HHS Public Access

Author manuscript

J Stat Comput Simul. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:

J Stat Comput Simul. 2015 ; 85(16): 3266–3275. doi:10.1080/00949655.2014.968159.

Behavior of the Gibbs Sampler When Conditional Distributions Are Potentially Incompatible

Shyh-Huei Chen[a,*] and **Edward H. Ip**[a,b]

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

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

Abstract

The Gibbs sampler has been used extensively in the statistics literature. It relies on iteratively sampling from a set of compatible conditional distributions and the sampler is known to converge to a unique invariant joint distribution. However, the Gibbs sampler behaves rather differently when the conditional distributions are not compatible. Such applications have seen increasing use in areas such as multiple imputation. In this paper, we demonstrate that what a Gibbs sampler converges to is a function of the order of the sampling scheme. Besides providing the mathematical background of this behavior, we also explain how that happens through a thorough analysis of the examples.

Keywords

Gibbs chain; Gibbs sampler; Potentially incompatible conditional-specified distribution

1. INTRDUCTION

The Gibbs sampler is, if not singularly, one of the most prominent Markov chain Monte Carlo (MCMC)-based methods. Partly because of its conceptual simplicity and elegance in implementation, the Gibbs sampler has been increasingly used across a very broad range of subject areas including bioinformatics and spatial analysis. While its root dates back to earlier work (e.g., [1]), the popularity of the Gibbs sampling is commonly credited to Geman and Geman [2], in which the algorithm was used as a tool for image processing. Its use in statistics, especially Bayesian analysis, has since grown very rapidly [3–5]. For a quick introduction of the algorithm, see Casella and George [6].

One of the recent developments of the Gibbs sampler is in its application to potentially incompatible conditional-specified distributions (PICSD) [7, 8]. When statistical models involve high-dimensional data, it is often easier to specify conditional distributions instead of the entire joint distributions. However, the approach of specifying a conditional

*Corresponding address: Medical Center Boulevard, Winston-Salem, NC 27157, USA. schen@wakehealth.edu. Phone: +1-336-713-1554. Fax:+1-336-716-6427.

distribution has the risk of not forming a compatible joint model. Consider a system of d discrete random variables $\boldsymbol{X} = \{X_1, X_2, \cdots, X_d\}$, whose fully conditional model is specified by $\boldsymbol{F} = \{f_1, f_2, \cdots, f_d\}$, where $f_k \equiv f_{x_k|x_k^c} = f(x_k|x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_d)$. If the conditional models are individually specified, then there may not exist a joint distribution that will give rise to the specified set of conditional distributions. In such a case, we call $\boldsymbol{F}$ incompatible.

The study of PICSD is closely related to the Gibbs sampler because the latter relies on iteratively drawing samples from $\boldsymbol{F}$ to form a Markov chain. Under mild conditions, the Markov chain converges to the desired joint distribution, if $\boldsymbol{F}$ is compatible. However, if $\boldsymbol{F}$ is not compatible, then the Gibbs sampler could exhibit erratic behavior [9].

In this paper, our goal is to demonstrate the behavior of the Gibbs sampler (or the pseudo Gibbs sampler as it is not a true Gibbs sampler in the traditional sense of presumed compatible conditional distributions) for PICSD. By using several simple examples, we show mathematically that what a Gibbs sampler converges to is a function of the order of the sampling scheme in the Gibbs sampler. Furthermore, we show that if we follow a random order in sampling conditional distributions at each iteration—i.e., using a random-scan Gibbs sampler [10]—then the Gibbs sampling will lead to a mixture of the joint distributions formed by each combination of fixed-order (or more formally, fixed-scan) when $d = 2$ but the result is not true when $d > 2$. This result is a refinement of a conjecture put forward in Liu [11].

Two recent developments in the statistical and machine-learning literature underscore the importance of the current work. The first is in the application of the Gibbs sampler to a dependency network, which is a type of generalized graphical model specified by conditional probability distributions [7]. One approach to learning a dependency network is to first specify individual conditional models and then apply a (pseudo) Gibbs sampler to estimate the joint model. Heckerman et al. [7] acknowledged the possibility of incompatible conditional models but argued that when the sample size is large, the degree of incompatibility will not be substantial and the Gibbs sampler is still applicable. Yet another example is the use of the fully conditional specification for multiple imputation of missing data [12, 13]. The method, which is also called multiple imputation by chained equations (MICE), makes use of a Gibbs sampler or other MCMC-based methods that operate on a set of conditionally specified models. For each variable with a missing value, an imputed value is created under an individual conditional-regression model. This kind of procedure was viewed as combining the best features of many currently available multiple imputation approaches [14]. Due to its flexibility over compatible multivariate-imputation models [15] and ability to handle different variable types (continuous, binary, and categorical) the MICE has gained acceptance for its practical treatment of missing data, especially in high-dimensional data sets [16]. Popular as it is, the MICE has the limitation of potentially encountering incompatible conditional-regression models and it has been shown that an incompatible imputation model can lead to biased estimates from imputed data [17]. So far, very little theory has been developed in supporting the use of MICE [18]. A better

understanding of the theoretical properties of applying the Gibbs sampler to PICSD could lead to important refinements of these imputation methods in practice.

The article is organized as follows: First, we provide basic background to the Gibbs chain and Gibbs sampler and define the scan order of a Gibbs sampler. In Section 3, we offer several analytic results concerning the stationary distributions of the Gibbs sampler under different scan patterns and a counter-example to a surmise about the Gibbs sampler under a random order of scan pattern. Section 4 describes two simple examples to numerically demonstrate the convergence behavior of a Gibbs sampler as a function of scan order, both by applying matrix algebra to the transition kernel as well as using MCMC-based computation. Finally in Section 5 we provide a brief discussion.

