Received: 6 May 2020 | Revised: 14 June 2021 | Accepted: 19 June 2021

DOI: 10.1111/sjos.12549

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Conditional Monte Carlo revisited

Bo H. Lindqvist | Rasmus Erlemann | Gunnar Taraldsen

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
Bo H. Lindqvist, Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.
Email: bo.lindqvist@ntnu.no

Abstract

Conditional Monte Carlo refers to sampling from the conditional distribution of a random vector $\mathbf{X}$ given the value $T(\mathbf{X}) = \mathbf{t}$ for a function $T(\mathbf{X})$. Classical conditional Monte Carlo methods were designed for estimating conditional expectations of functions $\phi(\mathbf{X})$ by sampling from unconditional distributions obtained by certain weighting schemes. The basic ingredients were the use of importance sampling and change of variables. In the present paper we reformulate the problem by introducing an artificial parametric model in which $\mathbf{X}$ is a pivotal quantity, and next representing the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$ within this new model. The approach is illustrated by several examples, including a short simulation study and an application to goodness-of-fit testing of real data. The connection to a related approach based on sufficient statistics is briefly discussed.

KEYWORDS

change of variables, conditional distribution, exponential family, goodness-of-fit testing, Monte Carlo simulation, pivotal quantity, sufficiency

1 | INTRODUCTION

Suppose we want to sample from the conditional distribution of a random vector $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$ for a function $T(\mathbf{X})$ of $\mathbf{X}$. The condition $T(\mathbf{X}) = \mathbf{t}$ represents a surface in the space where $\mathbf{X}$ takes its values, and direct sampling of $\mathbf{X}$ may then be difficult. Trotter and Tukey (1956) presented an interesting technique which they named *conditional Monte Carlo.* Their idea was to determine a weight $w_{\mathbf{t}}(\mathbf{X})$ and a modified sample $\mathbf{X}_{\mathbf{t}}$ such that

$$\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X})=\mathbf{t}]=\mathrm{E}[\phi(\mathbf{X}_{\mathbf{t}})w_{\mathbf{t}}(\mathbf{X})] \tag{1}$$

for any function ϕ, thus replacing conditional expectations by ordinary expectations and allowing Monte Carlo computation.

Although the authors were aware that the method had generalizations, they confined themselves to rather special cases. Hammersley (1956) used their idea in a slightly more general and flexible analytic setting, see also chapter 6 of the monograph by Hammersley and Handscomb (1964). Wendel (1957) gave an alternative explanation, wherein the group-theoretic aspect of the problem played the dominant role. Later, Dubi and Horowitz (1979) gave an explanation of conditional Monte Carlo in terms of importance sampling and change of variables. Their approach provides a framework by which in principle any conditional sampling problem can be handled, and is the survivor in textbooks (Evans & Swartz, 2000; Ripley, 1987). Conditional Monte Carlo, in the form as introduced in the 1950s and the following nearest decades, has apparently received little attention in the later literature and has seemingly remained theoretically underdeveloped.

In this paper we present a method that can be seen as a reformulation of the main ideas of the classical concept of conditional Monte Carlo. The basic idea was essentially the introduction of new coordinates. Trotter and Tukey (1956) made a point of the "skullduggery" related to such arbitrary new variables which had "nothing to do with the way our samples were drawn." This "trick" was, however, the successful ingredient of the method, and is basically also the way our method works.

The main idea of our method is to represent $\mathbf{X}$ as a *pivotal quantity* (Casella & Berger, 2002, p. 427) in an artificial statistical model consisting of a random vector $\mathbf{U}^{\theta}$ indexed by a parameter θ, such that a transformation $\chi(\mathbf{U}^{\theta},\theta)$ has the same distribution as $\mathbf{X}$ for each θ. Then, by considering θ as the realization of a random parameter Θ, it follows that $\chi(\mathbf{U},\Theta)$ has the same distribution as $\mathbf{X}$, where $\mathbf{U}$ conditional on $\Theta=\theta$ is distributed as $\mathbf{U}^{\theta}$. Defining $\tau(\mathbf{U},\Theta)=T(\chi(\mathbf{U},\Theta))$, we hence have

$$\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X})=\mathbf{t}]=\mathrm{E}[\phi(\chi(\mathbf{U},\Theta))|\tau(\mathbf{U},\Theta)=\mathbf{t}]. \tag{2}$$

As compared with the method of Trotter and Tukey (1956), the new coordinate of our method is the parameter θ and its distribution.

The practical application of (2) essentially involves the characterization of the condition $\tau(\mathbf{U},\Theta)=\mathbf{t}$. Special attention has been given to the case when there is a unique solution $\theta=\hat{\theta}(\mathbf{u},\mathbf{t})$ of the equation $\tau(\mathbf{u},\theta)=\mathbf{t}$. Cases with multiple roots of the equation $\tau(\mathbf{u},\theta)=\mathbf{t}$ will also be briefly considered; this occurs, for example, when $\mathbf{X}$ is discrete.

Typical applications of conditional sampling are in statistical inference problems involving sufficient statistics (Lehmann & Casella, 1998; Lehmann & Romano, 2005). Engen and Lillegård (1997) considered the general problem of Monte Carlo computation of conditional expectations given a sufficient statistic. Their approach was further studied and generalized by Lindqvist and Taraldsen (2005); see also Lindqvist et al. (2003) and Lindqvist and Taraldsen (2007).

