Published in final edited form as:
Comput Stat Data Anal. 2011 April 1; 55(4): 1760–1769. doi:10.1016/j.csda.2010.11.006.

Gibbs Ensembles for Nearly Compatible and Incompatible Conditional Models

Shyh-Huei Chen,
Department of Industrial Management, National Yunlin University of Science and Technology, Douliu, Yunlin 640, Taiwan

Edward H. Ip, and
Department of Biostatistical Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina 27157, U.S.A

Yuchung J. Wang
Department of Mathematical Sciences, Rutgers University, Camden, New Jersey 08102, U.S.A

Shyh-Huei Chen: chensh@yuntech.edu.tw; Edward H. Ip: eip@wfubmc.edu; Yuchung J. Wang: yuwang@crab.rutgers.edu

Abstract

Gibbs sampler has been used exclusively for compatible conditionals that converge to a unique invariant joint distribution. However, conditional models are not always compatible. In this paper, a Gibbs sampling-based approach — Gibbs ensemble —is proposed to search for a joint distribution that deviates least from a prescribed set of conditional distributions. The algorithm can be easily scalable such that it can handle large data sets of high dimensionality. Using simulated data, we show that the proposed approach provides joint distributions that are less discrepant from the incompatible conditionals than those obtained by other methods discussed in the literature. The ensemble approach is also applied to a data set regarding geno-polymorphism and response to chemotherapy in patients with metastatic colorectal

Keywords

Gibbs sampler; Conditionally specified distribution; Linear programming; Ensemble method; Odds ratio

1. Introduction

Since Besag (1974), the majority of the modeling for spatial observations has taken the conditional approach. For a finite system of random variables $(X_1, \cdots, X_J)$, the conditional likelihood is defined as

$$L_i(\theta)=P_\theta\{X_i=x_i|X_j=x_j, j \neq i\}, \quad (1)$$

Correspondence to: Edward H. Ip, eip@wfubmc.edu.

where θ is the model parameter. The product of all such conditional likelihoods, $\Pi L_i(\theta)$, called the pseudolikelihood, is used to replace the conventional likelihood in inference. Interestingly, maximum pseudolikelihood has developed into a franchise of its own, even though it was originally proposed as an expedient objective function for spatial data (Besag, 1974). The alteration of the likelihood function is partially motivated by the fact that conventional likelihood often contains an intractable partition function, whereas $L_i(\theta)$ can be modeled in a rather simple form such as a regression or a classifier. The main disadvantage of maximum pseudolikelihood estimator is that it can be biased and is an inefficient estimator — with a reported efficiencies (ratios of variances) ranging from 100% for models with weak dependence to 10% for models with strong dependence (Besag, 1977; Jensen and Künsch, 1994; Jensen and Møller, 1991). A good approximation of the conventional likelihood derived from $L_i(\theta)$, $1 \le j \le J$ can improve the efficiency and reduce the bias.

Similar formulations of likelihood occur in network models. Conditional models are individually created and fitted to the observed data with the objective of computing predictive probabilities. Heckerman et al. (2000) call the collection of such conditional models a dependence network and propose using a (random scan) Gibbs sampler to learn a joint distribution $\hat{p}(x_1, \cdots, x_J; \theta)$ from $L_i(\theta)$, $1 \le j \le J$. An important motivation of the dependency networks is to address the possible high dimensionality of data and to reduce the inference problem associated with univariate conditional distributions.

For several discrete variables, a statistical model may be formulated either as a joint model or via a system of conditional models such as (1). In general, when a joint distribution is to be recovered from a collection of conditional distributions the joint is said to be *conditionally specified* or *conditionally modeled*. Moreover, the conditional models are said to be *compatible* when there exists a joint distribution capable of generating all of the conditional distributions. Nevertheless conditional models are not always compatible when conditional models are non-saturated; usually no mutual consistency criteria are incorporated in the models for the individual conditional distributions. In fact, unlike joint models, mutually consistent conditional models are rather difficult to articulate. A “remedy” for incompatibility is to search for a joint distribution that deviates least from — however defined — the prescribed set of conditional distributions. One of the hurdles of conditional modeling is to calculate a reasonable joint from conflicting conditional distributions.

Arnold and Gokhale (1998) and Arnold et al. (2002) are among the first making efforts to find the most nearly compatible (or the minimally incompatible) joint distributions given a family of conditionals. Arnold and Gokhale (1998) considered Kullback-Leibler divergence and the L^2-distance as the measures of incompatibility, while Arnold et al. (2002) used absolute deviations as the criterion. Understandably, constrained optimization algorithms using different objective functions are involved. Under such circumstances, two issues need to be considered. First, the constrained optimization becomes more difficult as the number of conditional models or the number of variables (the dimension of the problem) increases. Arnold et al. (2002, p. 251) expressed their concern in the statement that “In practice, the number of equations, constraints and variables will limit consideration to cases in which the coordinated random variables have very few values.” For example, in a joint distribution of 4 variables, each with 3 categories, a constrained optimization formulation (linear programming) would involve 162 variables, 648 inequalities, and one equality. Second, the performance measures that had been documented are rather limited and may not be generally applicable to all problems. Some divergence measures can further complicate the algorithm in searching for an optimal solution. For example, Arnold and Gokhale (1998, p. 386) noted that switching divergence away from Kullback-Leibler divergence “clearly leads to a much less tractable objective function.”

The goal of this paper is to propose an algorithm — the Gibbs ensemble (GE) —that overcomes both difficulties. The algorithm has two important properties: (a) ability to scale up to handle large data sets of high dimensionality, and (b) ability to accommodate different performance measures. The algorithm first uses Gibbs sampler to formulate a collection of joint distributions, which is termed an *ensemble* (Bühlmann, 2004). Then a weighted average of all the distributions in the ensemble is taken as a solution. The *combination weights* are inversely proportional to the divergences of the members in the ensemble. Using the ubiquitous Gibbs sampler as the *base procedure* is highly suitable for this application because the procedure is robust, scalable, and relatively easy to program. To illustrate the ensemble approach, consider the following example.

Example 1.1 Arnold, Castillo, and Sarabia (2002, Example 10) considered

$$\mathcal{P}_1=\begin{pmatrix} \frac{1}{4} & \frac{1}{3} \\ \frac{3}{4} & \frac{2}{3} \end{pmatrix}, \quad \mathcal{P}_2=\begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{10} & \frac{9}{10} \end{pmatrix}.$$

Here $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively correspond to $L_1(x_1|x_2)$ and $L_2(x_2|x_1)$. The odds ratios of $\mathcal{P}_1$ and $\mathcal{P}_2$ are 2/3 and 9/2, respectively, and they are not compatible. Arnold et al. (2002) obtained the following optimal joint distribution via linear programming (LP):

$$\widehat{\pi}_{\mathrm{LP}}=\begin{pmatrix} 0.0749 & 0.2439 \\ 0.0995 & 0.5817 \end{pmatrix}.$$

Table 1 compares selected divergences of the different joint distributions obtained through several different methods: linear programming (LP), $\hat{\pi}_{\mathrm{LP}}$, random scan Gibbs, $\tilde{\pi}_{\mathrm{HC}}$, equal-weight Gibbs ensemble, π_{EQ}, and differentially weighted Gibbs ensembles, π_L, π_I, π_X, and π_F. The definitions of these divergences are given in Section 2. Relative to $\hat{\pi}_{\mathrm{LP}}$, the differentially weighted Gibbs ensemble reduces the L^2 and F^2 divergences by 20% and 11%, respectively. The random scan Gibbs sampler and the equally weighted Gibbs ensemble are more or less on par with $\hat{\pi}_{\mathrm{LP}}$.