2. GIBBS CHAIN AND GIBBS SAMPLER

Continuing the notation in the previous section, let $\boldsymbol{a} = (a_1, a_2, \cdots, a_d)$ denote a permutation of $\{1, 2, \cdots, d\}$, $\boldsymbol{x} = \{x_1, x_2, \cdots, x_d\}$ denote a realization of $\boldsymbol{X}$ with $x_k \in \{1, 2, \cdots, C_k\}$, where C_k is the number of categories of the k^{th} variable. Thus, $\boldsymbol{x}_{\boldsymbol{a}} \equiv (x_{a1}, x_{a2}, \cdots, x_{ad})$ is a realization of X defined in the order of $\boldsymbol{a}$. We also denote $\boldsymbol{x}^c_{a_k} = (x_{a_1}, \cdots, x_{a_{k-1}}, x_{a_{k+1}}, \cdots, x_{a_d})$, the relative complement of x_{ak} with respect to $\boldsymbol{x}_{\boldsymbol{a}}$.

For a specified $\boldsymbol{F}$, the associated fixed (systemic)-scan Gibbs chain governed by a scan pattern $\boldsymbol{a}$ can be implemented as follows:

1. Pick an arbitrary starting vector $\boldsymbol{x}^{(0)}_{\boldsymbol{a}} = (x^{(0)}_{a_1}, x^{(0)}_{a_2}, \cdots, x^{(0)}_{a_d})$.
2. On the t^{th} cycle, successively draw from the full conditional distributions according to scan pattern $\boldsymbol{a} = (a_1, a_2, \cdots, a_d)$ as follows:

$$X^{(td+1)}_{a_1} \sim f_{a_1}\left(x_{a_1} \middle| (\boldsymbol{x}^{(td)}_{a_1})^c\right)$$

$$X^{(td+2)}_{a_2} \sim f_{a_2}\left(x_{a_2} \middle| (\boldsymbol{x}^{(td+1)}_{a_2})^c\right)$$

$$\vdots$$

$$X^{(td+d-1)}_{a_{d-1}} \sim f_{a_{d-1}}\left(x_{a_{d-1}} \middle| (\boldsymbol{x}^{(td+d-2)}_{a_{d-1}})^c\right)$$

$$X^{((t+1)d)}_{a_d} \sim f_{a_d}\left(x_{a_d} \middle| (\boldsymbol{x}^{((t+1)d-1)}_{a_d})^c\right)$$

The series $\boldsymbol{x}^{(0)}_{\boldsymbol{a}}, \boldsymbol{x}^{(1)}_{\boldsymbol{a}}, \boldsymbol{x}^{(2)}_{\boldsymbol{a}}, \cdots, \boldsymbol{x}^{(s)}_{\boldsymbol{a}}, \cdots$ obtained by a single *draw* (iteration) is called a realization of Gibbs chain defined by $\boldsymbol{F}$ with scan pattern $\boldsymbol{a}$; and the series $\boldsymbol{x}^{(0)}_{\boldsymbol{a}}, \boldsymbol{x}^{(d)}_{\boldsymbol{a}}, \boldsymbol{x}^{(2d)}_{\boldsymbol{a}}, \cdots, \boldsymbol{x}^{(td)}_{\boldsymbol{a}}, \cdots$ obtained by a single *cycle* is a realization of the associated Gibbs sampler. For example, let = $\{X_1, X_2, X_3, X_4\}$, $\boldsymbol{a} = (2, 4, 1, 3)$, and given initial value

$\boldsymbol{x}_{\boldsymbol{a}}^{(0)}=(x_2^{(0)}, x_4^{(0)}, x_1^{(0)}, x_3^{(0)})$, the Gibbs chain in cycle 1 performs the following draws and produces the corresponding states:

$$X_2^{(1)}\sim f_2\left(x_2\middle|X_4=x_4^{(0)}, X_1=x_1^{(0)}, X_3=x_3^{(0)}\right), \boldsymbol{x}_{\boldsymbol{a}}^{(1)}=(x_2^{(1)}, x_4^{(0)}, x_1^{(0)}, x_3^{(0)});$$

$$X_4^{(2)}\sim f_4\left(x_4\middle|X_2=x_2^{(1)}, X_1=x_1^{(0)}, X_3=x_3^{(0)}\right), \boldsymbol{x}_{\boldsymbol{a}}^{(2)}=(x_2^{(1)}, x_4^{(2)}, x_1^{(0)}, x_3^{(0)});$$

$$X_1^{(3)}\sim f_1\left(x_1\middle|X_2=x_2^{(1)}, X_4=x_4^{(2)}, X_3=x_3^{(0)}\right), \boldsymbol{x}_{\boldsymbol{a}}^{(3)}=(x_2^{(1)}, x_4^{(2)}, x_1^{(3)}, x_3^{(0)});\text{and}$$

$$X_3^{(4)}\sim f_3\left(x_3\middle|X_2=x_2^{(1)}, X_4=x_4^{(2)}, X_1=x_1^{(3)}\right), \boldsymbol{x}_{\boldsymbol{a}}^{(3)}=\left(x_2^{(1)}, x_4^{(2)}, x_1^{(3)}, x_3^{(4)}\right).$$

In this example, the series $\boldsymbol{x}_{\boldsymbol{a}}^{(0)}, \boldsymbol{x}_{\boldsymbol{a}}^{(4)}, \boldsymbol{x}_{\boldsymbol{a}}^{(8)}\left(=\left(x_2^{(5)}, x_4^{(6)}, x_1^{(7)}, x_3^{(8)}\right)\right), \cdots$, is the realization of Gibbs sampler defined by $\boldsymbol{F}$ with scan pattern $\boldsymbol{a}$.

We can also express a Gibbs sampler of random scan order as a Gibbs chain. Let $\boldsymbol{r} = \{r_1, r_2, \cdots, r_d\}$ be the set of selection probabilities, where $r_k > 0$ is the probability of visiting a conditional f_k, and $\sum_{k=1}^{d} r_k=1$. The random-scan Gibbs sampler can be stated as follows [19]:

1. Pick an arbitrary starting vector $\boldsymbol{x}^{(0)}=(x_1^{(0)}, x_2^{(0)}, \cdots, x_d^{(0)})$.
2. At the s^{th} iteration, $s = 1, 2, \cdots$,
   i. Randomly choose $k \in \{1,2, \cdots, d\}$ with probability r_k;
   ii. $X_k^{(s)}\sim f_k\left(x_k\middle|(\boldsymbol{x}_k^{(s-1)})^c\right)$
3. Repeat step 2 until a convergence criterion is reached.