The situation considered in these papers is that $\mathbf{X}$ is the observation in a statistical model indexed by a parameter θ, and where a statistic $T(\mathbf{X})$ is sufficient for θ. A key assumption is here that $\mathbf{X}$ under the parameter θ can be simulated by a function $\chi(\mathbf{U}, \theta)$, where $\mathbf{U}$ has a known distribution not depending on θ. Now, with $\tau(\mathbf{U}, \theta) = T(\chi(\mathbf{U}, \theta))$, it is clear that $\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] = \mathrm{E}[\phi(\chi(\mathbf{U}, \theta))|\tau(\mathbf{U}, \theta) = \mathbf{t}]$ for all θ, which does not depend on θ by sufficiency. Hence equality holds also if the parameter θ is replaced by a random parameter Θ, stochastically independent of $\mathbf{U}$. We are thus again led to Equation (2), where the meaning of the ingredients has changed, though. Further developments of formulas and algorithms based on (2) for the sufficiency case are developed in Lindqvist and Taraldsen (2005).

The recent literature contains several other approaches to conditional sampling. For example, Lockhart et al. (2007) and Lockhart et al. (2009) studied the use of Gibbs sampling to generate samples from the conditional distribution given the minimal sufficient statistic for the gamma distribution and the von Mises distribution, respectively. Gracia-Medrano and O'Reilly (2005) and O'Reilly and Gracia-Medrano (2006) constructed corresponding sampling methods based on the Rao–Blackwell theorem, while Santos and Filho (2019) suggested a method using the Metropolis–Hastings algorithm. An older reference for conditional sampling in the inverse Gaussian distribution is Cheng (1984).

Other work related to the present paper is Diaconis et al. (2013) and Brubaker et al. (2012), who developed algorithms for sampling from probability distributions on submanifolds of $\mathbb{R}^n$. Bornn et al. (2019) considered models phrased through moment conditions, where posteriors are supported on manifolds. By careful representation of appropriate densities, the authors were able to sample from the posteriors by conventional simulation methods such as Markov chain Monte Carlo and importance sampling.

In the present paper we give several examples in order to demonstrate the applicability of the new approach, as well as to illustrate different aspects of the theoretical assumptions and derivations. In particular, our examples include a new method for sampling of uniformly distributed random variables conditional on their sum, where a method considered in Lindqvist and Taraldsen (2005) is apparently less attractive. Other examples study conditional sampling given sufficient statistics in the gamma and inverse Gaussian models, and we reconsider a classical example from Trotter and Tukey (1956) involving the normal distribution. While the main focus in the paper is on the case where the $\mathbf{X}$ are absolutely continuous, we also consider briefly the discrete case with an example that is relevant for logistic regression.

The paper is structured as follows. In Section 2 we give a detailed explanation of the approach and prove some basic results. The highlight of the section is Theorem 1, which can be interpreted as our version of (1). Specific methods for simulation and computation based on Theorem 1 are briefly described at the end of the section. Section 3 is devoted to examples, simulations, and a real data example involving goodness-of-fit testing. The first examples aim at illustrating the theory of Section 2. An application to a general two-parameter exponential family of positive variables is given next and provides algorithms of practical interest for conditional sampling of the gamma and inverse Gaussian distributions. An extension of the approach of Section 2 to cases of multiple solutions of the equation $\tau(\mathbf{u}, \theta) = \mathbf{t}$ is given in Section 4. Particular interest is here in the case of discrete $\mathbf{X}$. Section 5 sums up some issues regarding the construction of pivots for $\mathbf{X}$, and also considers the case of a general dimension of $T(\mathbf{X})$. Some final remarks, in particular comparing the present approach to the one of Lindqvist and Taraldsen (2005), are given in Section 6. The paper is concluded by an Appendix containing pseudocodes of the algorithms described in Section 2, and two lemmas referred to earlier in the paper.

2 | THE MAIN METHOD

2.1 | Representing the conditional distribution

Let $\mathbf{X}$ be a random vector taking values in $\mathcal{X} \subseteq \mathbb{R}^n$. Let further $T : \mathcal{X} \to \mathbb{R}^k$ be a function and consider the random vector $T(\mathbf{X})$. (Here and elsewhere in the paper, any function is assumed to be measurable). Our aim is to calculate conditional expectations or sample from the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$.

The basic assumption of our approach is that $\mathbf{X}$ is a *pivotal quantity* in a statistical model defined by a family of densities $f(\mathbf{u}|\theta)$, $\theta \in \Omega$, for the random vectors $\mathbf{U}^\theta$ taking values in an open set $\mathcal{U} \subseteq \mathbb{R}^n$ and where Ω is a parameter set. We formalize this as follows, letting "$\sim$" mean "having the same distribution as."

Assumption 1. Let $\mathbf{U}^\theta \in \mathcal{U}$ for $\theta \in \Omega$ be as defined above. Then there is a function $\chi(\mathbf{u}, \theta)$ defined for $\mathbf{u} \in \mathcal{U}$, $\theta \in \Omega$, with values in $\mathbb{R}^n$, such that

$$\chi(\mathbf{U}^\theta, \theta) \sim \mathbf{X} \quad \text{for each } \theta \in \Omega. \tag{3}$$

For examples of such a construction, see Section 3 for absolutely continuous $\mathbf{X}$ and Section 4.1.1 for discrete $\mathbf{X}$. Lemma 1 in Section 2.2 gives a general recipe for deriving $f(\mathbf{u}|\theta)$ from a given function $\chi(\mathbf{u}, \theta)$ in the continuous case.

