Ann Inst Stat Math (2019) 71:93–105
https://doi.org/10.1007/s10463-017-0625-x

Pseudo-Gibbs sampler for discrete conditional distributions

Kun-Lin Kuo[1] · **Yuchung J. Wang**[2]

Received: 26 October 2016 / Revised: 13 September 2017 / Published online: 24 October 2017

Abstract Conditionally specified models offers a higher level of flexibility than the joint approach. Regression switching in multiple imputation is a typical example. However, reasonable-seeming conditional models are generally not coherent with one another. Gibbs sampler based on incompatible conditionals is called pseudo-Gibbs sampler, whose properties are mostly unknown. This article investigates the richness and commonalities among their stationary distributions. We show that Gibbs sampler replaces the conditional distributions iteratively, but keep the marginal distributions invariant. In the process, it minimizes the Kullback–Leibler divergence. Next, we prove that systematic pseudo-Gibbs projections converge for every scan order, and the stationary distributions share marginal distributions in a circularly fashion. Therefore, regardless of compatibility, univariate consistency is guaranteed when the orders of imputation are circularly related. Moreover, a conditional model and its pseudo-Gibbs distributions have equal number of parameters. Study of pseudo-Gibbs sampler provides a fresh perspective for understanding the original Gibbs sampler.

Keywords Incompatibility · Iterative conditional replacement · Kullback–Leibler information divergence · Multiple imputation · Scan order · Stationary distribution

✉ Yuchung J. Wang
yuwang@camden.rutgers.edu

Kun-Lin Kuo
klkuo@nuk.edu.tw

[1] Institute of Statistics, National University of Kaohsiung, Kaohsiung 811, Taiwan

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

1 Introduction

For high-dimensional data, a reduced model may be formulated in two ways: a joint model in which all variables are considered simultaneously, or a system of univariate conditional models. The conditional approach allows more flexibility in modeling and expands the range of models. See Kuo and Wang (2011, p. 2457) for a brief survey of the statistical applications of conditionally specified models. In machine learning, Heckerman et al. (2000) propose dependence network which fits a classification or regression model for every variable using the remaining variables as the predictors: $\mathcal{F} = \{f_i(x_i|x_j, j \neq i)\}$. A dependence network then makes inference about the joint distribution from $\mathcal{F}$. However, the freedom to model each variable separately comes with a price: "in general, reasonable-seeming conditional models will not be compatible with any single joint distribution." (see Gelman and Raghunathan 2001, p. 268). That is, there does not exist a joint distribution that is coherent with every conditional.

Similarly in multiple imputation, the two mutually exclusive approaches are joint modeling (JM) and fully conditional specifications (FCS). According to van Buuren et al. (2006), advantages of FCS are (1) FCS is more flexible; (2) conditional models that are outside the limited parametric distributions are easier to specify; (3) it is easier to model certain bounds, patterns, and constraints conditionally; and (4) generalization to nonignorable missing data might be easier in FCS. But a major impairment of FCS is how to formulate a joint distribution such that all of the imputations are mutually consistent observations from one joint distribution. Both dependence network and FCS use the same computational method. Heckerman et al. (2000) use "the machinery of Gibbs sampling to define a joint distribution", while Rubin coined the acronym PIGS for potentially incompatible Gibbs sampler (Drechsler and Rässler 2008). Heckerman et al. (2000) call PIGS pseudo-Gibbs sampler (PGS). Hughes et al. (2014) noted that the order effects of PIGS are ubiquitous, but suggested that the real effects may be negligible. However, Chen et al. (2011, 2013) indicated that different scan orders can produce very different joint distributions. van Buuren et al. (1999) stated that "The subject of incompatible conditionals is, however, still an open research problem.", which remains true today. Other application of FCS for biomedical data can be found in Hughes et al. (2014).

This paper provides theoretical findings which can guide the practice of PGS. Using matrix representations, we show that Gibbs sampler is an algorithm that iteratively replaces the conditional distributions by the model conditionals until the marginal distributions are matched (Lemma 2). And every conditional replacement is an I^*-projection that minimizes the Kullback–Leibler divergence (Theorem 2). Next, we prove that systematic pseudo-Gibbs projections converge for every scan order (Theorem 3). Moreover, the stationary distributions share marginal distributions in a circularly fashion (Theorem 5). In short, PGS produces a large number of stationary distributions; some share conditionals and some share marginal distributions. We organize these distributions according to their commonalities, and show the degrees of freedom are balanced between a PGS and its stationary distributions (Theorem 6). Our characterizations help the practitioners of multiple imputation to answer some of their concerns (Drechsler and Rässler 2008): (1) does there exist a unique underlying joint

pdf? (PGS does converge but different scan order produces different joint distribution.), (2) what would happen to imputed values if such a joint pdf does not exist? (different degree of marginal consistency exists when the scan orders are circularly related; that is, different joint PGS distributions share the same marginal distributions.), and (3) can such conflicts be avoided and how? (yes, the conflicts, though intrinsic, can be alleviated by knowing the effect of imputation order on imputed values.). Moreover, study of PGS provides a fresh perspective for the compatible or nearly compatible FCS.