Example 1

Consider the following bivariate 2 × 2 joint distribution π and for (X_1, X_2) defined on the domain {1, 2}, with its corresponding conditional distributions $f_1(x_1|x_2)$ and $f_2(x_2|x_1)$ [20, page 242]:

$$\pi=\frac{1}{10}\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}, \text{and} f_1=\frac{1}{12}\begin{bmatrix}3 & 4\\ 9 & 8\end{bmatrix}, f_2=\frac{1}{21}\begin{bmatrix}7 & 14\\ 9 & 12\end{bmatrix}.$$

There are 4 possible states, (1, 1), (1, 2), (2, 1), and (2, 2) for the Gibbs chain. The transition from one state to another is governed by the conditional matrices f_1 and f_2 as shown in Figure 1. As a shorthand, we denote an entry in the matrix as $f_1(\cdot,\cdot)$; e.g. $f_1(1,2) = 1/3$. In order to keep track of the scan order, we denote the state in the Gibbs chain as $(\boldsymbol{x}^{(s)}|f_k)$, if the current state $\boldsymbol{x}^{(s)}$ at iteration s is the result of drawing from the conditional f_k. To fix ideas, we use a fixed-scan Gibbs sampler with $\boldsymbol{a} = (1,2)$ and the conditional distributions $\{f_1, f_2\}$. In such case, the realization of this Gibbs sampler is the collection of states $\{(\boldsymbol{x}^{(2t)}|f_2)|t = 1,2,\cdots\}$.

The local transition probability P_{a_k} [21, page 71] for two successive states of Gibbs chain, $(\boldsymbol{x}^{(s-1)}|f_{a_{k-1}})$ and $(\boldsymbol{x}^{(s)}|f_{a_k})$, can be defined by

$$P_{a_k}\left(\boldsymbol{x}^{(s-1)},\boldsymbol{x}^{(s)}\right)=\begin{cases} f_{a_k}\left(\boldsymbol{x}^{(s)}\right), & \text{if}(\boldsymbol{x}_{a_k}^{(s)})^c=(\boldsymbol{x}_{a_k}^{(s-1)})^c; \\ 0, & \text{otherwise.}\end{cases}$$

For example, as shown in Figure 1, the local transition probability from $(\boldsymbol{x}^{(2t)} = (1,1)|f_2)$ to $(\boldsymbol{x}^{(2t+1)} = (1,1)|f_1)$ is $f_1(1,1) = 1/4$, and $(\boldsymbol{x}^{(2t)} = (1,1)|f_2)$ to $(\boldsymbol{x}^{(2t+1)} = (1,2)|f_1)$ is 0.

By arranging the state in lexicographical order such that the first index changes the fastest and the last index the slowest (as shown in Figure 1), a transition probability matrix T_{a_k} can be constructed according to local transition probability P_{a_k}. In Example 1, the transition probability matrices T_1 and T_2 that correspond respectively to matrices P_1, and P_2:

$$T_1=\frac{1}{12}\begin{bmatrix}3&9&0&0\\3&9&0&0\\0&0&4&8\\0&0&4&8\end{bmatrix}\text{and}T_2=\frac{1}{21}\begin{bmatrix}7&0&14&0\\0&9&0&12\\7&0&14&0\\0&9&0&12\end{bmatrix}.$$

The matrices T_1 and T_2 have two pairs of identical rows and are idempotent but not irreducible.

As the above example illustrates, the local transition probability matrices defined by the given conditional distributions are in general not identical. Thus a Gibbs chain implemented using either fixed or random scan order is not homogeneous. However, if one defines a surrogate transition probability matrix $T_{\boldsymbol{a}} \equiv T_{a_1}T_{a_2}T_{a_3}\cdots T_{a_d}$, then a homogeneous chain with transition matrix $T_{\boldsymbol{a}}$ can be formed for the scan pattern $\boldsymbol{a} = (a_1, a_2, \cdots, a_d)$ [21]. In other words, for a collection of full conditional distributions $\boldsymbol{F}$ and a scan pattern $\boldsymbol{a}$ the fixed-scan Gibbs sampler is a homogeneous Markov chain with transition matrix $T_{\boldsymbol{a}}$. Analogously, a random-scan Gibbs sampler with selection probability $\boldsymbol{r} = (r_1, r_2, \cdots, r_d)$ can be also transferred to a homogeneous Markov chain by defining $T_{\boldsymbol{r}} \equiv \sum_{k=1}^{d} r_k T_k$ as the surrogate transition probability matrix [21]. Their corresponding stationary distributions $\pi_{\boldsymbol{a}}$ and $\pi_{\boldsymbol{r}}$ can be directly computed (under a mild condition) respectively by evaluating $\lim_{m\to\infty}(T_{\boldsymbol{a}})^m=\mathbf{1}_{\mathbf{C}}\pi_{\boldsymbol{a}}^T$ and $\lim_{m\to\infty}(T_{\boldsymbol{r}})^m=\mathbf{1}_{\mathbf{C}}\pi_{\boldsymbol{r}}^T$, where $\mathbf{C}=\prod_{k=1}^{d}C_k$, and $\mathbf{1}_{\mathbf{C}}$ is a $\mathbf{C}$-dimensional vector of 1's [22].

3. SOME ANALYTIC RESULTS

In this section, we offer several general results regarding the behaviors of the fixed-scan and the random-scan Gibbs sampler for discrete variables in which the transition matrices are finite. For these results, it is not necessary to assume compatibility, unless stated otherwise. Besides providing theoretical underpinning to the previous illustrative examples, the results here allow a closer look at the mechanisms through which incompatibility impacts the behaviors of the different Gibbs sampling schemes. Note that these results are special cases that can be derived from more general theories for Markov chains, but for our purpose, focusing on the special case of discrete variables and scan patterns makes it easier to examine the dynamics of convergence. General results regarding convergence of Markov chains can be found elsewhere [5, 24]. All of the proofs of the following results are included in the Appendix.