A key observation is that statement (3) of Assumption 1 will continue to hold if the parameter θ is given a distribution on Ω, thus treating it as a random variable $\Theta \in \Omega$. Then $f(\mathbf{u}|\theta)$ will represent the conditional density of a random vector $\mathbf{U} \in \mathcal{U}$ given $\Theta = \theta$ and Assumption 1 implies that

$$\chi(\mathbf{U}, \Theta) \sim \mathbf{X}. \tag{4}$$

This is seen since, for any bounded function ϕ on $\mathbb{R}^n$,

$$\mathrm{E}[\phi(\chi(\mathbf{U}, \Theta))] = \mathrm{E}\left[\mathrm{E}[\phi(\chi(\mathbf{U}, \Theta))|\Theta\right] = \mathrm{E}[\phi(\mathbf{X})].$$

Defining $\tau(\mathbf{U}, \Theta) = T(\chi(\mathbf{U}, \Theta))$ it follows that under Assumption 1, we have

$$(\mathbf{X}, T(\mathbf{X})) \sim (\chi(\mathbf{U}, \Theta), \tau(\mathbf{U}, \Theta))$$

and from this, that the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ equals the conditional distribution of $\chi(\mathbf{U}, \Theta)$ given $\tau(\mathbf{U}, \Theta)$, that is,

$$\mathbf{X}|T(\mathbf{X}) \sim \chi(\mathbf{U}, \Theta)|\tau(\mathbf{U}, \Theta). \tag{5}$$

2.2 | Calculation of conditional expectations

Let notation and assumptions be as in Section 2.1. Let Assumption 1 be satisfied and assume now that $\mathbf{X}$ has an absolutely continuous distribution, with density $f_{\mathbf{X}}(\mathbf{x})$ on the open set $\mathcal{X} \subseteq \mathbb{R}^n$. Let further Θ have the density $\pi(\theta)$ on the open set $\Omega \subseteq \mathbb{R}^k$.

We next introduce an assumption regarding the function $\tau(\mathbf{u}, \theta)$.

Assumption 2. For any fixed $\mathbf{u} \in \mathcal{U}$, the function $\theta \mapsto \tau(\mathbf{u}, \theta)$ is one-to-one and differentiable with a differentiable inverse $\mathbf{t} \to \hat{\theta}(\mathbf{u}, \mathbf{t})$.

Assumption 2 hence assumes that the equations $\tau(\mathbf{u}, \theta) = \mathbf{t}$ can be uniquely solved for θ by $\theta = \hat{\theta}(\mathbf{u}, \mathbf{t})$ for any given $\mathbf{u}$ and $\mathbf{t}$. For possible relaxations of this assumption, see Section 4.

The following result is a key result in the derivation of the conditional distribution in (5).

Proposition 1. *Under Assumption 2 we have, for any bounded function ϕ on $\mathbb{R}^n$,*

$$\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] = \mathrm{E}[\phi(\chi(\mathbf{U}, \hat{\theta}(\mathbf{U}, \mathbf{t})))|\tau(\mathbf{U}, \Theta) = \mathbf{t}].$$

Proof. We have

$$\begin{aligned}\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] &= \mathrm{E}[\phi(\chi(\mathbf{U}, \Theta))|\tau(\mathbf{U}, \Theta) = \mathbf{t}] \\ &= \mathrm{E}[\phi(\chi(\mathbf{U}, \hat{\theta}(\mathbf{U}, \tau(\mathbf{U}, \Theta))))|\tau(\mathbf{U}, \Theta) = \mathbf{t}] \\ &= \mathrm{E}[\phi(\chi(\mathbf{U}, \hat{\theta}(\mathbf{U}, \mathbf{t})))|\tau(\mathbf{U}, \Theta) = \mathbf{t}].\end{aligned}$$

The first equality here is by (5). The second equality follows from Assumption 2. Indeed, since $\tau(\mathbf{u}, \theta) = t \Leftrightarrow \theta = \hat{\theta}(\mathbf{u}, \mathbf{t})$, we obtain the identity $\theta = \hat{\theta}(\mathbf{u}, \tau(\mathbf{u}, \theta))$. The last equality is a consequence of the substitution principle of Bahadur and Bickel (1968), noting that under the assumptions, there is a regular conditional distribution of $(\mathbf{U}, \Theta)$ given $\tau(\mathbf{U}, \Theta) = \mathbf{t}$. Here and elsewhere in the paper we tacitly assume that the conditional expectations like the ones above should be considered as functions of $\mathbf{t}$. ■

Proposition 1 shows that in order to calculate the conditional expectations, we need the conditional distribution of $\mathbf{U}$ given $\tau(\mathbf{U}, \Theta)$. This distribution is obtained from a standard transformation from $(\mathbf{U}, \Theta)$ to $(\mathbf{U}, \tau(\mathbf{U}, \Theta))$, as shown in the proof of Theorem 1 below. The main result of the theorem is Equation (6), which gives an explicit formula for calculation of conditional expectations, and essentially shows the connection to classical Monte Carlo as represented by Equation (1).