2 The two-dimensional pseudo-Gibbs Sampler

Let $x = (x_1, x_2)$ where x_1 and x_2 are discrete random variables with n_1 and n_2 values, respectively. Assume that the support Ω of x is $\{(i, j) : 1 \le i \le n_1, 1 \le j \le n_2\}$. Let $f_{1|2}$ and $f_{2|1}$ be the two conditional pdfs of a PGS. We assume that both conditional pdfs are strictly positive. Set $u = (u_1, u_2)$ and $v = (v_1, v_2)$, then the two $(n_1 n_2) \times (n_1 n_2)$ transition matrices based on $f_{1|2}$ and $f_{2|1}$ are

$$T_1 = [t_{uv}]_{u,v\in\Omega} \text{ where } t_{uv} = \begin{cases} f_{1|2}(v_1|v_2), & \text{if } u_2 = v_2, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$T_2 = [t_{uv}]_{u,v\in\Omega} \text{ where } t_{uv} = \begin{cases} f_{2|1}(v_2|v_1), & \text{if } u_1 = v_1, \\ 0, & \text{otherwise,} \end{cases}$$

respectively. Both T_1 and T_2 are idempotent.

For any pdf $g = (g(1, 1), g(2, 1), \ldots, g(n_1, n_2))$ with $g(i, j) > 0$, and let $g^* = gT_1$, then $g^*(i, j) = f_{1|2}(i|j)g_2(j)$, where g_2 is the x_2-marginal of g. We observe the following results: (1) g^* is a pdf; (2) $g^*_{1|2} = f_{1|2}$; and (3) $g^*_2 = g_2$. Because of $g(i, j) = g_{1|2}(i|j)g_2(j)$ and $g^*(i, j) = f_{1|2}(i|j)g_2(j)$, T_1 replaces $g_{1|2}$ with $f_{1|2}$ but keeps the original marginal pdf g_2. By the same token, T_2 replaces $g_{2|1}$ with $f_{2|1}$ but keeps the original marginal pdf g_1.

Consider the systematic scan Gibbs sampler and let $g^{[0]}$ be an initial pdf, the process of the Gibbs sampler sequentially produces pdfs: $g^{[2k+1]} = g^{[2k]}T_1$ and $g^{[2k+2]} = g^{[2k+1]}T_2$. Therefore, every $g^{[2k+1]}$ has $f_{1|2}$ and every $g^{[2k+2]}$ has $f_{2|1}$ as its conditional. Define $\mathcal{C}_1$ ($\mathcal{C}_2$) as the set of all joint pdfs having $f_{1|2}$ ($f_{2|1}$) as their $(x_1|x_2)$-conditional ($(x_2|x_1)$-conditional), thus, $g^{[2k+1]} \in \mathcal{C}_1$ and $g^{[2k+2]} \in \mathcal{C}_2$.

Let $I(g; \tau)$ be the Kullback–Leibler information divergence of τ at g. When $\tau \in \mathcal{C}_1$, we write $\tau(i, j) = f_{1|2}(i|j)\tau_2(j)$ and

$$\begin{aligned} I(g; \tau) &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} g(i, j) \log \frac{g(i, j)}{\tau(i, j)} \\ &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} g(i, j) \log \frac{g(i, j)}{f_{1|2}(i|j)\tau_2(j)} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} g(i,j)\log\frac{g(i,j)}{f_{1|2}(i|j)g_2(j)} + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} g(i,j)\log\frac{f_{1|2}(i|j)g_2(j)}{f_{1|2}(i|j)\tau_2(j)} \\
&= I(g; gT_1) + \sum_{j=1}^{n_2} g_2(j)\log\frac{g_2(j)}{\tau_2(j)} \\
&= I(g; gT_1) + I(g_2; \tau_2).
\end{aligned}
$$

Because

$$
\begin{aligned}
I(g_2;\tau_2) &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} f_{1|2}(i|j)g_2(j)\log\frac{g_2(j)}{\tau_2(j)} \\
&= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} f_{1|2}(i|j)g_2(j)\log\frac{f_{1|2}(i|j)g_2(j)}{f_{1|2}(i|j)\tau_2(j)},
\end{aligned}
$$

we have the following Pythagoras equality:

$$I(g;\tau) = I(g; gT_1) + I(gT_1;\tau).$$

Therefore,

$$\min_{\tau\in\mathcal{C}_1} I(g;\tau) = I(g; gT_1).$$

We call gT_1 the I^*-projection of g into $\mathcal{C}_1$, while the I-projection of Csiszár (1975) minimizes $I(\tau; g)$ over τ. Similarly, the I^*-projection of g into $\mathcal{C}_2$ is gT_2, because

$$\min_{\tau\in\mathcal{C}_2} I(g;\tau) = I(g; gT_2).$$

This projection is restated in the following proposition, and we use Fig. 1 to geometrically describe the convergence of PGS.

Proposition 1 *The $g^{[2k+1]}$ is the I^*-projection of $g^{[2k]}$ into $\mathcal{C}_1$, and $g^{[2k+2]}$ is the I^*-projection of $g^{[2k+1]}$ into $\mathcal{C}_2$. In addition, $g_2^{[2k]} = g_2^{[2k+1]}$ and $g_1^{[2k+2]} = g_1^{[2k+1]}$.*

For a two-variable systematic scan, the transition matrices are T_1T_2 and T_2T_1. Because $f_{1|2}(i|j) > 0$ and $f_{2|1}(j|i) > 0$, both T_1T_2 and T_2T_1 are transition matrices without absorbing state. Let the invariant distributions be denoted by $\pi^{(1,2)} = \pi^{(1,2)}T_1T_2$ and $\pi^{(2,1)} = \pi^{(2,1)}T_2T_1$, respectively. Hereafter, such invariant distributions are called *pseudo-Gibbs distributions* with their scan orders indicated. The following theorem shows that $\{g^{[2k+1]}\}$ and $\{g^{[2k+2]}\}$ converge to $\pi^{(2,1)}$ and $\pi^{(1,2)}$, respectively, in terms of divergences.