Theorem 1

If $\boldsymbol{F}$ *is positive* ($f_k > 0, \forall\, k$) *then the Gibbs sampler, either fixed-scan with a scan pattern* $\boldsymbol{a}$ *or random scan with selection probability* $\boldsymbol{r} > 0$, *converges to a unique stationary distribution* $\pi_{\boldsymbol{a}}$ *and* $\pi_{\boldsymbol{r}}$ *respectively*.

Note that Theorem 1 does not require $\boldsymbol{F}$ to be compatible. The result assures that when $\boldsymbol{F}$ is positive—a stronger condition than $\boldsymbol{F}$ being non-negative—any scan pattern can have one and only one stationary distribution. Furthermore, the transition for any fixed-scan pattern is governed by the following theorem:

Theorem 2

If $\boldsymbol{F}$ *is positive then for each state set,* $(x_1, x_2, \cdots, x_d|f_{a_k})$, $k = 1, \cdots, d$, *of the Gibbs chain with scan pattern* $\boldsymbol{a} = (a_1, a_2, \cdots, a_d)$ *has exactly one stationary distribution* π_{a_k}. *In particular,* $\pi_{a_1}^T = \pi_{a_d}^T T_{a_1}$ *and* $\pi_{a_k}^T = \pi_{a_{k-1}}^T T_{a_k}$, $k = 2, \cdots, d$, *and* $\pi_{a_d} = \pi_{\boldsymbol{a}}$.

A direct consequence of Theorem 2 is that for any fixed-scan pattern, one of the specified conditional distributions in $\boldsymbol{F}$ can always be derived from its stationary distribution. This is summarized in the following corollary:

Corollary 1

If $\boldsymbol{F}$ *is positive then the stationary distribution* $\pi_{\boldsymbol{a}}$ *of the Gibbs sampler has* f_{a_d} *as one of its conditional distributions for the scan pattern* $\boldsymbol{a} = (a_1, a_2, \cdots, a_d)$, *i.e.*, $\pi_{\boldsymbol{a}}(x_{a_d}|x_{a_1}, x_{a_2}, \cdots, x_{a_{d-1}}) = f_{a_d}$.

When $\boldsymbol{F}$ is compatible, all scan patterns converge to the same joint distribution. The following theorem provides a formal statement.

Theorem 3

Given $\boldsymbol{F}$ *is positive.* $\boldsymbol{F}$ *is compatible if and only if there exists a unique joint distribution* π *with either* $\pi_{\boldsymbol{a}} = \pi, \forall\, \boldsymbol{a}$ *or* $\pi_{\boldsymbol{r}} = \pi, \forall\, \boldsymbol{r}$. *Furthermore,* π *is the joint distribution characterized by* $\boldsymbol{F}$.

An interesting observation about the random scan is that it forms a mixture of the fixed-scan patterns only for $d = 2$. We state the Corollary for the case $d = 2$ and give a counter-example for $d = 3$.

Corollary 2

If $\boldsymbol{F} > 0$ *and* $d = 2$ *then* $\pi_{\boldsymbol{r}}$, $\boldsymbol{r} = (r, 1 - r)$, *is formed by the convex combination of* $\pi_{\boldsymbol{a}_2 = (2,1)}$ *and* $\pi_{\boldsymbol{a}_1 = (1,2)}$; *i.e., for all* $r \in [0, 1]$, $\pi_{\boldsymbol{r}} = (1 - r)\,\pi_{\boldsymbol{a}_1} + r\pi_{\boldsymbol{a}_2}$.

A three-dimensional counter example to Corollary 2 for the case $d = 3$ is presented in Table 1. In this example, $\boldsymbol{F} = \{f_1, f_2, \cdots, f_3\}$ is positive but not compatible. There are a total of six scan patterns and for each scan pattern, the solution to which the individual Gibbs sampler converges (using matrix multiplication) is shown as a row in Table 1. Convergence is defined here as all cell-wise differences between estimates from two consecutive iterations to be less than 0.5×10^{-4}. The average of all six fixed-scan Gibbs sampler $\overline{\pi} = \frac{1}{6}\sum_{i=1}^{6} \pi_{\boldsymbol{a}_i}$ is provided, as well as a reference. In order to solve for a non-negative linear combination (mixture) of the fixed-scan distributions,

$$\pi_{\boldsymbol{r}_0} = \sum_{i=1}^{6} c_i \pi_{\boldsymbol{a}_i} \quad (1)$$

where $\boldsymbol{r}_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, we treated equation (1) as a system of linear equations and solved for $\boldsymbol{c} = (c_i)$, $i = 1, \cdots, 6$. As it turned out, our result indicated that there was no solution that satisfied $\boldsymbol{c} > 0$. This observation led us to believe that the surmise [11] that the stationary distribution for a random-scan Gibbs sampler is a mixture of the stationary distributions for all systematic scan Gibbs samplers is not true in general. It only holds for $d = 2$.

4. NUMERICL EXAMPLES

We illustrate the mathematical results using two numerical examples. In Example 1, the transition matrices for fixed- and random-scans are respectively $T_{\boldsymbol{a}_1} = T_1 T_2$ for $\boldsymbol{a}_1 = (1,2)$ and $T_{\boldsymbol{a}_2} = T_2 T_1$ for $\boldsymbol{a}_2 = (2,1)$ and $T_{\boldsymbol{r}} = (T_1 + T_2)/2$. Table 2 directly compares the joint distributions obtained from the following computations: (i) direct MCMC Gibbs sampler for the only two possible fixed-scan patterns $\boldsymbol{a}_1 = (1,2)$ and $\boldsymbol{a}_2 = (2,1)$; (ii) direct MCMC Gibbs sampler for random-scan patterns with the following selection probabilities: $\boldsymbol{r}_0 = \frac{1}{2}(1, 1)$; $\boldsymbol{r}_1 = \frac{1}{3}(1, 2)$, and $\boldsymbol{r}_2 = \frac{1}{3}(2, 1)$, and (iii) matrix multiplication using $(T_{\boldsymbol{a}})^m$ and $(T_{\boldsymbol{r}})^m$ with power m. To achieve numerical convergence, we used the first 5,000 cycles as burn-in and the subsequent 1,000,000 cycles for sampling for both (i) and (ii). And the smallest $m \in \{2^k | k = 0,1,2,\cdots\}$ such that all cell-wise differences between estimates from two consecutive iterations to be less than 0.5×10^{-4} is adopted for (iii).