Theorem 1. *Let $\chi(\mathbf{u}, \theta)$ be given as in Assumption 1 and let $\tau(\mathbf{u}, \theta)$ satisfy Assumption 2. Let $\mathbf{U}$ for a given $\theta \in \Omega$ have density $f(\mathbf{u}|\theta)$, and let Θ have density $\pi(\theta)$ for $\theta \in \Omega$. Then, for any bounded function ϕ on $\mathbb{R}^n$, we have*

$$\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] = \frac{\int \phi(\chi(\mathbf{u}, \hat{\theta}(\mathbf{u}, \mathbf{t})))h(\mathbf{u}, \mathbf{t})d\mathbf{u}}{\int h(\mathbf{u}, \mathbf{t})d\mathbf{u}}. \tag{6}$$

Here, $h(\mathbf{u}, \mathbf{t})$ is the joint density of $(\mathbf{U}, \tau(\mathbf{U}, \Theta))$, which is given by

$$\begin{aligned}h(\mathbf{u}, \mathbf{t}) &= f(\mathbf{u}|\hat{\theta}(\mathbf{u}, \mathbf{t}))\pi(\hat{\theta}(\mathbf{u}, \mathbf{t}))|\det \partial_{\mathbf{t}}\hat{\theta}(\mathbf{u}, \mathbf{t})| \\ &= f(\mathbf{u}|\hat{\theta}(\mathbf{u}, \mathbf{t}))\pi(\hat{\theta}(\mathbf{u}, \mathbf{t}))|\det \partial_{\theta}\tau(\mathbf{u}, \theta)|^{-1}_{\theta=\hat{\theta}(\mathbf{u}, \mathbf{t})}.\end{aligned} \tag{7}$$

Proof. We first prove (7). Note that the joint density of $(\mathbf{U}, \Theta)$ is

$$f_{\mathbf{U},\Theta}(\mathbf{u}, \theta) = f(\mathbf{u}|\theta)\pi(\theta).$$

Consider now the transformation $(\mathbf{u},\theta) \to (\mathbf{u},\tau(\mathbf{u},\theta)) \equiv (\mathbf{u},\mathbf{t})$. By Assumption 2 this is one-to-one, with inverse given as $(\mathbf{u},\mathbf{t}) \to (\mathbf{u},\hat{\theta}(\mathbf{u},\mathbf{t}))$. Assumption 2 allows the use of the standard transformation formula (Rudin, 1987, theorem 7.26),

$$f_{\mathbf{U},\tau(\mathbf{U},\Theta)}(\mathbf{u},\mathbf{t}) = f_{\mathbf{U},\Theta}(\mathbf{u},\hat{\theta}(\mathbf{u},\mathbf{t})) \cdot |J|$$

where J is the Jacobi-determinant of the inverse transformation. Since the first block of the transformed vector is the identity, $\mathbf{u} \to \mathbf{u}$, it is readily seen that $J = \det \partial_{\mathbf{t}}\hat{\theta}(\mathbf{u},\mathbf{t})$. The second equality of (7) is a well-known property for inverse transformations, and gives moreover the most useful version of the density $h(\mathbf{u},\mathbf{t})$.

Now, by Proposition 1 we have

$$\begin{aligned} \mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] &= \mathrm{E}[\phi(\chi(\mathbf{U},\hat{\theta}(\mathbf{U},\mathbf{t})))|\tau(\mathbf{U},\Theta) = \mathbf{t}] \\ &= \int \phi(\chi(\mathbf{u},\hat{\theta}(\mathbf{u},\mathbf{t})))h(\mathbf{u}|\mathbf{t})d\mathbf{u} \\ &= \frac{\int \phi(\chi(\mathbf{u},\hat{\theta}(\mathbf{u},\mathbf{t})))h(\mathbf{u},\mathbf{t})d\mathbf{u}}{\int h(\mathbf{u},\mathbf{t})d\mathbf{u}}, \end{aligned}$$

where $h(\mathbf{u}|\mathbf{t})$ denotes the conditional density of $\mathbf{U}$ given $\tau(\mathbf{U},\Theta)$, which equals $h(\mathbf{u}|\mathbf{t}) = h(\mathbf{u},\mathbf{t})/\int h(\mathbf{u},\mathbf{t})d\mathbf{u}$. This proves (6). ■

Lemma 1 below shows how to derive the density $f(\mathbf{u}|\theta)$ from a given transformation $\chi(\mathbf{u},\theta)$ and a given density $f_{\mathbf{X}}(\mathbf{x})$ of $\mathbf{X}$, in order that Assumption 1 holds. Let notation and assumptions otherwise be as above.