Theorem 1 *Both $I(\pi^{(2,1)}; g^{[2k+1]})$ and $I(\pi^{(1,2)}; g^{[2k]})$ decrease to 0 as $k \to \infty$.*

Proof Let $T_2T_1 = [t_{uv}]_{u,v\in\Omega}$, where t_{uv} represents the transition probability from u to v.

$$
\begin{aligned}
&I(\pi^{(2,1)}; g^{[2k-1]}) - I(\pi^{(2,1)}; g^{[2k+1]}) \\
&= \sum_u \pi^{(2,1)}(u) \log \frac{\pi^{(2,1)}(u)}{g^{[2k-1]}(u)} - \sum_v \pi^{(2,1)}(v) \log \frac{\pi^{(2,1)}(v)}{g^{[2k+1]}(v)} \\
&= \sum_u \sum_v \pi^{(2,1)}(u) t_{uv} \log \frac{\pi^{(2,1)}(u) t_{uv}}{g^{[2k-1]}(u) t_{uv}} \left(\text{since} \sum_v t_{uv} = 1 \text{for all } u \right) \\
&\quad - \sum_u \sum_v \pi^{(2,1)}(u) t_{uv} \log \frac{\pi^{(2,1)}(v)}{g^{[2k+1]}(v)} \left(\text{since} \sum_u \pi^{(2,1)}(u) t_{uv} = \pi^{(2,1)}(v) \right) \\
&= \sum_u \sum_v \pi^{(2,1)}(u) t_{uv} \log \frac{\pi^{(2,1)}(u) t_{uv}}{\frac{\pi^{(2,1)}(v) g^{[2k-1]}(u) t_{uv}}{g^{[2k+1]}(v)}}.
\end{aligned}
$$

Let $h_1(u, v) = \pi^{(2,1)}(u) t_{uv}$ and $h_2(u, v) = \frac{\pi^{(2,1)}(v) g^{[2k-1]}(u) t_{uv}}{g^{[2k+1]}(v)}$ where $(u, v) \in \Omega \times \Omega$. Because

$$
\sum_u \sum_v h_1(u, v) = \sum_u \sum_v \pi^{(2,1)}(u) t_{uv} = \sum_u \pi^{(2,1)}(u) = 1
$$

and

$$
\sum_u \sum_v h_2(u, v) = \sum_u \sum_v \frac{\pi^{(2,1)}(v) g^{[2k-1]}(u) t_{uv}}{g^{[2k+1]}(v)} = \sum_v \frac{\pi^{(2,1)}(v) g^{[2k+1]}(v)}{g^{[2k+1]}(v)} = 1.
$$

Therefore, both h_1 and h_2 are pdfs on $\Omega \times \Omega$. This implies that

$$
I(\pi^{(2,1)}; g^{[2k-1]}) - I(\pi^{(2,1)}; g^{[2k+1]}) = I(h_1; h_2),
$$

which is strictly positive until $g^{[2k-1]} = g^{[2k+1]}$ and the iterative I^*-projections have converged. Hence, $g^{[2k+1]}$ converges to $\pi^{(2,1)}$. By the same token, $g^{[2k+2]}$ converges to $\pi^{(1,2)}$. □

Because $\pi^{(2,1)}$ and $\pi^{(1,2)}$ are stationary distributions of T_2T_1 and T_1T_2, respectively, we have $\lim_{k\to\infty}(T_2T_1)^k = 1_{n_1n_2}\pi^{(2,1)}$ and $\lim_{k\to\infty}(T_1T_2)^k = 1_{n_1n_2}\pi^{(1,2)}$ where $1_{n_1n_2}$ is the $(n_1n_2) \times 1$ vector with all entries equal to 1.

The following proposition characterizes the relationship between $\pi^{(1,2)}$ and $\pi^{(2,1)}$.

Proposition 2 *Pseudo-Gibbs pdfs $\pi^{(2,1)}$ and $\pi^{(1,2)}$ satisfy* (1) $\pi^{(1,2)}T_1 = \pi^{(2,1)}$, $\pi^{(2,1)}T_2 = \pi^{(1,2)}$, $\pi^{(1,2)}T_2 = \pi^{(1,2)}$ *and* $\pi^{(2,1)}T_1 = \pi^{(2,1)}$; (2) $\pi^{(2,1)} \in \mathcal{C}_1$ *and* $\pi^{(1,2)} \in \mathcal{C}_2$; *and* (3) $\pi^{(2,1)}$ *and* $\pi^{(1,2)}$ *have the same x_1-marginal pdf and the same x_2-marginal pdf.*

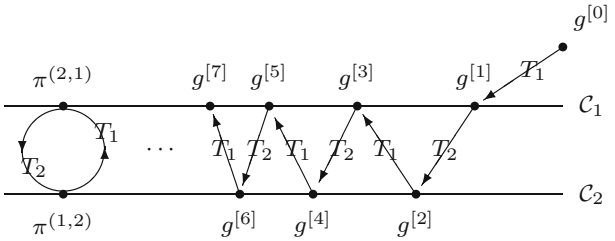

Fig. 1 The iterative I^*-projections and the two stationary PGS distributions

Proof 1. Because $\pi^{(1,2)}T_1 - \pi^{(2,1)} = (\pi^{(1,2)}T_1T_2)T_1 - \pi^{(2,1)}T_2T_1 = (\pi^{(1,2)}T_1 - \pi^{(2,1)})T_2T_1$, we must have either $\pi^{(1,2)}T_1 - \pi^{(2,1)} = \pi^{(2,1)}$ or $\pi^{(1,2)}T_1 - \pi^{(2,1)} = 0$. But $\pi^{(1,2)}T_1 = 2\pi^{(2,1)}$ is impossible because $2\pi^{(2,1)}$ is not a pdf. Thus, we have $\pi^{(1,2)}T_1 = \pi^{(2,1)}$. In addition, because T_2 is idempotent, we have $\pi^{(1,2)}T_2 = (\pi^{(1,2)}T_1T_2)T_2 = \pi^{(1,2)}T_1T_2 = \pi^{(1,2)}$.