As expected, both the fixed-scan, regardless of scan order, and the random-scan Gibbs samplers numerically converge to the same joint distribution. Table 2 also demonstrates that direct matrix multiplication of the transition probabilities produces rapid convergence even for a small m and different values of $\boldsymbol{r}$. However, we also observed that if $\boldsymbol{r}$ was heavily

imbalanced, it took many more iterations to achieve numerical convergence (not shown). For example, if $\boldsymbol{r}=\left(\frac{1}{10},\frac{9}{10}\right)$, it took $m > 120$ to achieve the same numerical convergence (up to 4 decimal places).

Example 2

(Incompatible conditional distributions). Consider a pair of 2 × 2 conditional distributions $f_1(x_1|x_2)$ and $f_2(x_2|x_1)$ defined on the domain {1, 2} as follows [20, page 242]:

$$f_1=\frac{1}{12}\begin{bmatrix}3 & 4\\ 9 & 8\end{bmatrix} \text{ and } f_2=\frac{1}{30}\begin{bmatrix}10 & 20\\ 3 & 27\end{bmatrix}.$$

These two conditional distributions are not compatible. Table 3 shows the results for the joint distributions derived from the simulated Gibbs samplers and matrix-multiplication of Example 2 for conditions that are identical to those presented in Table 2. Several observations can be made here: (1) The Gibbs samplers that use the fixed-scan pattern $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ respectively converge to two distinct joint distributions; (2) Each individual fixed-scan Gibbs sampler converges to the corresponding solution computed from the matrix-multiplication method; (3) The random-scan Gibbs sampler converges to the mixture distribution of the individual fixed-scan distributions—i.e., $\pi_{\boldsymbol{r}} = (1 - r)\,\pi_{\boldsymbol{a}_1} + r\pi_{\boldsymbol{a}_2}$; and (4) Regardless $\boldsymbol{F}$ is compatible or not, a random-scan Gibbs sampler always needs much larger m (using matrix-multiplication) to achieve numerical convergence than a fixed-scan Gibbs sampler does. Such phenomena should result from the idempotent property of the transition probability matrices, and it also implies slower convergence of random-scan Gibbs sampler should be expected in a MCMC simulation.

5. DISUSSION

In this paper, we show that for a given scan pattern, a homogeneous Markov chain is formed by the Gibbs sampling procedure and under mild conditions, the Gibbs sampler converges to a unique stationary distribution. Unlike compatible distributions, different scan patterns lead to different stationary distributions for PICSD. The random-scan Gibbs sampler generally converges to “something in between” but the exact weighted equation only holds for simple cases – i.e., when the dimension is two.

Our findings have several implications for the practical application of the Gibbs sampler, especially when they operate on PICSD. For example, the MICE often relies on a single fixed-scan pattern. This implies that the imputed missing values could change beyond expected statistical bounds when a seemingly innocuous change in the order of the variable is being made. Although in this paper we have not studied the issue of which fixed-scan pattern produces the “best” joint distribution, some recent work has been done in that direction. For example, Chen, Ip, and Wang [8] proposed using an ensemble approach to derive an optimal joint density. The authors also showed that the random-scan procedure generally produces promising joint distributions. It is possible that in some cases the gain from using multiple Gibbs chains, as in the case of random-scan, is marginal. As argued by

Heckerman et al. [7], the single-chain fixed-scan (pseudo) Gibbs sampler asymptotically works well when the extent to which the specified conditional distributions are incompatible is minimal. This may be true for models that are applied to one single data set with a large sample size. However, the extent of incompatibility could be much higher when multiple data sets are used and when multiple sets of conditional models are specified. While it is likely that even in more complex applications a brute-force implementation of the (pseudo) Gibbs sampler will still provide some kinds of solutions, the qualities and behaviors of such "solutions" will need to be rigorously evaluated.

Acknowledgements

The work was supported by NIH grants 1R21AG042761-01 and 1U01HL101066-01.

REFERENCES

1. Hastings WK. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika. 1970; 87:97–109.
2. Geman S, Geman D. Stochastic Relaxation, Gibbs Distribution, and the Bayes Restoration of Images. IEEE Trans Pattern Anal. 1984; 6:721–741.
3. Gelfand AE, Smith FM. Sampling-based Approaches to Calculating Marginal Densities. J Am Stat Assoc. 1990; 85:398–409.
4. Smith AFM, Roberts GO. Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. J Roy Stat Soc Ser B. 1993; 55:3–23.
5. Gilks, WR.; Richardson, S.; Spiegelhalter, D. Markov Chain Monte Carlo in Practice. London: Chapman & Hall; 1996.
6. Casella G, George E. Explaining the Gibbs Sampler. Am Stat. 1992; 46:167–174.
7. Heckerman D, Chickering DM, Meek C, Rounthwaite R, Kadie C. Dependence Networks for Inference, Collaborative Filtering, and Data Visualization. J Mach Learn Res. 2000; 1:49–57.
8. Chen SH, Ip EH, Wang Y. Gibbs Ensembles for Nearly Compatible and Incompatible Conditional Models. Comput Stat Data Anal. 2011; 55:1760–1769. [PubMed: 21286232]
9. Hobert JP, Casella G. Functional Compatibility, Markov Chains and Gibbs Sampling with Improper Posteriors. J Comput Graph Stat. 1998; 7:42–60.
10. Liu JS, Wong HW, Kong A. Correlation Structure and Convergence Rate of the Gibbs Sampler with Various Scans. J Roy Stat Soc Ser B. 1995; 57:157–169.
11. Liu JS. Discussion on Statistical Inference and Monte Carlo Algorithms, by G. Casella. Test. 1996; 5:305–310.
12. Van Buuren S, Boshuizen HC, Knook DL. Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. Stat Med. 1999; 18:681–694. [PubMed: 10204197]
13. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully Conditional Specifcation in Multivariate Imputation. J Stat Comput Sim. 2006; 76:1049–1064.
14. Rubin DB. Nested Multiple Imputation of NMES via Partially Incompatible MCMC. Stat Neerl. 2003; 57:3–18.
15. Schafer, JL. Analysis of Incomplete Multivariate Data. London: Chapman & Hall; 1997.
16. Rässler, S.; Rubin, DB.; Zell, ER. Incomplete Data in Epidemiology and Medical Statistics. In: Rao, CR.; Miller, JP.; Rao, DC., editors. Handbook of Statistics 27: Epidemiology and Medical Statistics. The Netherlands: Elsevier; 2008. p. 569-601.
17. Drechsler, J.; Rässler, S. Does Convergence Really Matter?. In: Shalabh; Heumann, C., editors. Recent Advances in Linear Models and Related Areas. Heidelberg: Physica-Verlag; 2008. p. 341-355.
18. White IR, Royston P, Wood AM. Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. Stat Med. 2011; 30:377–399. [PubMed: 21225900]