Lemma 1. *Let $\chi : \mathcal{U} \times \Omega \to \mathbb{R}^n$ be such that, for each fixed $\theta \in \Omega$, the function $\mathbf{u} \mapsto \chi(\mathbf{u},\theta)$ is one-to-one and differentiable at every point of $\mathcal{U}$. Assume further that, for each $\theta \in \Omega$, the range $\chi(\mathcal{U},\theta)$ contains the support $\mathcal{X}$ of $\mathbf{X}$. Then Assumption 1 holds if the density of $\mathbf{U}^{\theta}$ for $\theta \in \Omega$ is given as*

$$f(\mathbf{u}|\theta) = f_{\mathbf{X}}(\chi(\mathbf{u},\theta))\,|\det \partial_{\mathbf{u}}\chi(\mathbf{u},\theta)|\,.$$

Proof. Let ϕ be an arbitrary bounded function on $\mathbb{R}^n$ and fix a $\theta \in \Omega$. Then (Rudin, 1987, theorem 7.26) we have

$$\begin{aligned} \mathrm{E}[\phi(\chi(\mathbf{U}^{\theta},\theta))] &= \int_{\mathcal{U}} \phi(\chi(\mathbf{u},\theta))f(\mathbf{u}|\theta)d\mathbf{u} \\ &= \int_{\mathcal{U}} \phi(\chi(\mathbf{u},\theta))f_{\mathbf{X}}(\chi(\mathbf{u},\theta)) \cdot |\det \partial_{\mathbf{u}}\chi(\mathbf{u},\theta)|\,d\mathbf{u} \\ &= \int_{\chi(\mathcal{U},\theta)} \phi(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \\ &= \int_{\mathcal{X}} \phi(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \\ &= \mathrm{E}[\phi(\mathbf{X})]. \end{aligned}$$

The result of the lemma then holds since ϕ was arbitrarily chosen. ■

2.3 | Methods of computation and simulation from the conditional distribution

The integrals in formula (6) of Theorem 1 will usually have an intractable form. The calculation of (6) or simulation of samples from the conditional distribution hence needs to be done by suitable numerical techniques. Some approaches are briefly described in the following, with pseudocodes given in the Appendix.

Importance sampling appears to be the traditional method used in conditional Monte Carlo, see for example Dubi and Horowitz (1979). It is particularly useful for calculating conditional expectations of the form given in (6).

In order to obtain explicit *samples* from the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$, we may first sample $\mathbf{U} = \mathbf{u}$ from a density proportional to $h(\mathbf{u}, \mathbf{t})$; then solve the equation $\tau(\mathbf{u}, \theta) = \mathbf{t}$ for $\hat{\theta}(\mathbf{u}, \mathbf{t})$; and finally return the conditional sample $\hat{\mathbf{x}} = \chi(\mathbf{u}, \hat{\theta}(\mathbf{u}, t))$. With such a recipe, rejection sampling can be used to produce independent samples from the conditional distributions, while Markov chain Monte Carlo methods produce approximate samples. We shall also below consider an alternative approximate method, which we name the *naive sampler* and use for benchmarking.

2.3.1 | Importance sampling

Consider the computation of (6). The numerator and denominator may be calculated separately by importance sampling, noting that if $\mathbf{U}$ is distributed with density $g(\mathbf{u})$, then (6) can be written

$$\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] = \frac{\mathrm{E}[\phi(\chi(\mathbf{U}, \hat{\theta}(\mathbf{U}, \mathbf{t})))h(\mathbf{U}, \mathbf{t})/g(\mathbf{U})]}{\mathrm{E}[h(\mathbf{U}, \mathbf{t})/g(\mathbf{U})]}.$$

Algorithm 1 in the Appendix gives the recipe for simulation. Note, however, the bias that is introduced by division of two averages.

2.3.2 | Rejection sampling

In rejection sampling (Ripley, 1987, p. 60) one samples from a density $g(\mathbf{u})$ with support which includes the support of $\mathbf{u} \mapsto h(\mathbf{u}, \mathbf{t})$ and for which we can find a bound $M < \infty$ such that $h/g \leq M$, see Algorithm 2 in the Appendix. It should be noted that for each new proposal $\mathbf{u}$ one needs to solve the equations leading to $\hat{\theta}(\mathbf{u}, \mathbf{t})$.

2.3.3 | Markov chain Monte Carlo

A disadvantage of rejection sampling is the need for the bound M which may be difficult to obtain. The Metropolis–Hastings algorithm (Hastings, 1970) needs no such bound but, on the other hand, produces dependent samples. Algorithm 3 in the Appendix describes a version where proposals of the Metropolis–Hastings algorithm are independent samples $\mathbf{u}$ from a density $g(\mathbf{u})$, where g, as for the rejection sampling method, needs to have a support which includes the support of $\mathbf{u} \mapsto h(\mathbf{u}, \mathbf{t})$.

2.3.4 | The naive sampler

In order to check algorithms for conditional sampling, a type of benchmark might be to use a naive sampler. Then (see Algorithm 4) $\mathbf{x}$ is sampled from $f_{\mathbf{X}}(\mathbf{x})$ and is accepted if and only if $||T(\mathbf{x}) - \mathbf{t}|| < \epsilon$ for an a priori chosen (small) $\epsilon > 0$ and an appropriate norm $|| \cdot ||$. The method has clear resemblances to the ABC-method of Bayesian statistics (Sunnåker et al., 2013). Here, parameter values are drawn from the prior distribution, and data samples $\mathbf{x}$ are then drawn from the corresponding statistical model. A sample $\mathbf{x}$ is accepted if it is close enough to the observed sample, often measured with respect to some summary statistic, which naturally corresponds to $T(\mathbf{x})$ in our approach.