The remainder of this paper is organized as follows. The idea of generating an ensemble from incompatible Gibbs sampler is discussed in Section 2. In Section 3, we present numerical comparisons of several methods using both examples in the literature and simulation data. In Section 4, the algorithm is applied to a data set regarding geno-polymorphism and response to treatment in patients with metastatic colorectal cancer. A brief conclusion is provided in Section 5.

2. Gibbs Ensemble

Consider a system of J discrete random variables $\{x_1, \cdots, x_J\}$ whose conditional model is specified by $\mathcal{G} = \{\mathcal{P}_i = \pi_{\alpha_i|\bar{\alpha}_i}, 1 \le i \le m\}$, with $\alpha_i \cap \bar{\alpha}_i = \varnothing$ and $\alpha_i \cup \bar{\alpha}_i = \aleph = \{x_1, \cdots, x_J\}$. Here, $\mathcal{P}_i$, is called a full conditional because all J variables are involved. Hence $L_i(\theta) = \pi_{i|\bar{\imath}}$, $1 \le i \le J$ is a special case. In practice, both Besag (1974) and Heckerman et al. (2000) used the conditional model $\pi_{\alpha|\beta}$, $\beta \subset \bar{\alpha}$, but this change does not affect the Gibbs samplers as long as $\cup_{i=1}^{m} \alpha_i = \aleph$.

Let S_m be the symmetric group of all possible permutations of $(1, \cdots, m)$ and $|S_m| = m!$. For every $(k_1, \cdots, k_m) \in S_m$, arrange the given conditional distributions of $\mathcal{G}$ in the following sequence:

$$\mathcal{P}_{k_1} \to \mathcal{P}_{k_2} \to \cdots \to \mathcal{P}_{k_m} \to \mathcal{P}_{k_1} \to \cdots\cdots,$$

and generate simulations using Gibbs sampler in the above order. Every $(k_1, \cdots, k_m) \in S_m$ represents a *scan pattern*. Often, this is called a *deterministic scan* (Liu, 1996) or *fixed scan*, in contrast to scanning $\mathcal{P}_i$ in a random order with prescribed probabilities, which is called a *random scan* in Levine and Casella (2006).

Let $X_{\aleph} = (x_1, \cdots, x_J)$ and $X_{\alpha_i} = (x_j, j \in \alpha_i)$. When $i = k_j$, we replace subscript α_{kj} with k_j and subscript $\bar{\alpha}_{kj}$ with $\bar{k}_j$ to avoid triple subscripts. Begin with a randomly selected $X_{\aleph}^{(0)}=(x_1^0, \cdots, x_J^0)$. Conditioned on $X_{\bar{k}_1}^{(0)}=(x_i^0, i \in \bar{\alpha}_{k1})$, draw an $X_{k_1}^{(1)}=(x_j^1, j \in \alpha_{k_1})$ from $\mathcal{P}_{k_1}$ and update $X_{\aleph}^0$ to $X_{\aleph}^{(1)}=(X_{\aleph}^{(0)} \backslash X_{k_1}^{(0)}) \cup X_{k_1}^{(1)}$. Next, conditioned on $X_{\bar{k}_2}^{(1)}$, draw $X_{k_2}^{(1)}$ from $\mathcal{P}_{k_2}$ and form $X_{\aleph}^{(2)}=(X_{\aleph}^{(1)} \backslash X_{k_2}^{(1)}) \cup X_{k_2}^{(2)}$. Successive sampling steps from $\mathcal{P}_{k_1}$ to $\mathcal{P}_{k_m}$ are called a cycle, and after one cycle every x_i^0 of $X_{\aleph}^{(0)}$ will be updated at least once, due to $\cup\alpha_i = \aleph$. Let $X_{\aleph}^{(m)}=(x_1^m, \cdots, x_J^m)$ be the result at the end of the first cycle. After w burn-in cycles, $X_{\aleph}^{(mw+1)}$ is harvested at the end of the next cycle as the first qualified simulation, and the process is repeated l times. Let the empirical distribution of $\{ X_{\aleph}^{(hmw+1)}, 1 \le h \le l\}$ be $\hat{\pi}^{(k_1, \cdots, k_m)}$, where the superscript indicates the scan pattern. Every $\hat{\pi}^{(k_1, \cdots, k_m)}$ is called a *Gibbs distribution* (Geman and Geman, 1984; Israel, 2005). The collection of $m!$ such Gibbs distributions, $\{\hat{\pi}^{(k_1, \cdots, k_m)}, (k_1, \cdots, k_m) \in S_m\}$, assuming that they all converge, is named *Gibbs ensemble* after its base procedure. To compute the error, convert each $\hat{\pi}^{(k_1, \cdots, k_m)}$ to $\{ \hat{\pi}_{\alpha_i|\bar{\alpha}_i}^{(k_1, \cdots, k_m)}, 1 \le i \le m\}$, and measure the separation of $\hat{\pi}_{\alpha_i|\bar{\alpha}_i}^{(k_1, \cdots, k_m)}$ from the corresponding $\mathcal{P}_i$ via a divergence measure. The sum of m divergences is the error of $\hat{\pi}^{(k_1, \cdots, k_m)}$. In theory, the error is zero if and only if $\mathcal{G}$ is compatible; but in simulation, such errors will not be exactly zero due to Monte Carlo variation.

There are several commonly used divergences for measuring the closeness between two distributions $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{p}$ (e.g., Bishop et al., 1975, p. 348–9). The divergences adopted in this study are: $L^2 := \Sigma(\hat{\pi}_i - p_i)^2$ (Euclidean); $F^2 := 4 \sum (\sqrt{\hat{p}_i} - \sqrt{p_i})^2$ (Freeman-Turkey); $X^2 := \Sigma(\hat{\pi}_i - p_i)^2/\hat{\pi}_i$ (Pearson's chi-square); $N^2 := \Sigma(\hat{\pi}_i - p_i)^2/p_i$ (Neyman's chi-square); $I^2 := \Sigma\hat{\pi}_i \log(\hat{\pi}_i/p_i)$ (Information); and $G^2 := \Sigma p_i \log(p_i/\hat{\pi}_i)$ (Kullback-Leibler). To calculate the divergence between two distributions that are represented by multidimensional matrices, each matrix is cast into a row vector, and the formula listed above is applied.

Let D represent one of the divergence measures, and $e_D^{(k_1, \cdots, k_m)} = \sum_{i=1}^m D(\hat{\pi}_{\alpha_i|\bar{\alpha}_i}^{(k_1, \cdots, k_m)}, \mathcal{P}_i)$ measure the D-error between $\hat{\pi}^{(k_1, \cdots, k_m)}$ and $\mathcal{G}$. The sum of D-errors over the entire ensemble is $E_D = \sum (1/e_D^{(k_1, \cdots, k_m)})$. Also, let $\varepsilon_D^{(k_1, \cdots, k_m)} = (1/e_D^{(k_1, \cdots, k_m)})/E_D$ be the weight assigned to $\hat{\pi}^{(k_1, \cdots, k_m)}$. The differentially weighted Gibbs ensemble is

$$\tilde{\pi}_D = \sum_{(k_1, \cdots, k_m) \in S_m} \varepsilon_D^{(k_1, \cdots, k_m)} \hat{\pi}^{(k_1, \cdots, k_m)}, \tag{2}$$

which is shortened to *Gibbs solution*. For example, Gibbs solution $\tilde{\pi}_L$ is weighted by the L^2-divergence. In addition, we use $\hat{\pi}_{\mathrm{EQ}}$ and $\hat{\pi}_{\mathrm{HC}}$, respectively, for the equally weighted Gibbs

solution ($\varepsilon^{(k_1,\cdots,k_m)} = 1/m!$) and the random scan Gibbs distribution (Heckerman et al., 2000). Figure 1 illustrates the process for obtaining the Gibbs solution.