2. Because $\pi^{(2,1)} = \pi^{(2,1)}T_1$, we have $\pi^{(2,1)} \in \mathcal{C}_1$.
3. Because $\pi^{(2,1)}T_2 = \pi^{(1,2)}$, $\pi^{(1,2)}$ is the I^*-projection of $\pi^{(2,1)}$ into $\mathcal{C}_2$, therefore they have the same x_1-marginal pdf. Similarly, $\pi^{(1,2)}T_1 = \pi^{(2,1)}$ implies that $\pi^{(1,2)}$ and $\pi^{(2,1)}$ also have the same x_2-marginal pdf. □

Figure 1 illustrates the iterative I^*-projections, and the iterations stop when the two stationary distributions have the same marginal pdfs. When $f_{1|2}$ and $f_{2|1}$ are compatible, $\mathcal{C}_1$ and $\mathcal{C}_2$ will intersect at $\pi^{(1,2)} = \pi^{(2,1)}$. Conditional model has $2n_1n_2 - n_1 - n_2$ parameters, while $\pi^{(1,2)}$ and $\pi^{(2,1)}$ collectively have $2n_1n_2 - 2$ parameters. Due to shared marginal pdfs that number is reduced by $n_1 + n_2 - 2$, and the two sides are balanced.

3 The d-dimensional pseudo-Gibbs distributions

Without loss of generality, we assume that each constituent x_i of $x = (x_1, \ldots, x_d)$ is univariate. Let the supports of x_i and x be, respectively, $\Omega_i = \{1, \ldots, n_i\}$ and $\Omega = \Omega_1 \times \cdots \times \Omega_d$. For the rest of the paper, we assume that $x_1, \ldots, x_d$ are variation-independent so that the joint pdf $g(x) > 0$ for all $x \in \Omega$. For $a \subseteq \{1, \ldots, d\}$, we use g_a to represent the $(x_i : i \in a)$-marginal pdf of g. Let the symbol "$-i$" mean $\{1, \ldots, i-1, i+1, \ldots, d\}$, hence $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$ has marginal pdf g_{-i}, and the full conditional of x_i is $g_{i|-i} = g/g_{-i}$.

We assume that the d strictly positive full conditionals, $f_{j|-j}$, $1 \le j \le d$, have been calculated. Let I_n be the identity matrix of size n, let 1_n be the $n \times 1$ vector of 1's, and let $\delta_{n,j}$ be the $n \times n$ matrix whose (j, j)th entry is 1, and the rest are all 0's. The transition matrix T_j, based on $f_{j|-j}$, can be expressed as

$$T_j = \sum_{x_{-j} \in \Omega_{-j}} \delta_{n_d, x_d} \otimes \cdots \otimes \delta_{n_{j+1}, x_{j+1}} \otimes [1_{n_j} f_{j|-j}(\cdot|x_{-j})] \otimes \delta_{n_{j-1}, x_{j-1}} \otimes \cdots \otimes \delta_{n_1, x_1}$$

where $f_{j|-j}(\cdot|x_{-j}) \equiv (f_{j|-j}(1|x_{-j}), \ldots, f_{j|-j}(n_j|x_{-j}))$ and $\otimes$ is the Kronecker product. Also, define matrix $M_a = Q_d \otimes Q_{d-1} \otimes \cdots \otimes Q_1$, where $Q_j = I_{n_j}$ if $j \in a$, and $Q_j = 1_{n_j}$ if $j \notin a$, and it can be shown that $g_a = gM_a$. Sometimes it is easier to use the a-complement as the subscript, such as $g_{-a} = gM_{-a}$. For $d = 4$, $g_{124} = gM_{-3} = gM_{124}$ and $g_{13} = gM_{13} = gM_{-24}$. Using the above matrix representations, we can prove the following lemma.

Lemma 1 *For* $1 \le j \le d$, *we have* (1) $T_j T_j = T_j$; *and* (2) *for* $j \in a$, $T_j M_{-a} = M_{-a}$, *thus,* $T_j M_{-j} = M_{-j}$.

Let $\mathcal{D}$ be the space of joint pdfs of x, and let $\mathcal{C}_j$ be the set of all the joint pdfs having $f_{j|-j}$ as their $(x_j|x_{-j})$-conditional. The following Lemma 2 shows that T_j replaces the $(x_j|x_{-j})$-conditional of g by $f_{j|-j}$, but keeps its original g_{-j} unaltered.

Lemma 2 *For* $g \in \mathcal{D}$, *we have* $gT_j \in \mathcal{C}_j$ *and the the* x_{-j}*-marginal pdf of* gT_j *and* g *are the same.*

Proof Consider the x_{-j}-marginal pdf of gT_j,

$$(gT_j)_{-j} = gT_jM_{-j} = g(T_jM_{-j}) = gM_{-j} = g_{-j}.$$

Moreover, for every $x \in \Omega$,

$$gT_j(x) = \sum_{\substack{y \in \Omega \\ y_{-j} = x_{-j}}} g(y) f_{j|-j}(x_j|x_{-j}) = g_{-j}(x_{-j}) f_{j|-j}(x_j|x_{-j}).$$

□

Lemma 2 implies that $T_j(\mathcal{D}) \equiv \{gT_j : g \in \mathcal{D}\} \subset \mathcal{C}_j$ which in conjunction with T_j being an idempotent implies that $T_j(\mathcal{D}) = \mathcal{C}_j$. Mapping T_j is clearly not one-to-one because $g^1 T_j = g^2 T_j$ whenever $g^1_{-j} = g^2_{-j}$.