3 | EXAMPLES

3.1 | Conditional sampling of uniforms

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be an i.i.d. sample from $U(0, 1)$, where $U(0, a)$ is the uniform distribution on $(0, a)$, and let $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$. Suppose one wants to sample from the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$, where $0 < t < n$. There appears to be no simple expression for this conditional distribution. Lindqvist and Taraldsen (2005) considered an approach where the uniform distribution is embedded in a parametric family involving truncated exponential distributions and utilized the sufficiency of $T(\mathbf{X})$ in this model. The resulting method is, however, surprisingly complicated. A Gibbs sampling method was devised by Lindqvist and Rannestad (2011), apparently being much quicker than the former method, and much easier to implement.

We now present a solution to the problem using the approach of Section 2. An advantage as compared with the Gibbs sampling algorithm is that the present method produces independent samples.

Let $\mathbf{U}^{\theta} = (U_1^{\theta}, U_2^{\theta}, \dots, U_n^{\theta})$ be i.i.d. realizations from $U(0, \theta)$, where $\theta > 0$. Then the U_i^{θ}/θ are i.i.d. from $U(0, 1)$, so Assumption 1 is satisfied with

$$\chi(\mathbf{u}, \theta) = \left(\frac{u_1}{\theta}, \frac{u_2}{\theta}, \dots, \frac{u_n}{\theta} \right), \tag{8}$$

defined for $\mathbf{u} \in \mathcal{U} \equiv (0, \infty)^n$ and $\theta \in \Omega \equiv (0, \infty)$, and

$$f(\mathbf{u}|\theta) = \frac{1}{\theta^n} I(\max_i u_i \leq \theta); \mathbf{u} \in \mathcal{U}, \theta \in \Omega,$$

where $I(\cdot)$ is the indicator function. Furthermore, we have

$$\tau(\mathbf{u}, \theta) = \frac{\sum_{i=1}^{n} u_i}{\theta},$$

so there exists a unique solution for θ of the equation $\tau(\mathbf{u}, \theta) = t$, given by

$$\hat{\theta}(\mathbf{u}, t) = \frac{\sum_{i=1}^{n} u_i}{t}. \tag{9}$$

Assumption 2 is hence satisfied.

Let Θ have density $\pi(\theta)$ on $\Omega = (0, \infty)$. Then by (7), the joint density of $\mathbf{U}$ and $\tau(\mathbf{U}, \Theta)$ is

$$\begin{aligned} h(\mathbf{u}, t) &= f(\mathbf{u}|\hat{\theta}(\mathbf{u}, t))\pi(\hat{\theta}(\mathbf{u}, \mathbf{t}))|\det \partial_t \hat{\theta}(\mathbf{u}, \mathbf{t})| \\ &= \frac{1}{(\hat{\theta}(\mathbf{u}, t))^n} I(\max u_i \leq \hat{\theta}(u, t)) \cdot \pi(\hat{\theta}(\mathbf{u}, \mathbf{t})) \frac{\sum_{i=1}^n u_i}{t^2} \\ &= \left(\frac{t}{\sum_{i=1}^n u_i} \right)^{n-1} I\left(\max u_i \leq \frac{\sum_{i=1}^n u_i}{t} \right) \cdot \pi \left(\frac{\sum_{i=1}^n u_i}{t} \right) \left(\frac{1}{t} \right). \end{aligned} \tag{10}$$

We then have to choose the density $\pi(\theta)$. The above expression suggests $\pi(\theta) \propto \theta^{n-1}$. In order to have a proper density $\pi(\theta)$, we need to bound its support, say, to the interval $(0, a)$ for some $a > 0$. It turns out that we may without loss of generality let $a = 1$, in which case $\pi(\theta) = n\theta^{n-1} I(0 < \theta < 1)$. In order to obtain samples from the conditional distribution of $\mathbf{X}$ given $\sum_{i=1}^n X_i = t$, we shall hence sample $\mathbf{u} \in \mathcal{U}$ from a density in $\mathbf{u}$ proportional to

$$\begin{aligned} h(\mathbf{u}, t) &= I\left(\max u_i \leq \frac{\sum_{i=1}^n u_i}{t} \right) \cdot I\left(\sum_{i=1}^n u_i \leq t \right) \left(\frac{n}{t} \right) \\ &\propto I\left(t \cdot \max u_i \leq \sum_{i=1}^n u_i \leq t \right). \end{aligned} \tag{11}$$

The resulting conditional density of $\mathbf{U}$ given $\tau(\mathbf{U}, \Theta) = t$ is hence seen to be uniform on the set of $\mathbf{u} \in \mathcal{U}$, satisfying the restriction given by the final indicator function in (11). Note here that if $t \leq 1$, then the left inequality is always satisfied. Moreover, although the conditional density in principle is defined for all positive u_i, the inequalities inside the indicator function imply that it is positive only if $\max u_i \leq 1$. We may hence sample the u_i independently from $U(0, 1)$ and accept the sample if and only if the restriction is satisfied. Finally, for the accepted samples we conclude from (8) and (9) that the resulting conditional sample is

$$\hat{\mathbf{x}} = \left(t \frac{u_1}{\sum_{i=1}^n u_i}, t \frac{u_2}{\sum_{i=1}^n u_i}, \ldots, t \frac{u_n}{\sum_{i=1}^n u_i} \right). \tag{12}$$

It can be verified that (11) is needed in addition to (12) in order to have samples from the correct conditional distribution. Still the algorithm is very simple, and simpler than the corresponding algorithms of Lindqvist and Taraldsen (2005) and Lindqvist and Rannestad (2011) that were mentioned above.