For $J = 2$, let $\mathcal{P}_1$ and $\mathcal{P}_2$ correspond to the conditional distributions $(x_1|x_2)$ and $(x_2|x_1)$, respectively. The Gibbs ensemble has two distributions: $\hat{\pi}^{(1,2)}$ is the empirical distribution of $\{(x_1^{(2hw+1)}, x_2^{(2hw+1)}), 1 \le h \le l\}$ sampled in the following manner:

$$(\cdot|x_2^{(0)}) \xrightarrow{\mathcal{P}_1} (x_1^{(1)}|x_2^{(0)}) \xrightarrow{\mathcal{P}_2} (x_2^{(1)}|x_1^{(1)}) \xrightarrow{\mathcal{P}_1} (x_1^{(2)}|x_2^{(1)}) \xrightarrow{\mathcal{P}_2} \cdots,$$

and $\hat{\pi}^{(2,1)}$ is derived from $\{(x_2^{(2hw+1)}, x_1^{(2hw+1)}), 1 \le h \le l\}$, which are sampled from $\mathcal{P}_2 \to \mathcal{P}_1 \to \cdots \mathcal{P}_2 \to \mathcal{P}_1 \to \cdots$. The Gibbs solution is $\tilde{\pi}_D = \varepsilon_D^{(1,2)}\hat{\pi}^{(1,2)} + \varepsilon_D^{(2,1)}\hat{\pi}^{(2,1)}$.

For $J = 3$ and $\mathcal{G} = \{\pi_{i|\{\bar{i}\}}\}$, there are six Gibbs distributions. Let $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$ respectively denote $(x_1|x_2, x_3)$, $(x_2|x_1, x_3)$, and $(x_3|x_1, x_2)$. For permutation (1, 3, 2), the sampling scheme runs as follows:

$$\begin{aligned}(\cdot|x_2^{(0)},\ & x_3^{(0)}) \xrightarrow{\mathcal{P}_1} (x_1^{(1)}|x_2^{(0)},\\ & x_3^{(0)}) \xrightarrow{\mathcal{P}_3} (x_3^{(1)}|x_1^{(1)},\\ & x_2^{(0)}) \xrightarrow{\mathcal{P}_2} (x_2^{(1)}|x_1^{(1)}, x_3^{(1)}) \xrightarrow{\mathcal{P}_1} (x_1^{(2)}|x_2^{(1)},\\ & x_3^{(1)}) \xrightarrow{\mathcal{P}_3} (x_3^{(2)}|x_1^{(2)},\\ & x_2^{(1)}) \xrightarrow{\mathcal{P}_2} (x_2^{(2)}|x_1^{(2)},\\ & x_3^{(2)}) \to \cdots.\end{aligned}$$

Then, $\hat{\pi}^{(1,3,2)}$ is the empirical distribution of $\{(x_1^{(3hw+1)}, x_3^{(3hw+1)}, x_2^{(3hw+1)}), 1 \le h \le l\}$, and the three-dimensional Gibbs solution is the following weighted average of six Gibbs distributions:

$$\tilde{\pi}_D = \varepsilon_D^{(1,2,3)}\hat{\pi}^{(1,2,3)} + \varepsilon_D^{(1,3,2)}\hat{\pi}^{(1,3,2)} + \cdots + \varepsilon_D^{(3,2,1)}\hat{\pi}^{(3,2,1)}.$$

3. Numerical Comparisons and Simulations

We use both numerical examples and simulated data to compare the performances of LP and GE. First, we consider a 3 × 4, two-dimensional, conditional models and its incompatible variations. Using divergence measures as criteria, we compare GE and LP for this example. The second example compares the errors of GE and of LP on a 3 × 3 × 3 conditional model and its incompatible variations. Finally, we report results from a simulation study. One hundred pairs $\{\mathcal{P}_1, \mathcal{P}_2\}$ of 3×4 two-dimensional conditional distributions are randomly generated. The reported means and standard deviations of the divergences are based on 100 replications of Monte Carlo simulations, each of sample size 100, 000. The programs for simulation and analysis are developed in Matlab and can be download from the link (provided upon publication).

3.1 A Two-Dimensional Example

This example is taken from Ip and Wang (2009). The conditional distributions $\mathcal{P}_1(c, d)$ and $\mathcal{P}_2(a, b)$ are defined as follows:

$$\mathcal{P}_1(c,d)=\begin{pmatrix} \frac{1}{7} & \frac{1}{4} & \frac{3}{7}+c & \frac{1}{7}+d \\ \frac{2}{7} & \frac{2}{4} & \frac{1}{7} & \frac{2}{7} \\ \frac{4}{7} & \frac{1}{4} & \frac{3}{7}-c & \frac{4}{7}-d \end{pmatrix},\ \mathcal{P}_2(a,b)=\begin{pmatrix} \frac{1}{6}-a & \frac{1}{6}+a & \frac{3}{6} & \frac{1}{6} \\ \frac{2}{7} & \frac{2}{7} & \frac{1}{7} & \frac{2}{7} \\ \frac{2}{6}+b & \frac{1}{12}+b & \frac{1}{4}-2b & \frac{1}{3} \end{pmatrix}.$$

When $a = b = c = d = 0$, $\mathcal{P}_1$ and $\mathcal{P}_2$ are compatible. The perturbation parameters a, b, c, and d are used to locate and control the degree of incompatibility. We study the following cases:

Case (i) $a = -1/12$, $b = c = d = 0$;

Case (ii) $c = -1/7$, $d = 1/7$, $a = b = 0$;

Case (iii) $c = -1/7$, $d = 1/7$, $a = -1/12$, $b = 0$; and

Case (iv) $c = -1/7$, $d = 1/7$, $a = -1/12$, $b = 1/12$.

Case (i) and Case (ii) represent incompatibility arising from deviations in $\mathcal{P}_2(x_2|x_1 = 1)$ and deviations in $\mathcal{P}_1(x_1|x_2 = 3, 4)$, respectively. Case (iii) represents the situation when Case (i) and Case (ii) occur simultaneously. Case (iv) represents the situation in which $\mathcal{P}_2$ of Case (iii) deviates further from compatibility. Our purpose is to access the performance of GE as $\mathcal{P}_1$, and $\mathcal{P}_2$ increasingly depart from compatibility.

We follow the LP method described in Arnold et al. (2002) to compute the joint probability p_{ij} with minimum ε_{ij} such that $|p_{ij} - a_{ij}p_{\cdot j}| \le \varepsilon_{ij}$, $|p_{ij} - b_{ij}p_{i\cdot}| \le \varepsilon_{ij}$ and $\Sigma_{i,j}\, p_{ij} = 1$. Here, $\mathcal{P}_1 = (a_{ij})$, $\mathcal{P}_2 = (b_{ij})$, $p_{i\cdot} = \Sigma_j\, p_{ij}$, and $p_{\cdot j} = \Sigma_i\, p_{ij}$. Because the LP formulation contains many unknowns to be solved, the restrictions $\varepsilon_{ij} = \varepsilon$ are imposed. After such restrictions, there are 48 inequalities, 1 equality, and 13 unknowns.

