Journal of Multivariate Analysis 167 (2018) 171–180

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Simulating conditionally specified models

Kun-Lin Kuo [a,*], Yuchung J. Wang [b]

[a] Institute of Statistics, National University of Kaohsiung, Kaohsiung, Taiwan
[b] Department of Mathematical Sciences, Rutgers University, Camden, NJ 08102, USA

ARTICLE INFO

Article history:
Received 15 May 2017
Available online 4 May 2018

AMS subject classifications:
primary 65C05
65C60
secondary 68T05

Keywords:
Dependence network
Faster convergence
Multiple imputation
Non-full conditional specification
Partially collapsed Gibbs sampler
Valid scan order

ABSTRACT

Expert systems routinely use conditional reasoning. Conditionally specified statistical models offer several advantages over joint models; one is that Gibbs sampling can be used to generate realizations of the model. As a result, full conditional specification for multiple imputation is gaining popularity because it is flexible and computationally straightforward. However, it would be restrictive to require that every regression/classification must involve all of the variables. Feature selection often removes some variables from the set of predictors, thus making the regression local. A mixture of full and local conditionals is referred to as a partially collapsed Gibbs sampler, which often achieves faster convergence due to reduced conditioning. However, its implementation requires choosing a correct scan order. Using an invalid scan order will bring about an incorrect transition kernel, which leads to the wrong stationary distribution. We prove a necessary and sufficient condition for Gibbs sampling to correctly sample the joint distribution. We propose an algorithm that identifies all of the valid scan orders for a given conditional model. A forward search algorithm is discussed. Checking compatibility among conditionals of different localities is also discussed.

1. Introduction

Expert systems routinely use conditional reasoning. In medical diagnosis, for example, the probability of a specific disease is always stated conditionally, based on the results of various clinical tests and risk factors. Likewise in multiple imputation, missing values are imputed conditionally on the observed values of a subject. The advent of graphical tools facilitates the depiction of conditional reasoning.

In directed acyclic-graph (DAG) modeling [13], risk factors and tests results are presented as parental nodes that have a direct causal effect on the disease, which is called a child node. Subsequently, the disease, along with treatments, become the parental nodes toward associated complications. Diagnostic systems based on DAGs can become highly complex, making the learning of the joint distribution computationally burdensome. Because DAG does not allow feedbacks from child nodes, Heckerman et al. [7] argue for the use of a dependence network (DN) in the context of machine learning.

Graphically, a DN is represented by a directed cyclic graph, which is in essence a collection of conditional distributions. Briefly, a DN is built using a two-step approach: (i) creating a conditional model for every variable, given the remaining variables, and (ii) "gluing" the conditionally specified models together to form the joint distribution. As an example, consider three variables X, Y, Z. In Step (i), three separate regression models – say using backward variable selection – are built, viz. $p_1(x|y,z)$, $p_2(y|z)$, $p_3(z|x)$. Such a DN is depicted in Fig. 1. Unlike DAGs, DNs allow loops like $X \to Z \to Y \to X$. Important advantages of a DN over a DAG are its flexibility and the convenience in building one model at a time. For Step (ii),

* Corresponding author.
E-mail addresses: klkuo@nuk.edu.tw (K.-L. Kuo), yuwang@camden.rutgers.edu (Y.J. Wang).

https://doi.org/10.1016/j.jmva.2018.04.012
0047-259X/

Gibbs sampling (GS) has been suggested for approximating the joint distribution; see [7]. For the DN of Fig. 1 and an initial (x^0, y^0, z^0), GS draws y^1 from $p_2(y|z^0)$, followed by x^1 from $p_1(x|y^1, z^0)$, and finally z^1 from $p_3(z|x^1)$. The sequential order of sampling here is $Y \to X \to Z \to Y$. Is this the only correct sequence? And are the simulated (x^t, y^t, z^t) correct samples of the joint distribution? Based on Example 1 in Section 2, the answers may surprise some practitioners of GS.

Let $y = (y_1, \ldots, y_d)$ be a permutation of $x = (x_1, \ldots, x_d)$ whose joint probability density function (JPD), $f(x)$, is needed for inference. The recursive factorization in the sequence of y is $f(x) = \prod_{i=1}^{d} p_i(y_i|y_{>i})$, where $y_{>i} = \{y_{i+1}, \ldots, y_d\}$ and $y_{>d} = \emptyset$. This factorization is also called the hierarchical Bayesian model where only $y_k \to y_i$, $k > i$ is allowed. However, specifying hierarchically d such conditional probability densities (CPD) is not always practical. For example, missing value imputations are not hierarchical. Thus, DN attempts to define $f(x)$ from $\mathcal{M} = \{p_i(y_i|C_i) : i \in \{1, \ldots, d\}\}$ with $C_i \subseteq y_{-i} \equiv y \backslash \{y_i\}$. Hereafter, $\mathcal{M}$ is called a conditionally specified statistical model (CSSM). In the statistical literature, CPD $p_i(y_i|y_{-i})$ is called a full conditional, and $p_i(y_i|C_i)$ is called a collapsed conditional when C_i is a proper subset of y_{-i}. We would suggest global and local conditionals, respectively, as more suitable names.