Let $(i_1, \ldots, i_d)$ be any permutation of $(1, \ldots, d)$. For a systematic Gibbs sampler with scan order $(i_1, \ldots, i_d)$ (i.e., the variables are repeatedly updated in the order $x_{i_1} \to \cdots \to x_{i_d} \to x_{i_1} \to \cdots$) and initial pdf $g^{[0]}$, the Gibbs sampler sequentially produces pdfs: $g^{[1]} = g^{[0]}T_{i_1}, g^{[2]} = g^{[1]}T_{i_2}, \ldots, g^{[d]} = g^{[d-1]}T_{i_d}, g^{[d+1]} = g^{[d]}T_{i_1}, \ldots$, and $g^{[dk+j]} = g^{[dk+j-1]}T_{i_j}$ for $k = 0, 1, 2, \ldots$ and $1 \le j \le d$.

The Kullback–Leibler information divergence of $\tau = f_{j|-j}\tau_{-j} \in \mathcal{C}_j$ at $g \in \mathcal{D}$ is

$$I(g; \tau) = I(g; gT_j) + I(gT_j; \tau).$$

Therefore,

$$\min_{\tau \in \mathcal{C}_j} I(g; \tau) = I(g; gT_j).$$

Thus, we have the following theorem.

Theorem 2 *For a scan order* $(i_1, \ldots, i_d)$, $g^{[dk+j]}$ *is the* I^**-projection of* $g^{[dk+j-1]}$ *into* $\mathcal{C}_{i_j}$ *for every nonnegative integer* k *and* $1 \le j \le d$.

For scan order $(i_1, \ldots, i_d)$, its transition matrix is $T_{i_1} \cdots T_{i_d}$, and its pseudo-Gibbs distribution is denoted by $\pi^{(i_1,\ldots,i_d)} = \pi^{(i_1,\ldots,i_d)} T_{i_1} \cdots T_{i_d}$. By Lemma 2, $\pi^{(i_1,\ldots,i_d)} \in \mathcal{C}_{i_d}$. When there exists a $g \in \cap_{j=1}^d \mathcal{C}_j$, we have $gT_j = g$ for all j, which implies $gT_{i_1} \cdots T_{i_d} = g$ for all $d!$ scan orders. Hence, the model conditionals are compatible if and only if $\cap_{j=1}^d \mathcal{C}_j \neq \emptyset$.

Let $\mathcal{S}$ denote the symmetric group of all possible permutations of $(1, \ldots, d)$. Circular operator $\sigma : \mathcal{S} \to \mathcal{S}$ is defined as $\sigma(i_1, \ldots, i_d) = (i_2, \ldots, i_d, i_1)$. Applying the circular operator j times, $\sigma^j(i_1, \ldots, i_d) = (i_{j+1}, \ldots, i_d, i_1, \ldots, i_j)$, and $\sigma^0 = \sigma^d$ is the identity operator. Two scan orders u and v are said to be *circularly related* if $v = \sigma^j(u)$ for some j, and $\mathcal{S}$ can be partitioned into circularly related equivalence classes.

Theorem 3 *For a scan order* $u = (i_1, \ldots, i_d)$ *and* $1 \le j \le d$, $I(\pi^{\sigma^j(u)}; g^{[dk+j]}) > I(\pi^{\sigma^j(u)}; g^{[d(k+1)+j]})$, *unless* $g^{[dk+j]} = g^{[d(k+1)+j]}$. *Therefore,* $g^{[dk+j]}$ *converges to* $\pi^{\sigma^j(u)}$ *as* k *approaches* ∞.

The proof of Theorem 3 is similar to that of Theorem 1 so we omit it.

Let $\mathcal{G} = \{\pi^{(i_1,\ldots,i_d)} : (i_1, \ldots, i_d) \in \mathcal{S}\}$ be the collection of $d!$ pseudo-Gibbs distributions, which may be partitioned by two different ways. First, $\mathcal{G}$ is partitioned into $\cup_{j=1}^d (\mathcal{C}_j \cap \mathcal{G})$, where the $(d-1)!$ distributions of $\mathcal{C}_j \cap \mathcal{G}$ share the same $(x_j|x_{-j})$-conditional $f_{j|-j}$. Second, $\mathcal{G}$ is partitioned into $(d-1)!$ equivalent classes of $\mathcal{G}^u = \{\pi^{\sigma^j(u)} : 0 \le j \le d-1\}$, collection of d circularly related pseudo-Gibbs distributions.

Theorem 4 *In a* $\mathcal{G}^u$ *where* $u = (i_1, \ldots, i_d)$, *we have* (1) $\pi^{\sigma^j(u)} T_{i_{j+1}} = \pi^{\sigma^{j+1}(u)}$, *and* $\pi^{\sigma^{j+1}(u)}$ *is the* I^**-projection of* $\pi^{\sigma^j(u)}$ *into* $\mathcal{C}_{i_{j+1}}$; *and* (2) $\pi^{\sigma^j(u)} T_{i_j} = \pi^{\sigma^j(u)}$.

Proof 1. Because $\pi^{\sigma^j(u)} T_{i_{j+1}} - \pi^{\sigma^{j+1}(u)} = (\pi^{\sigma^j(u)} T_{i_{j+1}} - \pi^{\sigma^{j+1}(u)}) T_{i_{j+2}} \cdots T_{i_d} T_{i_1} \cdots T_{i_{j+1}}$, we have $\pi^{\sigma^j(u)} T_{i_{j+1}} - \pi^{\sigma^{j+1}(u)} = 0$ or $\pi^{\sigma^j(u)} T_{i_{j+1}} - \pi^{\sigma^{j+1}(u)} = \pi^{\sigma^{j+1}(u)}$. The latter is impossible because $\pi^{\sigma^j(u)} T_{i_{j+1}}$ is a pdf.