We first consider the compatible case, which has a unique solution, to validate our programs. We found that the joint distribution of LP optimization, denoted as $\hat{\pi}_{\mathrm{LP}}$, indeed had zero divergence, as expected. Table 2 shows the more interesting results regarding the accuracy of Gibbs sampler, which is subject to simulation variations. Gibbs samplers, both fixed scan and random scan, were run with 5, 000 burn-in cycles, and retain the subsequent 100, 000 pairs of (x_1, x_2). The resulting joint distributions are denoted as $\hat{\pi}^{(1,2)}$ and $\hat{\pi}_{(1/2,1/2)}$, respectively, whereas π_F is the GE solution of (2), with D being the Freeman-Tukey divergence. To assess the Monte Carlo errors, the same Gibbs samplers are repeated 100 times so that the means and standard deviations of the divergences can be computed. From Table 2 we observe that the $\tilde{\pi}_F$ has considerably smaller mean divergences than both $\hat{\pi}^{(1,2)}$ and $\hat{\pi}_{(1/2,1/2)}$. For example, the mean in L^2 is reduced by more than 50% from 1.81E-4 to 7.44E-5. All divergence measures are computed with respect to the unique joint distribution. It is also observed that while the fixed scan and random scan have similar accuracies, GE solutions reduce all of the divergences by more than one half.

For the incompatible Case (i) to Case (iv), Table 3 lists the divergences of two fixed scans and the GE solution weighted by L^2, G^2, and F^2, respectively. In general, GE solutions achieve significant reductions of divergence, while the performance of LP is slightly worse than $\hat{\pi}^{(2,1)}$. For Case (i), the F -divergence of $\tilde{\pi}_F$ is 40.8% of that of $\hat{\pi}^{(2,1)}$; the L^2-divergence of $\tilde{\pi}_L$ is only 33.5% of that of $\hat{\pi}^{(2,1)}$. For Case (iv), $\tilde{\pi}_F$ and $\tilde{\pi}_L$ reduce the F^2 and L^2 of $\hat{\pi}^{(2,1)}$ by 31.6% and 39.8%, respectively, while the reductions of F^2 and L^2 of $\hat{\pi}^{(1,2)}$ are 60.1% (by $\tilde{\pi}_F$) and 39.8% ($\tilde{\pi}_L$), respectively. Moreover, the F^2 of $\hat{\pi}_{\mathrm{LP}}$ is 148.1% of that of $\tilde{\pi}_F$, and the L^2 of $\hat{\pi}_{\mathrm{LP}}$ is 152.2% of that of $\tilde{\pi}_L$. Later, we will show that such reductions are generally

observed in randomly generated, two-dimensional — and incompatible — conditional models.

3.2 A Three-Dimensional Example

The three conditional distributions, $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$, are listed in Table 4. We summarize the comparison of the compatible case plus cases (i) and (iv) in Table 5. Besides listing the L^2 and F^2, the percentages of reduction relative to $\hat{\pi}^{(1,2,3)}$ are used to demonstrate the performance of GE. The compatible case allows us to examine the Monte Carlo simulation variations and to explore ways to reduce such variations. For the six supposedly identical Gibbs distributions, their L^2 ranges from 1.466E-3 to 1.551E-3. Here, π_{EW} denotes the equally weighted average of three Gibbs distributions: $\hat{\pi}^{(1,2,3)}$, $\hat{\pi}^{(2,3,1)}$, and $\hat{\pi}^{(3,1,2)}$. From Table 5, $\tilde{\pi}_{\mathrm{EW}}$ reduces the L^2 and F^2 of $\hat{\pi}^{(1,2,3)}$ by 65.6% and 64.5%, respectively, and $\hat{\pi}_{\mathrm{EQ}}$, which uses all six Gibbs distributions, further reduces L^2 and F^2 of $\tilde{\pi}_{\mathrm{EW}}$ by an additional 51.4%. The best GE, $\tilde{\pi}_F$, reduces the L^2 and F^2 of $\hat{\pi}_{\mathrm{EQ}}$ by another 7%. In summary, combining six Gibbs distributions is better than combining only three, and the use of different kinds of weights seems to play a relatively minor role. Case (i) to Case (iv) represent gradual departure from compatibility. The reductions of L^2 for $\tilde{\pi}_F$ over $\hat{\pi}^{(1,2,3)}$ are around 38% to 52% for Case (i) and Case (iv), respectively. In all four cases, the divergence of $\hat{\pi}_{\mathrm{LP}}$ is considerable larger than that of $\hat{\pi}^{(1,2,3)}$ (except L^2 in Case (iv)).

3.3 Simulation Experiments

In both Examples 3.1 (four cases) and 3.2 (another four cases), we further conducted 100 replications of Gibbs sampler. We try to address the question: Is it possible that LP outperforms GE in some of the replications? We found that for all eight different conditional models, GE outperformed LP in *every* replication. This implies that even the worst-case GE — due to Monte Carlo variation — still has smaller divergences than LP. Furthermore, in Example 11 of Arnold et al. (2002) (details not reported here), for a 3 × 4 models with two structural zeros, we also observed that out of 100 replications, GE, without exception, had smaller divergence than LP.

To confirm that the superiority of GE over LP and Gibbs distributions is not just an artifact of the examples we selected, we randomly generate pairs of 3 × 4 matrices with positive integers between 1 and 100 as entries. Then we normalize one matrix so that its column entries sum to 1, and we use it as $\mathcal{P}_1$. The other matrix is similarly normalized (in rows) into $\mathcal{P}_2$. For every randomly generated pair, LP, fixed scan Gibbs distribution and GE solution with different weights are computed. The mean and standard deviation of the percentages of reduction of the LP and GE solutions with respect to $\hat{\pi}^{(1,2)}$ and $\hat{\pi}^{(2,1)}$ are summarized in Table 6. In the course of the simulation, we noticed that the LP approach sometimes produced unreasonable joint distributions when there were multiple optimal solutions. Because the randomly generated conditional models do not contain any zero entry, their joint distribution is not expected to contain a zero entry. However, LP produces 11 joints (out of 100) with some zero entries, and one of the joints has an entire column of zeros. Those with any zero entry are excluded from comparison.

The results in Table 6 indicate that the GE approach outperforms the LP approach and that the GE solutions enjoy substantial divergence reduction with respect to both $\hat{\pi}^{(1,2)}$ and $\hat{\pi}^{(2,1)}$. For example, the average percentages of reductions in L^2 are 63.9% and 36.2% relative to the fixed scan $\hat{\pi}^{(1,2)}$ and $\hat{\pi}^{(2,1)}$, respectively, while the same averages for LP are 38.1% and −15.1%, respectively. Figure 2 shows the two boxplots of percentages of reductions in L^2 of $\tilde{\pi}_L$ and $\hat{\pi}_{\mathrm{LP}}$ relative to $\hat{\pi}^{(1,2)}$ (left panel), and to $\hat{\pi}^{(2,1)}$ (right panel). The other distributions are similar and not shown because of space limitation. These simulation results give us confidence that the weighted average of ensembles can produce a joint distribution that fits

the conditional models significantly better than LP and the Gibbs sampler. When we compare the coefficients of variation using values in Table 6, we find that the percentages of reduction for the LP approach is far less predictable than those of GE, whose coefficients of variation range from 8.0% to 26.3%.

4. A Real Example

Table 7 is taken from Toffoli et al. (2006), which constitutes one of the largest prospective studies conducted to date to investigate the relationship between polymorphism in the gene region UGT1A1*28 and response to irinotecan for metastatic colorectal cancer patients. Toffoli et al. (2006) observed a significant increased risk of developing severe hematologic toxicity among patients carrying the TA_7 allele. The hypothesis is that genetic testing for UGT1A1*28 polymorphism may have utility as a predictor of response to irinotecan. In Table 7, the row variable X_1 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 treatment of a combination of irinotecan flourourail and leucovorin, which is represented by the column variable X_2. The four categories of X_2 are complete response, partial response, stable disease, and progressive disease, respectively coded as 1–4.

