In: *Recent Advances in Linear Models and Related Areas*, Shalabh and C. Heumann, eds, Physica-Verlag Heidelberg, pp. 341–355. MR2523859
[14] Freeman, M. F. and Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics* **21** 607–611. MR0038028
[15] Gelman, A. and Raghuanthan, T. E. (2001). Comment of conditionally specified distributions: An introduction (with discussions). *Statistical Science* **16** 268–269. MR1874154
[16] Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. (2000). Dependence networks for inference, collaborative filtering and data visualization. *Journal of Machine Learning and Research* **1** 49–57.
[17] Ip, E. H. and Wang, Y. J. (2009). Canonical representation of conditionally specified multivariate discrete distributions. *Journal of Multivariate Analysis* **100** 1282–1290. MR2508387
[18] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd ed.* Chapman & Hall, London, UK. MR0727836
[19] Neville, J. and Jensen, D. (2007). Relational dependency networks. *The Journal of Machine Learning Research* **8** 653–692.
[20] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK. MR1744773
[21] Rässler, S., Rubin, D. B., and Zell, E. R. (2008). Incomplete data in epidemiology and medical statistics. In: *Handbook of Statistics 27: Epidemiology and Medical Statistics*, C.R. Rao, J.P. Miller, and D.C. Rao, eds, The Netherlands, Elsevier, pp. 569–601.
[22] Rijmen, F., Ip, E. H., Rapp, S., and Shaw, E. G. (2008). Qualitative longitudinal analysis of symptoms in patients with primary and metastatic brain tumour. *Journal of the Royal Statistical Society: Series A* **171** 739–753. MR2439857
[23] Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57** 3–18. MR2055518
[24] Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. London, Chapman and Hall. MR1692799
[25] Spiegelhalter, D., Dawid, A., Lauritzen, S., and Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science* **8** 579–605. MR1243594
[26] Toffoli, E., Cecchin, E., Corona, G., Russo, A., Buonadonna, A., D'Andrea, M., Pasetto, L., Pessa, S., Errante, D., De Pangher, V., Giusto, M., Medici, M., Gaion, F., Sandri, P., Galligioni, E., Bonura, S., Boccalon, M., Biason, P., and Frustaci, S. (2006). The role of UGT1A1*28 polymorphism in the pharmacodynamics and pharmacokinetics of irinotecan in patients with metastatic colorectal cancer. *Journal of Clinical Oncology* **24** 3061–3068.
[27] Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18** 681–694.
[28] Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoornm, C. G. M., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulations* **12** 1049–1064. MR2307507
[29] Wang, Y. J. and Ip, E. H. (2008). Conditionally specified continuous distributions. *Biometrika* **95** 735–746. MR2443187
[30] Weitzner, M. A., Meyers, C. A., Gelke, C. K., Byrne, K. S., Cella, D. F., and Levin, V. A. (1995). The functional assessment of cancer therapy (FACT) scale: Development of a brain subscale and revalidation of the general version (FACT-G) in patients with primary brain tumors. *Cancer* **75** 1151–1161.

Shyh-Huei Chen
Department of Biostatistical Sciences
Wake Forest School of Medicine
Winston-Salem, NC 27157
USA
E-mail address: schen@wakehealth.edu

Edward H. Ip
Department of Biostatistical Sciences
Wake Forest School of Medicine
Winston-Salem, NC 27157
USA
E-mail address: eip@wakehealth.edu

Yuchung J. Wang
Department of Mathematical Sciences
Rutgers University
Camden, NJ 08102
USA
E-mail address: yuwang@camden.rutgers.edu