19. Levine R, Casella G. Optimizing Random Scan Gibbs Samplers. J Multivar Anal. 2006; 97:2071–2100.
20. Arnold BC, Castillo E, Sarabia JM. Exact and Near Compatibility of Discrete Conditional Distributions. Comput Stat Data Anal. 2002; 40:231–252.
21. Madras, N. Lectures on Monte Carlo Methods, Providence. Rhode Island: American Mathematical Association; 2002.
22. Seneta, E. Non-negative Matrices and Markov Chains. 2nd ed.. New York: Springer; 1981.
23. Tierney L. Markov Chains for Exploring Posterior Distributions. Ann Stat. 1994; 22:1701–1728.
24. Norris, JR. Markov Chain. Cambridge, UK: Cambridge University Press; 1998.
25. Besag JE. Discussion of Markov Chains for Exploring Posterior Distributions. Ann Stat. 1994; 22:1734–1741.

APPENDIX: PROOFS OF ANALYTIC RESULTS

Proof of Theorem 1

We need a lemma to prove Theorem 1 about irreducibility (ability to reach all interesting points of the state-space) and aperiodicity (returning to a given state-space at irregular times).

Lemma 1

If $\boldsymbol{F}$ *is positive then* $T_{\boldsymbol{a}}$ *and* $T_{\boldsymbol{r}}$ *are irreducible and aperiodic w.r.t. any given* $\boldsymbol{a}$ *and* $\boldsymbol{r} > 0$.

Proof—Let $\boldsymbol{x^1}=(x_1^1, x_2^1, x_3^1, \cdots, x_d^1)$ d and $\boldsymbol{x^2}=(x_1^2, x_2^2, x_3^2, \cdots, x_d^2)$ be two states for the chain induced by $T_{\boldsymbol{a}}$ or $T_{\boldsymbol{r}}$, and define $(T)_{ij}$ as the $(ij)^{th}$ element in the matrix T. Without loss of generality, we also let $\boldsymbol{a}$ = (1,2, ⋯, d) and $\boldsymbol{r}=\frac{\mathbf{1}}{\boldsymbol{d}}(1, 1, \cdots, 1)$, $T_{\boldsymbol{a}} = T_1T_2T_3 \cdots T_d$ and $T_{\boldsymbol{r}}=\frac{1}{d}(T_1+T_2+T_3+\cdots+T_d)$.

To prove that $T_{\boldsymbol{a}}$ and $T_{\boldsymbol{r}}$ are aperiodic, we must have $(T_{\boldsymbol{a}})_{ii} > 0$ and $(T_{\boldsymbol{a}})_{ii} > 0$, $\forall\, i$. By the definition of local transition probability, we have $(T_k)_{ii} > 0$, $\forall\, k$, if $\boldsymbol{F}$ is positive.

Consequently, $(T_{\boldsymbol{a}})_{ii} \geq \prod_{k=1}^{d}(T_k)_{ii}>0$ and $(T_{\boldsymbol{r}})_{ii}=\frac{1}{d}\sum_{k=1}^{d}(T_k)_{ii}>0, \forall i$.

To prove that $T_{\boldsymbol{a}}$ and $T_{\boldsymbol{r}}$ are irreducible is equivalent to prove that $\boldsymbol{x^1}$ and $\boldsymbol{x^2}$ commute, i.e., to show the transition probability $P(\boldsymbol{x^1} \rightarrow \boldsymbol{x^2}) > \mathbf{0}$ and $P(\boldsymbol{x^2} \rightarrow \boldsymbol{x^1}) > \mathbf{0}$. Given the scan pattern $\boldsymbol{a}$ we have

$$P(\boldsymbol{x^1} \rightarrow \boldsymbol{x^2})=f_1(x_1^2, x_2^1, \cdots, x_d^1) f_2(x_1^2, x_2^2, x_3^1 \cdots, x_d^1) \cdots f_d(x_1^2, x_2^2, x_3^2, \cdots, x_d^2)>0$$

And

$$P(\boldsymbol{x^2} \rightarrow \boldsymbol{x^1})=f_1(x_1^1, x_2^2, \cdots, x_d^2) f_2(x_1^1, x_2^1, x_3^2 \cdots, x_d^2) \cdots f_d(x_1^1, x_2^1, x_3^1, \cdots, x_d^1)>0$$

Similarly, for the random-scan case we have

$$P(\boldsymbol{x^1} \to \boldsymbol{x^2}) \geq (\frac{1}{d})^d [f_1(x_1^2, x_2^1, \cdots, x_d^1) f_2(x_1^2, x_2^2, x_3^1 \cdots, x_d^1) \cdots f_d(x_1^2, x_2^2, x_3^2, \cdots, x_d^2)] > 0$$

and

$$P(\boldsymbol{x^2} \to \boldsymbol{x^1}) \geq (\frac{1}{d})^d [f_1(x_1^1, x_2^2, \cdots, x_d^2) f_2(x_1^1, x_2^1, x_3^2 \cdots, x_d^2) \cdots f_d(x_1^1, x_2^1, x_3^1, \cdots, x_d^1)] > 0.$$

It is well known that if a Markov chain is irreducible and aperiodic, then it converges to a unique stationary distribution [24]. Consequently, we have the uniqueness and existence theorem (Theorem 1) for the Gibbs sampler.

Proof of Theorem 2

We need a lemma to prove Theorem 2.