2. Because T_{i_j} is idempotent, we have

$$\pi^{\sigma^j(u)} T_{i_j} = (\pi^{\sigma^j(u)} T_{i_{j+1}} \cdots T_{i_d} T_{i_1} \cdots T_{i_j}) T_{i_j} = \pi^{\sigma^j(u)}.$$

□

The following corollary shows the cyclic nature of $\mathcal{G}^u$, thus, its pseudo-Gibbs distributions form a cycle.

Corollary 1 *For a scan order* $u = (i_1, \ldots, i_d)$ *and* $1 \le j \ne k \le d$, *we have*

$$\pi^{\sigma^k(u)} = \begin{cases} \pi^{\sigma^j(u)} T_{i_{j+1}} \cdots T_{i_k}, & \text{if } k > j, \\ \pi^{\sigma^j(u)} T_{i_{j+1}} \cdots T_{i_d} T_{i_1} \cdots T_{i_k}, & \text{if } k < j. \end{cases}$$

Next, we identify shared marginal pdfs between $\pi^{\sigma^j(u)}$ and $\pi^{\sigma^k(u)}$.

Theorem 5 *For a scan order $u = (i_1, \ldots, i_d)$ and $1 \le j \ne k \le d$, set $a = \{i_{j+1}, \ldots, i_k\}$ when $j < k$ or $a = \{i_{k+1}, \ldots, i_j\}$ when $j > k$. Among circularly related pseudo-Gibbs pdfs, both $\pi_a^{\sigma^j(u)} = \pi_a^{\sigma^k(u)}$, and $\pi_{-a}^{\sigma^j(u)} = \pi_{-a}^{\sigma^k(u)}$ hold.*

Proof Suppose that $j < k$, by Corollary 1 and (2) of Lemma 1, we have

$$\pi^{\sigma^k(u)} M_{-a} = \pi^{\sigma^j(u)} T_{i_{j+1}} \cdots T_{i_k} M_{-a} = \pi^{\sigma^j(u)} M_{-a},$$

and

$$\pi^{\sigma^j(u)} M_a = \pi^{\sigma^j(u)} T_{i_{j+1}} \cdots T_{i_d} T_{i_1} \cdots T_{i_j} M_a = \pi^{\sigma^j(u)} T_{i_{j+1}} \cdots T_{i_k} M_a = \pi^{\sigma^k(u)} M_a.$$

Suppose that $j > k$, we have

$$\pi^{\sigma^k(u)} M_a = \pi^{\sigma^j(u)} T_{i_{j+1}} \cdots T_{i_d} T_{i_1} \cdots T_{i_k} M_a = \pi^{\sigma^j(u)} M_a,$$

and

$$\begin{aligned} \pi^{\sigma^j(u)} M_{-a} &= \pi^{\sigma^j(u)} T_{i_{j+1}} \cdots T_{i_d} T_{i_1} \cdots T_{i_j} M_{-a} \\ &= \pi^{\sigma^j(u)} T_{i_{j+1}} \cdots T_{i_d} T_{i_1} \cdots T_{i_k} M_{-a} = \pi^{\sigma^k(u)} M_{-a}. \end{aligned}$$

□

Example 1 Suppose that $d = 5$ and $u = (1, 3, 5, 4, 2)$, consider $\pi^{\sigma^2(u)} = \pi^{(5,4,2,1,3)}$ and $\pi^{\sigma^4(u)} = \pi^{(2,1,3,5,4)}$. Let $a = \{5, 4\}$ then $\pi_{45}^{(5,4,2,1,3)} = \pi_{45}^{(2,1,3,5,4)}$ and $\pi_{123}^{(5,4,2,1,3)} = \pi_{123}^{(2,1,3,5,4)}$. In general, when two circularly related pseudo-Gibbs pdfs can be written as $\pi^{(a,-a)}$ and $\pi^{(-a,a)}$, they share the same x_a-marginal and x_{-a}-marginal distributions.

The immediate neighbors of π^u on the circle are π^v and π^w such that $\sigma(v) = u$ and $\sigma(u) = w$, and three of them share a $(d-2)$-dimensional marginal pdf. By induction, members of $\mathcal{G}^u$ share the same one-dimensional marginal distributions, though they belong to $\mathcal{C}_{i_d}, \mathcal{C}_{i_1}, \ldots,$ and $\mathcal{C}_{i_{d-1}}$, respectively. A natural application to multiple imputation is that one-dimensional consistency is guaranteed when the orders of imputation are circularly related.