The algorithm may be slow if t is close to 0 or n, due to the low acceptance rate of, respectively, the right and left inequality in (11) in these cases. It might then be better to use importance sampling by drawing the u_i from a density $g(u) = cu^{c-1}$ for $c > 0$, where c is small (large) if t is close to 0 (close to n). But note that this leads to sampling from a nonuniform density $h(\mathbf{u}|t)$.

We note that one may in principle use any probability density $\pi(\theta)$ on $\Omega = (0, \infty)$. As an example, let $\pi(\theta) = e^{-\theta}$ for $\theta \in (0, \infty)$. Then (10) becomes

$$h(\mathbf{u}, t) = \frac{t^{n-2}}{\left(\sum_{i=1}^n u_i\right)^{n-1}} e^{(1/t) \sum_{i=1}^n u_i} I\left(\max u_i \leq \frac{\sum_{i=1}^n u_i}{t} \right), \tag{13}$$

and the task would be to draw $\mathbf{u} \in (0, \infty)^n$ from a density proportional to this. Suppose for illustration that $n = 2$ and $t = 1$. Then (13) becomes

$$h(u_1, u_2, 1) = \frac{e^{-(u_1+u_2)}}{u_1 + u_2}, \tag{14}$$

which actually is itself a proper joint density in u_1 and u_2. By (12), the desired sample $(\hat{x}_1, \hat{x}_2)$ is hence $(u_1/(u_1 + u_2), u_2/(u_1 + u_2))$ when (u_1, u_2) is drawn from (14). A calculation shows that if (U_1, U_2) has density (14), then $U_1/(U_1 + U_2) \sim U(0, 1)$, where the latter is easily checked to be the correct conditional distribution of X_1 given $X_1 + X_2 = 1$ when X_1, X_2 are independent with $X_i \sim U(0, 1)$.

As a final remark on this example, suppose instead that we wanted to condition on $\sum_{i=1}^n X_i^r = t$ for some given $r > 0$. It is then straightforward to check that only a minor modification of the above derivation is needed. As a result, one may still sample u_i from $U(0, 1)$, but change the indicator of (11) into

$$I\left(t \cdot \max_i u_i^r \leq \sum_{i=1}^n u_i^r \leq t\right)$$

and return the samples $\hat{\mathbf{x}}$ where

$$\hat{x}_i = t^{1/r} \frac{u_i}{(\sum_{\ell=1}^n u_\ell^r)^{1/r}} \quad \text{for } i = 1, 2, \dots, n.$$

3.2 | Conditional sampling of normals

The following is a classical example in conditional Monte Carlo, see, for example, Trotter and Tukey (1956), Hammersley (1956), Granovsky (1981), and Ripley (1987). Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be i.i.d. from $N(0, 1)$. We wish to sample from the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$, where $T(\mathbf{X}) = \max_i X_i - \min_i X_i$ is the range of the sample.

Now let $\mathbf{U}^\theta = (U_1^\theta, U_2^\theta, \dots, U_n^\theta)$ be i.i.d. realizations from $N(0, \theta^2)$. Then the U_i^θ/θ are i.i.d. from $N(0, 1)$, so Assumption 1 is satisfied when

$$\chi(\mathbf{u}, \theta) = \left(\frac{u_1}{\theta}, \frac{u_2}{\theta}, \dots, \frac{u_n}{\theta}\right)$$

for $\mathbf{u} = (u_1, u_2, \dots, u_n) \in \mathcal{U} = \mathbb{R}^n$ and $\theta \in \Omega = (0, \infty)$. The situation is hence much like the one of the uniform example in Section 3.1, and by arguments similar to the ones used in that example, in particular choosing $\pi(\theta) = n\theta^{n-1} I(0 < \theta < 1)$, we arrive at

$$h(\mathbf{u}, t) = \frac{1}{(2\pi)^{n/2}} \left(\frac{n}{t}\right) \exp\left(-\frac{t^2}{2(\max_i\ u_i - \min_i\ u_i)^2} \sum_{i=1}^n u_i^2\right) I(\max_i u_i - \min_i u_i < t). \tag{15}$$

Noting that the right-hand side of (15) is less than or equal to

$$\exp\left(-\frac{1}{2}\sum_{i=1}^n u_i^2\right) I(\max_i u_i - \min_i u_i < t),$$

we can use rejection sampling (Section 2.3.2) based on sampling of i.i.d. standard normal variates. If t is small, then in order to increase the acceptance probability of the rejection sampling, it might be beneficial to use as the proposal distribution, a mixture of a standard normal and a normal distribution with small variance. For a sampled $\mathbf{u}$, the resulting conditional samples are then of the form

$$\hat{\mathbf{x}} = \left(t\frac{u_1}{\max_i u_i - \min_i u_i}, \dots , t\frac{u_n}{\max_i u_i - \min_i u_i} \right).$$

Suppose now that one wants to condition on the empirical median $\tilde{\mathbf{X}}$ of the sample $\mathbf{X}$, in addition to the range as considered above. This extension requires a transformation $\chi(\mathbf{u}, \theta)$ involving two parameters, where a natural choice in this case is

$$\chi(\mathbf{u}, \theta) = \left(\frac{u_1 - \alpha}{\beta}, \dots , \frac{u_n - \alpha}{\beta} \right),$$

with $\theta = (\alpha, \beta)$. Indeed, letting the U_i^θ be i.i.d. $N(\alpha, \beta^2)$, the pivotal condition of Assumption 1 is satisfied. Defining $T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X}))$ with $T_1(\mathbf{X}) = \tilde{\mathbf{X}}$ and $T_2(\mathbf{X}) = \max_i X_i - \min_i X_i$, it is seen that Assumption 2 is satisfied with the appropriately defined $\tau(\mathbf{u}, \theta)$. Thus by a slight extension of the argument above, we may use Theorem 1 as a basis for sampling standard normals conditional on the median and the range.