Clinicians commonly use two conditional models for such data: the diagnostic model $\mathcal{P}_1(x_1|x_2)$ and the treatment model $\mathcal{P}_2(x_2|x_1)$. Of practical interest are the following sets of parameters: the diagnostic odds $d_{ij} = P(x_1 = i|x_2 = j)/P(x_1 = i|x_2 = j + 1)$, $1 \le i \le 3$, $1 \le j \le 3$, and the response odds $t_{ij} = P(x_2 = j|x_1 = i)/P(x_2 = j|x_1 = i + 1)$, $1 \le i \le 2$, $1 \le j \le 4$. We consider the following conditional models for $\mathcal{P}_1$ and $\mathcal{P}_2$ in terms of d_{ij} and t_{ij}:

Model A $d_{ij}/d_{ij+1} = \delta$ and $t_{ij}/t_{i+1,j} = \xi$ for all permissible i, j;

Model B $d_{ij}/d_{ij+1} = \delta_i$ and $t_{ij}/t_{i+1,j} = \xi_j$ for all permissible i, j; and

Model C Logistic regression for both $\mathcal{P}_1$ and $\mathcal{P}_2$.

Using maximum-likelihood methods, we estimated the conditional distributions for all three models. Specifically, in Model C, $\mathcal{P}_1(x_1|x_2)$ was estimated by applying multinomial logistic regression of x_1 on x_2, and $\mathcal{P}_2(x_2|x_1)$ was estimated by applying ordinal logistic regression of x_2 on x_1. Model A produces compatible $\mathcal{P}_1$ and $\mathcal{P}_2$, while Models B and C do not. Ip and Wang (2009) show that when the odds ratios across the conditional distributions are identical, as in the case of Model A, then there exists a unique joint distribution. Table 8 shows the three pairs of respective conditional distributions under Models A, B, and C. For every conditional pairs, we compute the joint distributions of LP ($\hat{\pi}_{\mathrm{LP}}$), fixed scan ($\hat{\pi}^{(1,2)}$, and $\hat{\pi}^{(2,1)}$), and GE solutions via different weights ($\tilde{\pi}_L$, $\tilde{\pi}_F$, and $\tilde{\pi}_G$). Table 9 compares the means and standard deviations of G^2 divergence between the observed joint distribution (Table 7) and the estimated joint distributions out of 100 Monte Carlo replications.

Because Model A is compatible, both LP and Gibbs sampler converge to the same joint distribution, and GE offers no advantage over $\hat{\pi}_{\mathrm{LP}}$ and $\hat{\pi}^{(1,2)}$, as expected. For incompatible Model B, every GE-based joint distributions outperforms both LP and fixed scan solutions. The coefficients of variation in G^2 for $\tilde{\pi}_L$, $\tilde{\pi}_F$, and $\tilde{\pi}_G$ were all below 7%, suggesting that the divergences of GE-based distributions are consistent over Monte Carlo replications.

For Model C (Table 9), GE also outperforms LP and fixed scan. Several observations can be made. First, the standard deviations of the replications are slightly smaller than the standard deviations under Model B. The coefficients of variation for both Model B and Model C are less than 5%, again suggesting that GE are quite robust to Monte Carlo variations. Second, the divergences of LP in Model C are 20% to 70% larger than those in Model B across

different divergence measures. Note that the observed values of the conditional distribution for $P(x_1|x_2 = 2)$ are (0.400, 0.4706, 0.1294), which is close to the estimated values for the corresponding cell (0.3877, 0.5008, 0.1115) under Model C.

If only $\hat{\pi}_{\mathrm{LP}}$ were computed and we had adopted the G^2 divergence of $\hat{\pi}_{\mathrm{LP}}$ as the sole criterion for selecting model, then Model B would hands-down win over Model C. However, we have computed GE solution and according to their G^2 divergence, the advantage of Model B over Model C is only marginal (Table 9). This shows that goodness-of-fit indices can be quite different when using LP and GE.

The G^2-statistic for Model A, B and C are respectively 6.5069, 2.4119 and 2.5780; whereas the Akaike's information criterion (AIC) value are respectively 10.5069, 12.4119 and 12.5780. The G^2-statistic is the G^2-divergence multiplied by twice the sample size, here 2×238. Model A has the smallest value of AIC, which suggests that model A is more efficient (adjusted for model complexity) than models B and C. Using Model A as the basis, the diagnostic odds (d_{ij}) and response odds (t_{ij}) matrices calculated from its joint distribution are as follows:

$$(d_{ij}) = \begin{pmatrix} 0.8731 & 0.8861 & 0.8983 \\ 1.0520 & 1.0673 & 1.0822 \\ 1.2669 & 1.2871 & 1.3040 \end{pmatrix}, (t_{ij}) = \begin{pmatrix} 0.7143 & 0.8610 & 1.0372 & 1.2495 \\ 0.7375 & 0.8883 & 1.0698 & 1.2891 \end{pmatrix}.$$

Matrix (d_{ij}) is partitioned horizontally by genotype along the line of locus TA7; the values of d_{ij} in the row of TA6/TA6 are all less than 1, while the rows corresponding to TA6/TA7 and TA7/TA7 are all larger than 1. Moreover, matrix (t_{ij}) is partitioned vertically between patients who respond positively to irinotecan versus those who do not benefit from the treatment. That is, the response odds are less than 1 for $x_2 = 1$ and 2 and larger than 1 for $x_2 = 3$ (no response) and $x_2 = 4$ (worsening response).

An interesting thought from examining the goodness-of-fit indices is the possibility of using the Gibbs solution to replace the pseudolikelihood in inference. Under Model C of Table 8, the G^2-statistic for $\mathcal{P}_1(x_1|x_2)$ is 4.03 (p-value = 0.2578) and the G^2-statistic for $\mathcal{P}_2(x_2|x_1)$ is 6.31 (p-value = 0.0426). Therefore, the combined G^2-statistic for pseudolikelihood $\mathcal{P}_1(x_1|x_2)\mathcal{P}_2(x_2|x_1)$ is $(4.03 + 6.31)/2 = 5.17$, where the averaging is to compensate for using the observed table twice. The G^2-statistic of $\tilde{\pi}_G$ under Model C is 2.578 (p-value = 0.7647), which is less than half of 5.17. Hence, the estimated joint distribution $\tilde{\pi}_G$ is a closer representation of the observed table than the pseudolikelihood. The pros and cons of these two approaches deserve further study.

5. Discussions

For high-dimensional data, a reduced model may be formulated in at least two ways: as an undirected graphical model in which all variables are considered jointly or as a system of univariate conditional models, which can be depicted as a cyclic, directed graph (Heckerman et al., 2000). In the second approach, as observed by Dobra et al. (2004), the conditionally specified models almost surely do not cohere to a proper joint distribution. It is therefore important to study and compare the joint distributions obtained through different methods of estimation. In this paper, we compare the Gibbs distribution, LP, and GE when conditional models are not compatible.

Our results have been primarily based on empirical studies. Liu (1996) provided several observations about the behavior of Gibbs distributions that are consistent with our empirical work. As long as the transition matrix corresponding to (1) is irreducible and aperiodic, the

Gibbs sampler has a recurrent joint distribution. For two-dimensional conditional models that are not compatible, the two fixed scanned Gibbs (recurrent) distributions $\tilde{\pi}^{(1,2)}$ and $\tilde{\pi}^{(2,1)}$ are known to be the two extreme points in the space of all the Gibbs distributions. The usual random scan Gibbs distribution has two characteristics: (a) any of its odds ratios will be sandwiched between the corresponding odds ratios of $\tilde{\pi}^{(1,2)}$ and $\tilde{\pi}^{(2,1)}$, and (b) its one-dimensional marginal distributions are identical to those of $\tilde{\pi}^{(1,2)}$. Briefly, the space of all the Gibbs distributions consists of joints having the same marginals and having odds ratios that are confined to ranges determined by $\tilde{\pi}^{(1,2)}$ and $\tilde{\pi}^{(2,1)}$.