Proposition 3 *In a $\mathcal{G}^u$, all members have the same one-dimensional marginal distributions. That is, $\pi_j^u = \pi_j^{\sigma(u)} = \cdots = \pi_j^{\sigma^{d-1}(u)}$, $1 \le j \le d$.*

Example 2 For $d = 4$, $\mathcal{G}$ has $4! = 24$ distinguishable pseudo-Gibbs distributions. First, $\mathcal{G}$ is partitioned as $\cup_{j=1}^4 (\mathcal{C}_j \cap \mathcal{G})$ where

$$\mathcal{C}_1 \cap \mathcal{G} = \left\{ \pi^{(2,3,4,1)}, \pi^{(2,4,3,1)}, \pi^{(3,2,4,1)}, \pi^{(3,4,2,1)}, \pi^{(4,2,3,1)}, \pi^{(4,3,2,1)} \right\},$$

$$\mathcal{C}_2 \cap \mathcal{G} = \left\{\pi^{(1,3,4,2)}, \pi^{(1,4,3,2)}, \pi^{(3,1,4,2)}, \pi^{(3,4,1,2)}, \pi^{(4,1,3,2)}, \pi^{(4,3,1,2)}\right\},$$
$$\mathcal{C}_3 \cap \mathcal{G} = \left\{\pi^{(1,2,4,3)}, \pi^{(1,4,2,3)}, \pi^{(2,1,4,3)}, \pi^{(2,4,1,3)}, \pi^{(4,1,2,3)}, \pi^{(4,2,1,3)}\right\},$$
$$\mathcal{C}_4 \cap \mathcal{G} = \left\{\pi^{(1,2,3,4)}, \pi^{(1,3,2,4)}, \pi^{(2,1,3,4)}, \pi^{(2,3,1,4)}, \pi^{(3,1,2,4)}, \pi^{(3,2,1,4)}\right\}.$$

All pdfs in $\mathcal{C}_j \cap \mathcal{G}$ have the same $(x_j|x_{-j})$-conditional. The second partition divides $\mathcal{G}$ into six circularly related classes with shared marginals:

$$\mathcal{G}^{(1,2,3,4)} = \left\{\pi^{(1,2,3,4)}, \pi^{(2,3,4,1)}, \pi^{(3,4,1,2)}, \pi^{(4,1,2,3)}\right\},$$
$$\mathcal{G}^{(1,2,4,3)} = \left\{\pi^{(1,2,4,3)}, \pi^{(2,4,3,1)}, \pi^{(4,3,1,2)}, \pi^{(3,1,2,4)}\right\},$$
$$\mathcal{G}^{(1,3,2,4)} = \left\{\pi^{(1,3,2,4)}, \pi^{(3,2,4,1)}, \pi^{(2,4,1,3)}, \pi^{(4,1,3,2)}\right\},$$
$$\mathcal{G}^{(1,3,4,2)} = \left\{\pi^{(1,3,4,2)}, \pi^{(3,4,2,1)}, \pi^{(4,2,1,3)}, \pi^{(2,1,3,4)}\right\},$$
$$\mathcal{G}^{(1,4,2,3)} = \left\{\pi^{(1,4,2,3)}, \pi^{(4,2,3,1)}, \pi^{(2,3,1,4)}, \pi^{(3,1,4,2)}\right\},$$
$$\mathcal{G}^{(1,4,3,2)} = \left\{\pi^{(1,4,3,2)}, \pi^{(4,3,2,1)}, \pi^{(3,2,1,4)}, \pi^{(2,1,4,3)}\right\}.$$

Figure 2 illustrates the two partitions of $\mathcal{G}$.

A conditional model is compatible when every pair of conditionals are compatible, otherwise it is incompatible. When every pair of conditionals are incompatible, the model is said to be *maximally incompatible*. For a maximally incompatible model $\mathcal{P}$, the number of free parameters of all PGS distributions is equal to the number of free parameters of $\mathcal{P}$. The following theorem implies that shared conditionals within $\mathcal{C}_j \cap \mathcal{G}$ plus share marginal within $\mathcal{G}^u$ exhaust all possible commonalities.

Theorem 6 *Both $\mathcal{P}$ and $\mathcal{G}$ have the same number of free parameters when $\mathcal{P}$ is maximally incompatible.*

Proof For $1 \le j \le d$, the support of x_j is $\Omega_j = \{1, \ldots, n_j\}$; therefore, the total number of degrees of freedom of $\mathcal{P}$ is

$$\sum_{j=1}^{d} (n_j - 1) \frac{\prod_{i=1}^{d} n_i}{n_j}.$$

Without commonality, the total number of degrees of freedom of $\mathcal{G}$ would be $d!(\prod_{j=1}^{d} n_j - 1)$. Within each $\mathcal{C}_j \cap \mathcal{G}$, there is a $[(d-1)! - 1](n_j - 1)\prod_{i=1}^{d} n_i/n_j$ reduction of parameters due to shared conditionals. Within every $\mathcal{G}^u$, there is a $\sum_{j=1}^{d}(\prod_{i=1}^{d} n_i/n_j - 1)$ reduction parameters due to shared marginals. The total number of degrees of freedom of $\mathcal{G}$ is computed by:

$$d!\left(\prod_{j=1}^{d} n_j - 1\right) - \sum_{j=1}^{d} [(d-1)! - 1](n_j - 1)\frac{\prod_{i=1}^{d} n_i}{n_j}$$