After a $\mathcal{M} = \{p_i(y_i|C_i) : i \in \{1, \ldots, d\}\}$ is proposed to define a JPD, three mathematical issues need to be settled.

(i) Compatibility: Are the CPDs consistent with a JPD?
(ii) Sufficiency: Does $\mathcal{M}$ have sufficient information to define a d-dimensional JPD?
(iii) Implementation of Gibbs sampling: When $\mathcal{M}$ is both compatible and sufficient, can GS be used to simulate the JPD? If yes, then in what sequence?

The main contribution of this paper is to give a necessary and sufficient condition under which the iterative GS generates correct samples of JPD, and to propose an algorithm that identifies all the valid scan sequence. The proof of the main theorem is in Section 2, with illustrative examples. Applications of the theorems, reduced conditioning, and a compatibility check are discussed in Section 3.

2. Valid scan orders for implementing Gibbs sampling

Under compatibility and the positive condition of Besag [1], Gelman and Speed [5,6] establish the following sufficient condition.

Theorem 1. *A CSSM $\mathcal{M}$ determines a unique JPD for $x = (x_1, \ldots, x_d)$ if there exists a permutation $y = (y_1, \ldots, y_d)$ of x such that $\mathcal{M} = \{p_i(y_i|B_i, y_{>i}) : B_i \subseteq y_{<i}, i \in \{1, \ldots, d\}\}$.*

Theorem 1 requires that a sufficient $\mathcal{M}$ must contain at least one full conditional, $p_1(y_1|y_{>1})$. If every $B_i = \emptyset$, then $\mathcal{M}$ is a hierarchial Bayesian model, and when some B_is are nonempty, $\mathcal{M}$ is a DN with feedback loops, i.e., $y_j \to y_i$ for some $j < i$. The following example shows that the condition of Theorem 1 does not guarantee that GS can always be implemented.

Example 1. Consider the following JPD:

(x_1, x_2, x_3)	(0, 0, 0)	(1, 0, 0)	(0, 1, 0)	(1, 1, 0)	(0, 0, 1)	(1, 0, 1)	(0, 1, 1)	(1, 1, 1)
$f(x_1, x_2, x_3)$	1/20	3/20	4/20	2/20	3/20	3/20	3/20	1/20

Compute $\mathcal{M} = \{f(x_1|x_2, x_3), f(x_2|x_1, x_3), f(x_3)\}$ from the above $f(x_1, x_2, x_3)$. Let $y_i = x_i$ for $i \in \{1, 2, 3\}$. With $B_1 = \emptyset$, $B_2 = \{y_1\}$ and $B_3 = \emptyset$, $\mathcal{M}$ satisfies the sufficient condition of Theorem 1. The following table lists the six stationary JPDs along with their respective scan orders.

Scan Order	(0, 0, 0)	(1, 0, 0)	(0, 1, 0)	(1, 1, 0)	(0, 0, 1)	(1, 0, 1)	(0, 1, 1)	(1, 1, 1)
$y_1 \to y_2 \to y_3$	137/1360	207/1360	230/1360	106/1360	137/1360	207/1360	230/1360	106/1360
$y_1 \to y_3 \to y_2$	94/1700	228/1700	376/1700	152/1700	235/1700	285/1700	235/1700	95/1700
$y_2 \to y_1 \to y_3$	681/6800	1003/6800	1199/6800	517/6800	681/6800	1003/6800	1199/6800	517/6800
$y_2 \to y_3 \to y_1$	43/680	129/680	112/680	56/680	86/680	86/680	126/680	42/680
$y_3 \to y_1 \to y_2$	31/680	111/680	124/680	74/680	106/680	96/680	106/680	32/680
$y_3 \to y_2 \to y_1$	161/3400	483/3400	704/3400	352/3400	520/3400	520/3400	495/3400	165/3400

None of the stationary distributions is the original $f(x_1, x_2, x_3)$, therefore, GS cannot be used. □

The following theorem provides a necessary and sufficient condition for implementing GS. The additional requirement is that the B_i must be nested.

Theorem 2. *Assume a compatible CSSM $\mathcal{M} = \{f(x_i|B_i, x_{>i}) : B_i \subseteq x_{<i}, i \in \{1, \ldots, d\}\}$ is sufficient to define a unique JPD $f(x) = f(x_1, \ldots, x_d)$. Gibbs sampling in the order of $x_d \to x_{d-1} \to \cdots \to x_1$ can be used to generate correct samples of $f(x)$*