In this paper, we use divergences to compute the proper weights for combining the extreme Gibbs distributions. Thus, the GE solutions represent mixtures of the Gibbs distributions. For three variables with $\mathcal{P}_i(x_i|x_j, j \neq i)$, the space of all Gibbs distributions is more complicated: (a) only the invariant interactions (Ip and Wang, 2009), not all of the odds ratios, can reoccur in the Gibbs distributions; (b) there are four different two-dimensional marginal distributions; and (c) there are three different one-dimensional marginal distributions shared by 3! = 6 distinct fixed scan Gibbs distributions. Every GE solution is a mixture of $\tilde{\pi}^{(i,j,k)}$ and $(i, j, k) \in S_3$, and its combination weights can be computed via the selected divergence.

Two observations can be made about the mixture distributions. First, the number of odds ratios are $(K_1 - 1)(K_2 - 1)$ and $(K_1 - 1)(K_2 - 1)(K_3 - 1) + \Sigma(K_i - 1)(K_j - 1)$ for (X_1, X_2) and (X_1, X_2, X_3), respectively, where X_i assumes K_i values. The optimization problem becomes more challenging because the number of parameters grows quickly with the number of variables. GE reduces the optimizations to one-dimensional problems instead of searching for optimal odds ratios in every range. Second, the marginal distributions of GE are confined to those that can be generated by the Gibbs samplers. It is possible that some perturbations of the marginal distributions could reduce the divergence. However, we observe that in the two-dimensional case the marginals of LP are always close to the marginals of GE. For the example in Section 4, the three $\tilde{\pi}$'s under Model A, Model B, and Model C produce nearly identical X_1-marginal and X_2-marginal distributions. Thus, we conjecture that the marginals of Gibbs distributions are not far from optimal.

The GE offers a computationally feasible approach in obtaining optimal or nearly optimal solutions for conditionally specified models. Unlike methods based on LP, the GE approach described here can be generalized to high-dimensional problems in straightforward ways. The computational burden is also more manageable than with LP. Furthermore, we provide evidence that the performance of GE is at least comparable, if not better, than LP. Finally, we show that Gibbs distributions can be improved upon by using ensembles. By varying scan patterns to generate ensembles of Gibbs distributions, we believe that the power of GE can be further expanded. Instead of using pre-determined probabilities to perform random scan, one may be able to further refine the search in the space of a Gibbs distribution for a better solution for nearly compatible and incompatible conditional models.

References

Arnold BC, Castillo E, Sarabia JM. Exact and near compatibility of discrete conditional distributions. Computational Statistics and Data Analysis 2002;16:231–252.

Arnold BC, Gokhale DV. Distributions most nearly compatible with given families of conditional distributions. Test 1998;7:377–390.

Besag JE. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal Royal Statistical Society B 1974;36:192–236.

Besag JE. Efficiency of pseudo-likelihood estimators for simple Gaussian fields. Biometrika 1977;64:616–618.

Bühlmann, P. Bagging, boosting and ensemble methods. In: Gentle, J.; Härdle, W.; Mori, Y., editors. Handbook of Computational Statistics: Concepts and Methods. Springer-Verlag; Berlin, Heidelberg: 2004. p. 877-907.

Dobra A, Hans C, Jones B, Nevins J, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis 2004;90:196–212.

Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 1984;6:721–741.

Heckerman D, Chickering DM, Meek C, Rounthwaite R, Kadie C. Dependence networks for inference, collaborative filtering and data visualization. Journal of Machine Learning and Research 2000;1:49–57.

Ip EH, Wang YJ. Canonical representation of conditionally specified multivariate discrete distributions. Journal of Multivariate Analysis 2009;100:1282–1290.

Israel, RB. Gibbs distributions ii. In: Kotz, S.; Read, C.; Balakrishnan, N.; Vidakovic, B., editors. Encyclopedia of Statistical Sciences. 2. Vol. 4. John Wiley & Sons, Inc; New York: 2005. p. 2838-2843.

Jensen JL, Künsch HR. On asymptotic normality of pseudo likelihood estimates for pairwise onteraction processes. Annals of the Institute of Statistical Mathematics 1994;46:475–486.

Jensen JL, Møller J. Pseudolikelihood for exponential family models of spatial point processes. Annals of Applied Probability 1991;1:455–461.

Levine R, Casella G. Optimizing random scan Gibbs samplers. Journal of Multivariate Analysis 2006;97:2071–2100.

Liu JS. Discussions on statistical theory and monte carlo algorithms (with discussions). Test 1996;5:305–310.

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, Frustaci S. The role of UGT1A1*28 polymorphism in the pharmacodynamics and pharmacokinetics of irinotecan in patients with metastatic colorectal cancer. Journal of Clinical Oncolog 2006;24:3061–3068.

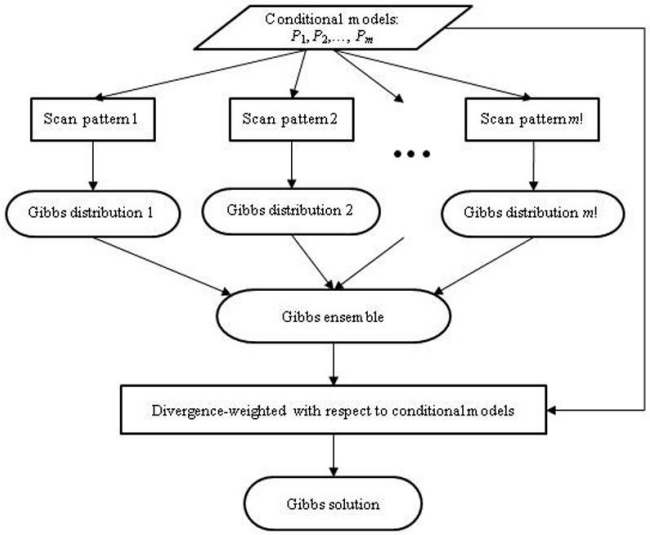

Figure 1.
Flow diagram for Gibbs ensemble

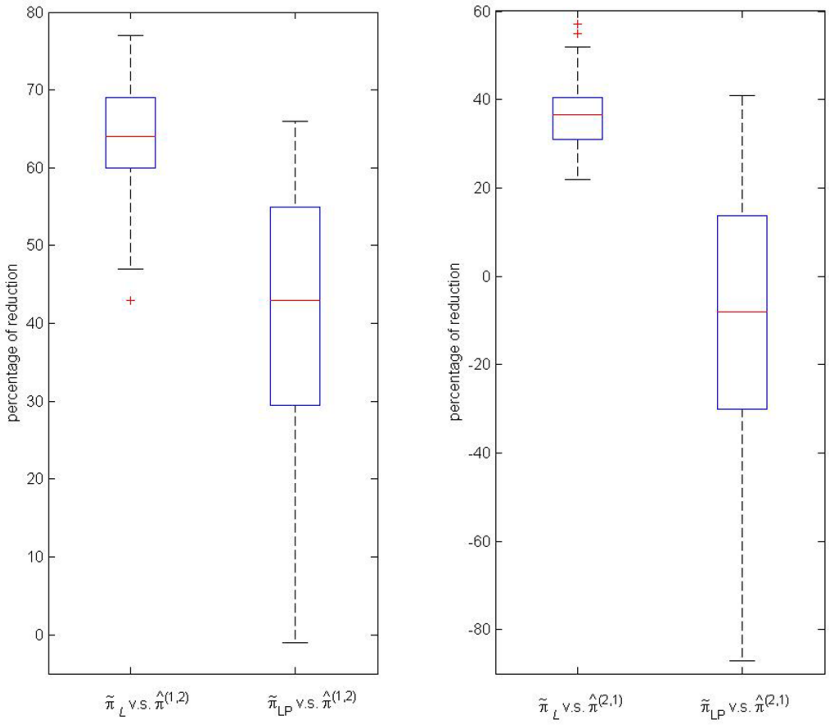

Figure 2.