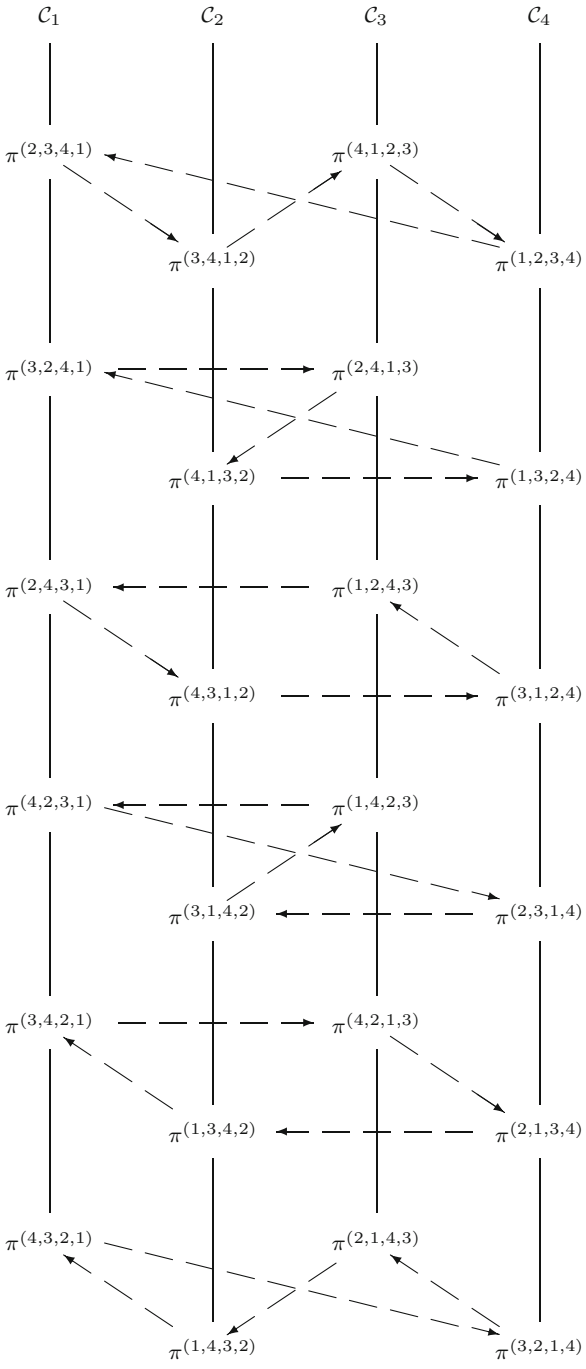

Fig. 2 Commonalities among pseudo-Gibbs distributions for $d = 4$. The vertical lines represent the spaces with common conditional distribution, while the circles of the dotted line depict the six equivalence classes where marginal distributions are shared in a circular fashion

$$-(d-1)!\sum_{j=1}^{d}\left(\frac{\prod_{i=1}^{d} n_i}{n_j}-1\right)=\sum_{j=1}^{d}(n_j-1)\frac{\prod_{i=1}^{d} n_i}{n_j}.$$

□

4 Conclusions

Incompatibility broadens not only the applications of Gibbs sampling but also the understanding of the algorithm itself. New insights are gained by viewing the compatible Gibbs sampler as a special case of pseudo-Gibbs samplers. The existence of $d!$ distinct pseudo-Gibbs distributions is under the assumption that the conditional model is maximally incompatible. For incompatible model which is not maximally incompatible, the total degree of freedom of pseudo-Gibbs distributions will be reduced to maintain balance of degrees of freedom. Accounting the additional commonality among pseudo-Gibbs distributions can be dealt with on a case-by-case base. However, balance of parameters between $\mathcal{P}$ and $\mathcal{G}$ should always be preserved, and such a balance will help us to determine the number of distinct pseudo-Gibbs distributions.

By showing that the effect of Gibbs sampling is an I^*-projection, and proving that successive projections reduce the information divergence, the convergence of pseudo-Gibbs sampler is thus guaranteed. Our approach mimics Darroch and Ratcliff's (1972) proof of the convergence of the iterative proportional fitting algorithm. They also proved that convergence in divergence implies convergence in L^2 norm for probability vectors. It is interesting to note the complimentary nature of the two algorithms. The Gibbs sampling replaces the conditional but leaves the marginal unaltered, while the iterative proportional fitting algorithm replaces the marginal but leaves the conditional invariant. Regardless of compatibility, Gibbs sampling is an algorithm that replaces its conditional distributions iteratively until convergence. Coincidentally, the convergence of both algorithms is proved via the information divergence.

This article provides a theoretical study about the richness and commonalities among pseudo-Gibbs distributions. Though there is a one-to-one correspondence between the scan orders and the pseudo-Gibbs distributions, the optimality in terms of divergence imposes maximal commonality among the stationary distributions. Pseudo-Gibbs sampler translates the incompatibility among conditional models into an ordered diversity. By organizing the ensemble of joint distributions in a systematical fashion, their commonality becomes easily understood. The knowledge about the effect of imputation order on imputed values can alleviate some of the conflicts caused by PIGS. For regression switching, the d imputations from d circularly related scan orders not only incorporate the dependence of every conditional but also achieve univariate marginal consistency. In addition, the d imputations provide good information to estimate the standard error.

Acknowledgements This work was supported in part by the Ministry of Science and Technology, Taiwan (MOST 104-2118-M-390-001 and MOST 105-2118-M-390-002). The authors thank two referees and one Associated Editor for their comments.

References

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

Chen, S.-H., Ip, E. H., Wang, Y. J. (2013). Gibbs ensembles for incompatible dependence networks. *WIREs Computational Statistics*, *5*, 478–485.

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, *3*, 146–158.

Darroch, J. N., Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, *43*, 1470–1480.

Drechsler, J., Rässler, S. (2008). Does convergence really matter? In Shalabh, C. Heumann (Eds.), *Recent advances in linear models and related areas* (pp. 341–355). Heidelberg: Physica-Verlag.

Gelman, A., Raghunathan, T. E. (2001). Comment on "Conditionally specified distributions" by B.C. Arnold, E. Castillo and J.M. Sarabia. *Statistical Science*, *16*, 268–269.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., Kadie, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, *1*, 49–75.

Hughes, R. A., White, I. R., Seaman, S. R., Cappenter, J. R., Tilling, K., Sterne, J. A. C. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, *14*, 28.

Kuo, K.-L., Wang, Y. J. (2011). A simple algorithm for checking compatibility among discrete conditional distributions. *Computational Statistics and Data Analysis*, *55*, 2457–2462.

van Buuren, S., Boshuizen, H. C., Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–94.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064.