Lemma 2

If $\boldsymbol{F}$ *is positive then the stationary distribution* $\pi_{\boldsymbol{a}}$ *of the Gibbs sampler has* f_{a_d} *as one of its conditional distribution for the scan pattern* $\boldsymbol{a} = (a_1, a_2, \cdots, a_d)$, *i.e.*, $\pi_{\boldsymbol{a}}(x_{a_d}|x_{a_1}, x_{a_2}, \cdots, x_{a_{d-1}}) = f_{a_d}$.

Proof—Since $X_{a_d}^{(td)} \sim f_{a_d}\left(x_{a_d} \middle| (\boldsymbol{x}_{a_d}^{(td-1)})^c\right)$, $t = 1,2,3, \cdots$, it follows $\pi_{\boldsymbol{a}} \propto f_{a_d}(x_{a_d}|x_{a_1}, x_{a_2}, \cdots, x_{a_{d-1}})$. Consequently, $\pi_{\boldsymbol{a}}(x_{a_d}|(\boldsymbol{x}_{a_d})^c) = f_{a_d}(x_{a_d}|(\boldsymbol{x}_{a_d})^c)$.

Theorem 2 easily follows from Lemma 2.

Proof of Theorem 3

"If" part. Since $\boldsymbol{F}$ is positive and compatible, there exists a positive joint distribution $\pi > 0$ characterized by $\boldsymbol{F}$. Under the positivity assumption of π, it is well known that the Gibbs sampler governed by $\boldsymbol{F}$ determines π [25].

"Only if" part. Let $\boldsymbol{a} = (a_1, a_2, \cdots, a_d)$ be a scan pattern with $a_d = i$, $i = 1, \cdots, d$. Assuming that there exists a π such that $\pi_{\boldsymbol{a}} = \pi$, $\forall\, \boldsymbol{a}$. From Theorem 1 and Lemma 2, it follows that $\pi_{\boldsymbol{a}}(x_{a_d}|(\boldsymbol{x}_{a_d})^c) = f_{a_d}(x_{a_d}|(\boldsymbol{x}_{a_d})^c)$, $\forall\, \boldsymbol{a}$. Thus, $\pi(x_i|(\boldsymbol{x}_i)^c) = \pi_{\boldsymbol{a_i}}(x_i|(\boldsymbol{x}_i)^c) = f_i(x_i|(\boldsymbol{x}_i)^c)$, $\forall\, i$. Hence $\boldsymbol{F}$ is compatible and π is the joint distribution of $\boldsymbol{F}$.

Assuming that there exists a π such that $\pi_{\boldsymbol{r}} = \pi$, $\forall\, \boldsymbol{r}$. We only need to prove that $\pi_{\boldsymbol{a}} = \pi$, $\forall\, \boldsymbol{a}$. From Theorem 1, we have $\pi_{\boldsymbol{a}} T_{\boldsymbol{a}} = \pi_{\boldsymbol{a}}$. By the definition of random-scan Gibbs sampler, we have $\pi_{\boldsymbol{r}} T_k = \pi_{\boldsymbol{r}} = \pi$, $\forall\, k$, $\forall\, \boldsymbol{r}$. It follows that

$$\pi = \pi T_{a_d} = (\pi T_{a_{d-1}}) T_{a_d} = \cdots = \pi(T_{a_1} T_{a_2} \cdots T_{a_{d-1}} T_{a_d}) = \pi T_{\boldsymbol{a}}.$$

Form Theorem 1, π is uniquely determined by $T_{\boldsymbol{a}}$. As a result, $\pi_{\boldsymbol{r}} = \pi = \pi_{\boldsymbol{a}}$, $\forall\, \boldsymbol{r}$, $\forall\, \boldsymbol{a}$.

Proof of Corollary 1

The proof follows directly from Theorem 2 and 3.

Proof of Corollary 2

Since $\boldsymbol{F}$ is positive, $\pi_{\boldsymbol{a}_1}$, $\pi_{\boldsymbol{a}_2}$ and π are stationary distributions uniquely determined by $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{r}$, respectively. Therefore,

$$\begin{aligned}
&[r\pi_{\boldsymbol{a}_2}^T+(1-r)\pi_{\boldsymbol{a}_1}^T][rT_1+(1-r)T_2]\\
&\quad=r^2\pi_{\boldsymbol{a}_2}^T T_1+r(1-r)\pi_{\boldsymbol{a}_1}^T T_1+r(1-r)\pi_{\boldsymbol{a}_2}^T T_2+(1-r)^2\pi_{\boldsymbol{a}_1}^T T_2\\
&\quad=r^2\pi_{\boldsymbol{a}_2}^T+r(1-r)\pi_{\boldsymbol{a}_2}^T+r(1-r)\pi_{\boldsymbol{a}_1}^T+(1-r)^2\pi_{\boldsymbol{a}_1}^T\\
&\quad=r\pi_{\boldsymbol{a}_2}^T+(1-r)\pi_{\boldsymbol{a}_1}^T
\end{aligned}$$

Because $T_{\boldsymbol{r}} = rT_1 + (1 - r)\, T_2$ is the transition kernel for the random-scan Gibbs chain with selection probability $\mathbf{r}$, we have the uniquely determined $\pi_{\boldsymbol{r}}$ which equals $r\pi_{\boldsymbol{a}_2} + (1 - r)\, \pi_{\boldsymbol{a}_1}$.■