Box plots for the percentages of reductions in L^2 of $\tilde{\pi}_L$ and $\hat{\pi}_{LP}$ relative to $\hat{\pi}^{(1,2)}$ (left panel), and to $\hat{\pi}^{(2,1)}$ (right panel)

The outliers of $\hat{\pi}_{LP}$ relative to $\hat{\pi}^{(1,2)}$ (3 models with percentages of reductions less than −1) and to $\hat{\pi}^{(2,1)}$ (3 models with percentages of reductions less than −87) are not shown in figure.

Table 1

Divergence of the joint distributions computed via different algorithms for Example 1.1

Errors	$\hat{\pi}_{\mathbf{LP}}$	$\hat{\pi}_{\mathbf{HC}}$	$\tilde{\pi}_{\mathbf{EQ}}$	$\tilde{\pi}_{L}$	$\tilde{\pi}_{I}$	$\tilde{\pi}_{X}$	$\tilde{\pi}_{F}$
L^2	0.0910	0.0939	0.0926	0.0700	0.0711	0.0771	0.0721
I^2	0.1134	0.1165	0.1148	0.0972	0.0963	0.1005	0.0968
X^2	0.2094	0.2140	0.2119	0.2137	0.2018	0.1977	0.1996
F^2	0.2201	0.2257	0.2229	0.1964	0.1928	0.1981	0.1930

Table 2

Performance of Gibbs ensemble for the compatible case of Example 3.1

	Mean Divergence (Standard Deviation)					
	L^2	I^2	G^2	X^2	N^2	F^2
$\tilde{\pi}_F$	7.44E-5	1.38E-4	1.38E-4	2.75E-4	2.75E-4	2.75E-4
Gibbs Ensemble	(3.38E-5)	(6.01E-5)	(6.01E-5)	(1.20E-4)	(1.20E-4)	(1.20E-4)
$\hat{\pi}_{(1/2,1/2)}$	1.57E-4	2.92E-4	2.92E-4	5.85E-4	5.84E-4	5.85E-4
Random Scan	(6.43E-5)	(1.21E-4)	(1.21E-4)	(2.41E-4)	(2.41E-4)	(2.41E-4)
$\hat{\pi}^{(1,2)}$	1.81E-4	3.30E-4	3.31E-4	6.61E-4	6.61E-4	6.61E-4
Fixed Scan	(8.74E-5)	(1.51E-4)	(1.51E-4)	(3.03E-4)	(3.02E-4)	(3.03E-4)

Table 3

Mean divergences for the incompatible cases of Example 3.1 based on 100 Monte Carlo replications

		L^2	G^2	F^2	I^2
Case (i)	$\hat{\pi}^{(1,2)}$	0.0224	0.0510	0.0971	0.0466
	$\hat{\pi}^{(2,1)}$	0.0114	0.0332	0.0680	0.0351
	$\tilde{\pi}_L$	0.0075	0.0201	0.0409	0.0208
	$\tilde{\pi}_G$	0.0076	0.0197	0.0397	0.0201
	$\tilde{\pi}_F$	0.0077	0.0196	0.0396	0.0200
	$\hat{\pi}_{\mathrm{LP}}$	0.0117	0.0280	0.0557	0.0277
Case (ii)	$\hat{\pi}^{(1,2)}$	0.0819	0.1237	0.2386	0.1163
	$\hat{\pi}^{(2,1)}$	0.0697	0.1116	0.2304	0.1203
	$\tilde{\pi}_L$	0.0376	0.0586	0.1190	0.0606
	$\tilde{\pi}_G$	0.0377	0.0584	0.1185	0.0603
	$\tilde{\pi}_F$	0.0378	0.0584	0.1182	0.0600
	$\hat{\pi}_{\mathrm{LP}}$	0.0706	0.1093	0.2191	0.1101
Case (iii)	$\hat{\pi}^{(1,2)}$	0.1038	0.1743	0.3349	0.1623
	$\hat{\pi}^{(2,1)}$	0.0808	0.1446	0.2987	0.1559
	$\tilde{\pi}_L$	0.0454	0.0787	0.1599	0.0815
	$\tilde{\pi}_G$	0.0455	0.0785	0.1592	0.0810
	$\tilde{\pi}_F$	0.0456	0.0784	0.1587	0.0806
	$\hat{\pi}_{\mathrm{LP}}$	0.0788	0.1356	0.2741	0.1391
Case (iv)	$\hat{\pi}^{(1,2)}$	0.3297	0.5376	1.0004	0.4776
	$\hat{\pi}^{(2,1)}$	0.1529	0.3210	0.6637	0.3517
	$\tilde{\pi}_L$	0.1046	0.1998	0.4140	0.2168
	$\tilde{\pi}_G$	0.1062	0.1948	0.4018	0.2090
	$\tilde{\pi}_F$	0.1078	0.1942	0.3995	0.2072

	L^2	G^2	F^2	I^2
$\hat{\pi}_{1,P}$	0.1592	0.3124	0.6163	0.3068

Table 4

Three conditional distributions of Example 3.2

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$
$(\cdot, \cdot, 1)$	$\begin{pmatrix} \frac{6}{15}+a & \frac{1}{5} & \frac{8}{14} \\ \frac{5}{15} & \frac{7}{10} & \frac{2}{14} \\ \frac{4}{15}-a & \frac{1}{10} & \frac{4}{14} \end{pmatrix}$	$\begin{pmatrix} \frac{6}{12} & \frac{2}{12} & \frac{4}{12} \\ \frac{5}{13}+b & \frac{7}{13} & \frac{1}{13}-b \\ \frac{4}{7} & \frac{1}{7} & \frac{2}{7} \end{pmatrix}$	$\begin{pmatrix} \frac{6}{15} & \frac{2}{18} & \frac{4}{12} \\ \frac{5}{17} & \frac{7}{16} & \frac{1}{8} \\ \frac{4}{12}+c & \frac{1}{13} & \frac{2}{12} \end{pmatrix}$
$(\cdot, \cdot, 2)$	$\begin{pmatrix} \frac{5}{13}+e & \frac{8}{17} & \frac{1}{11} \\ \frac{4}{13}+e & \frac{2}{17} & \frac{6}{11} \\ \frac{4}{13}-2e & \frac{7}{17} & \frac{4}{11} \end{pmatrix}$	$\begin{pmatrix} \frac{5}{14} & \frac{8}{14} & \frac{1}{14} \\ \frac{4}{12} & \frac{2}{12} & \frac{6}{12} \\ \frac{4}{15} & \frac{7}{15} & \frac{4}{15} \end{pmatrix}$	$\begin{pmatrix} \frac{5}{15} & \frac{8}{18} & \frac{1}{12} \\ \frac{4}{17} & \frac{2}{16} & \frac{6}{8} \\ \frac{4}{12}-c & \frac{7}{13} & \frac{4}{12} \end{pmatrix}$
$(\cdot, \cdot, 3)$	$\begin{pmatrix} \frac{4}{16}+d & \frac{8}{20} & \frac{7}{14} \\ \frac{8}{16} & \frac{7}{20} & \frac{1}{14} \\ \frac{4}{16}-d & \frac{5}{20} & \frac{6}{14} \end{pmatrix}$	$\begin{pmatrix} \frac{4}{19} & \frac{8}{19} & \frac{7}{19} \\ \frac{8}{16} & \frac{7}{16} & \frac{1}{16} \\ \frac{4}{15}-f & \frac{5}{15}-g & \frac{6}{15}+f+g \end{pmatrix}$	$\begin{pmatrix} \frac{4}{15} & \frac{8}{18} & \frac{7}{12} \\ \frac{8}{17} & \frac{7}{16} & \frac{1}{8} \\ \frac{4}{12} & \frac{5}{13} & \frac{6}{12} \end{pmatrix}$

When $a = b = c = d = e = f = g = 0$, the three conditional distributions are compatible. There are four cases:

(i) $a = -3/15$, $e = 1/13$, $d = -3/16$ and $b = c = f = g = 0$

(ii) $a = -3/15$, $e = 1/13$, $d = -3/16$, $b = -2/13$ and $c = f = g = 0$

(iii) $a = -3/15$, $e = 1/13$, $d = -3/16$, $b = -2/13$, $c = -2/12$ and $f = g = 0$

(iv) $a = -3/15$, $e = 1/13$, $d = -3/16$, $b = -2/13$, $c = -2/12$, $f = 7/30$ and $g = 4/15$.

Table 5

Mean of divergences reduction vs. $\hat{\pi}^{(1,2,3)}$ of Example 3.2

		L^2	Percentage of Reduction	F^2	Percentage of Reduction
Compatible	$\hat{\pi}^{(1,2,3)}$	1.47E-3	–	4.98E-3	–
	$\tilde{\pi}_{\mathrm{EW}}$	5.04E-4	65.63%	1.77E-3	64.49%
	$\tilde{\pi}_{\mathrm{EQ}}$	2.45E-4	83.30%	8.59E-4	82.74%
	$\tilde{\pi}_{L}$	2.25E-4	84.63%	8.06E-4	89.82%
	$\tilde{\pi}_{F}$	2.28E-4	84.47%	7.97E-4	83.99%
Case (i)	$\hat{\pi}^{(1,2,3)}$	0.1846	–	0.6943	–
	$\tilde{\pi}_{\mathrm{EW}}$	0.1195	35.26%	0.4394	36.71%
	$\tilde{\pi}_{\mathrm{EQ}}$	0.1128	38.89%	0.4214	39.31%
	$\tilde{\pi}_{L}$	0.1148	37.82%	0.4411	36.47%
	$\tilde{\pi}_{F}$	0.1145	38.01%	0.4287	36.82%
	$\hat{\pi}_{\mathrm{LP}}$	0.3143	−70.24%	1.1127	−60.27%
Case (iv)	$\hat{\pi}^{(1,2,3)}$	0.9191	–	2.9883	–
	$\tilde{\pi}_{\mathrm{EW}}$	0.4719	48.65%	1.6807	43.76%
	$\tilde{\pi}_{\mathrm{EQ}}$	0.4366	52.50%	1.5959	46.59%
	$\tilde{\pi}_{L}$	0.4369	52.46%	1.5968	46.56%
	$\tilde{\pi}_{F}$	0.4392	52.21%	1.6034	46.34%
	$\tilde{\pi}_{\mathrm{LP}}$	0.8791	4.36%	3.1423	−5.15%

Table 6

Mean and standard deviation of the percentage of reduction relative to $\hat{\pi}^{(1,2)}$ and $\hat{\pi}^{(2,1)}$ based on 100 randomly generated 3 × 4 conditional models

	L^2		F^2		G^2	
	$\hat{\pi}^{(1,2)}$	$\hat{\pi}^{(2,1)}$	$\hat{\pi}^{(1,2)}$	$\hat{\pi}^{(2,1)}$	$\hat{\pi}^{(1,2)}$	$\hat{\pi}^{(2,1)}$
$\tilde{\pi}_L$	63.89% (7.10)	36.19% (7.02)	55.74% (4.44)	40.30% (5.26)	59.19% (6.70)	47.15% (8.02)
$\tilde{\pi}_F$	34.52% (8.28)	63.12% (6.68)	40.98% (4.66)	56.18% (4.75)	48.31% (7.48)	60.00% (6.81)
$\tilde{\pi}_G$	62.40% (6.94)	33.26% (8.75)	55.16% (4.96)	39.58% (5.30)	59.53% (6.98)	47.72% (7.53)
$\hat{\pi}_{\mathrm{LP}}$*	38.08% (23.12)	−15.12% (52.87)	27.12% (27.83)	−0.10% (42.11)	26.86% (40.55)	4.62% (54.99)

* Based on 89 randomly generated conditional models.

Table 7

Cross-tabulation by UGT1A1*28 polymorphism and response to treatment (Toffoli et al., 2006)

Polymorphism (X_1)	Response to chemotherapy treatment (X_2)				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 8

Conditional models for genetic data

Model	Conditional Distributions	
	$\mathcal{P}_1(x_1\|x_2)$	$\mathcal{P}_2(x_2\|x_1)$
A	$\begin{pmatrix} 0.3634 & 0.4162 & 0.4697 & 0.5229 \\ 0.5038 & 0.4789 & 0.4487 & 0.4146 \\ 0.1329 & 0.1049 & 0.0815 & 0.0625 \end{pmatrix}$	$\begin{pmatrix} 0.0600 & 0.3246 & 0.2844 & 0.3310 \\ 0.0840 & 0.3770 & 0.2742 & 0.2649 \\ 0.1139 & 0.4244 & 0.2563 & 0.2055 \end{pmatrix}$
B	$\begin{pmatrix} 0.3953 & 0.4396 & 0.4686 & 0.4869 \\ 0.4005 & 0.4400 & 0.4633 & 0.4756 \\ 0.2042 & 0.1205 & 0.0681 & 0.0375 \end{pmatrix}$	$\begin{pmatrix} 0.0775 & 0.3027 & 0.2734 & 0.3463 \\ 0.0750 & 0.3889 & 0.2815 & 0.2547 \\ 0.0691 & 0.4762 & 0.2762 & 0.1785 \end{pmatrix}$
C	$\begin{pmatrix} 0.5052 & 0.3877 & 0.4476 & 0.5388 \\ 0.4226 & 0.5008 & 0.4631 & 0.3975 \\ 0.0722 & 0.1115 & 0.0893 & 0.0637 \end{pmatrix}$	$\begin{pmatrix} 0.0648 & 0.3379 & 0.2861 & 0.3112 \\ 0.0679 & 0.3469 & 0.2847 & 0.3005 \\ 0.1677 & 0.4944 & 0.2034 & 0.1345 \end{pmatrix}$

Table 9

Mean and standard deviation of G^2 divergence across 100 replications for the conditional models of genetic data

	G^2 divergence		
	Model A	**Model B**	**Model C**
$\tilde{\pi}^{(1,2)}$	1.367E-2 (4.626E-4)	1.021E-2 (4.449E-4)	6.733E-3 (3.671E-4)
$\hat{\pi}_L$	1.367E-2 (3.175E-4)	5.037E-3 (2.253E-4)	5.892E-3 (2.107E-4)
$\tilde{\pi}_F$	1.367E-2 (3.290E-4)	5.144E-3 (2.190E-4)	5.566E-3 (1.997E-4)
$\tilde{\pi}_G$	1.367E-2 (3.291E-4)	5.067E-3 (2.221E-4)	5.416E-3 (2.020E-4)
$\hat{\pi}_{\mathrm{LP}}$	1.369E-2	7.899E-3	1.213E-2