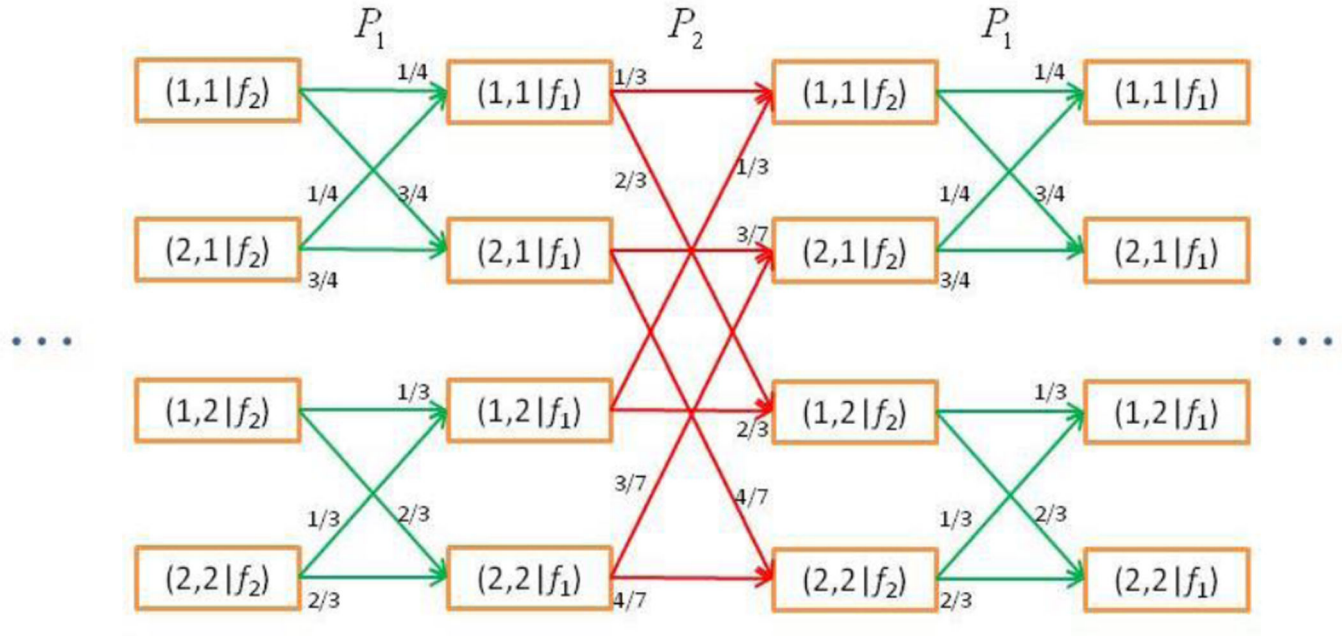

Figure 1.
Transition probabilities of the Gibbs chain in Example 1.

Table 1

A three-dimensional counter example for Corollary 2.

	(1,1,1)	**(2,1,1)**	**(1,2,1)**	**(2,2,1)**	**(1,1,2)**	**(2,1,2)**	**(1,2,2)**	**(2,2,2)**
f_1	0.1	0.9	0.2	0.8	0.3	0.7	0.4	0.6
f_2	0.5	0.6	0.5	0.4	0.7	0.8	0.3	0.2
f_3	0.9	0.1	0.1	0.9	0.1	0.9	0.9	0.1
$\pi_{\boldsymbol{a}_1} = (2,3,1)$	0.0199	0.1795	0.0411	0.1646	0.1484	0.3462	0.0401	0.0602
$\pi_{\boldsymbol{a}_2} = (3,1,2)$	0.0305	0.2064	0.0305	0.1376	0.1319	0.3251	0.0565	0.0813
$\pi_{\boldsymbol{a}_3} = (1,2,3)$	0.1462	0.0532	0.0087	0.1970	0.0162	0.4784	0.0784	0.0219
$\pi_{\boldsymbol{a}_4} = (3,2,1)$	0.0228	0.2050	0.0355	0.1421	0.1399	0.3263	0.0513	0.0770
$\pi_{\boldsymbol{a}_5} = (1,3,2)$	0.0775	0.1502	0.0775	0.1002	0.0661	0.4001	0.0283	0.1000
$\pi_{\boldsymbol{a}_6} = (2,1,3)$	0.1464	0.0531	0.0087	0.1972	0.0163	0.4782	0.0782	0.0219
$\bar{\pi}$	0.0739	0.1412	0.0337	0.1565	0.0865	0.3924	0.0555	0.0604
$\pi_{\boldsymbol{r}_0=(\frac{1}{3},\frac{1}{3},\frac{1}{3})}$	0.0728	0.1406	0.0331	0.1532	0.0873	0.3944	0.0575	0.0613

Table 2

Joint distributiuons produced by various Gibbs samplers for Example 1.

	(1,1)	**(2,1)**	**(1,2)**	**(2,2)**
π	0.1	0.3	0.2	0.4
$\boldsymbol{a}_1 = (1,2)$	0.1002	0.3002	0.2000	0.3997
$\boldsymbol{a}_1 = (1,2)$	0.0998	0.3004	0.1999	0.4000
$\boldsymbol{r}_0 = \frac{1}{2}(1,1)$	0.1007	0.2998	0.2000	0.3995
$\pi_{\boldsymbol{a}_1}(m = 4)$	0.1000	0.3000	0.2000	0.4000
$\pi_{\boldsymbol{a}_2}(m = 4)$	0.1000	0.3000	0.2000	0.4000
$\pi_{\boldsymbol{r}_0}(m = 32)$	0.1000	0.3000	0.2000	0.4000
$\pi_{\boldsymbol{r}_1}(m = 32)$	0.1000	0.3000	0.2000	0.4000
$\pi_{\boldsymbol{r}_2}(m = 32)$	0.1000	0.3000	0.2000	0.4000

Table 3

Joint distributiuons produced by various Gibbs samplers for Example 2.

	(1,1)	**(2,1)**	**(1,2)**	**(2,1)**
$\boldsymbol{a}_1 = (1,2)$	0.1062	0.0680	0.2128	0.6130
$\boldsymbol{a}_1 = (1,2)$	0.0435	0.1314	0.2753	0.5498
$\boldsymbol{r}_0 = \frac{1}{2}(1,1)$	0.0748	0.0990	0.2447	0.5815
$\pi_{\boldsymbol{a}_1}(m = 4)$	0.1063	0.0681	0.2125	0.6131
$\pi_{\boldsymbol{a}_2}(m = 4)$	0.0436	0.1308	0.2752	0.5504
$\pi_{\boldsymbol{r}_0}(m = 32)$	0.0749	0.0995	0.2439	0.5817
$\pi_{\boldsymbol{r}_1}(m = 32)$	0.0854	0.0890	0.2334	0.5922
$\pi_{\boldsymbol{r}_2}(m = 32)$	0.0645	0.1099	0.2543	0.5713