3.3 | Conditional sampling given sufficient statistics for two-parameter exponential families

The main motivation for the examples in this subsection is the need for simple techniques for simulation of conditional samples given sufficient statistics for the gamma and inverse Gaussian distributions. For other algorithms in the literature, see, for example, Lockhart et al. (2007), Gracia-Medrano and O'Reilly (2005), Cheng (1984), Lindqvist and Taraldsen (2007), and Diaconis et al. (2013). Typical applications are in goodness-of-fit testing, see the case study in Section 3.3.4 below. The present examples also demonstrate the use of Lemma 1 for derivation of $f(\mathbf{u}|\theta)$ from a given function $\chi(\mathbf{u}, \theta)$.

Suppose $\mathbf{X} = (X_1, X_2, \dots , X_n)$ is distributed as an i.i.d. sample from a two-parameter exponential family of *positive* continuously distributed random variables with density of the form

$$f(x; \eta_1, \eta_2) = c(\eta_1, \eta_2)h(x) \exp(\eta_1 s_1(x) + \eta_2 s_2(x)), \tag{16}$$

for $x \in (0, \infty)$, $(\eta_1, \eta_2) \in \mathcal{H}$, where $h(x) \geq 0$, $s_1(x), s_2(x)$ are real-valued functions defined for positive x, and $c(\eta_1, \eta_2) > 0$ for $(\eta_1, \eta_2) \in \mathcal{H}$, where $\mathcal{H}$ is the natural parameter space (Casella & Berger, 2002, chapter 3.4). The minimal sufficient statistic can then be written

$$\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = \left(\sum_{i=1}^n s_1(X_i), \sum_{i=1}^n s_2(X_i) \right).$$

Suppose that $\mathbf{t} = (t_1, t_2)$ is the observed value of $T(\mathbf{X})$, and that we want to sample $\mathbf{X} = (X_1, X_2, \dots , X_n)$ conditionally on $T(\mathbf{X}) = \mathbf{t}$. By sufficiency, samples from the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$ can be obtained by choosing any density from the given family (16) as the basic density. Let $f_X(x)$ be the chosen density and let $f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^n f_X(x_i)$.

Whereas in the examples of Sections 3.1 and 3.2 there were natural pivots in the problems, there might not be one in the present case. Thus we will take as the point of departure a function $\chi(\mathbf{u}, \theta)$ which is apparently useful for positive random variables, having two parameters since $T(\mathbf{X})$ has dimension 2.

Let

$$\chi(\mathbf{u}, \theta) = \left(\left(\frac{u_1}{\beta}\right)^\alpha, \left(\frac{u_2}{\beta}\right)^\alpha, \dots, \left(\frac{u_n}{\beta}\right)^\alpha \right), \tag{17}$$

where $\mathbf{u} = (u_1, u_2, \dots, u_n) \in \mathcal{U} = (0, \infty)^n$ and $\theta = (\alpha, \beta) \in \Omega = (0, \infty)^2$. Then Assumption 1 is satisfied by Lemma 1 if $\mathbf{U}^\theta$ has density

$$f(\mathbf{u}|\theta) = \prod_{i=1}^n \left\{ \frac{\alpha}{\beta} \left(\frac{u_i}{\beta}\right)^{\alpha-1} f_X\left(\left(\frac{u_i}{\beta}\right)^\alpha\right) \right\}. \tag{18}$$

Furthermore, Assumption 2 requires that there is a unique solution for θ of the equation

$$\tau(\mathbf{u}, \theta) = \mathbf{t},$$

which here means

$$\sum_{i=1}^n s_1\left(\left(\frac{u_i}{\beta}\right)^\alpha\right) = t_1,$$
$$\sum_{i=1}^n s_2\left(\left(\frac{u_i}{\beta}\right)^\alpha\right) = t_2.$$

Assume that there is such a unique solution $\hat{\theta}(\mathbf{u}, \mathbf{t}) = (\hat{\alpha}(\mathbf{u}, \mathbf{t}), \hat{\beta}(\mathbf{u}, \mathbf{t}))$ of these equations and that also the rest of Assumption 2 is satisfied.

If $\pi(\theta) \equiv \pi(\alpha, \beta)$ is the density of Θ, then (7) gives

$$\begin{aligned} h(\mathbf{u}, \mathbf{t}) &= f(\mathbf{u}|\hat{\theta}(\mathbf{u}, \mathbf{t}))\pi(\hat{\theta}(\mathbf{u}, \mathbf{t}))|\det \partial_\theta \tau(\mathbf{u}, \theta)|^{-1}_{\theta=\hat{\theta}(\mathbf{u},\mathbf{t})} \\ &= \frac{(\hat{\alpha}/\hat{\beta})^n \left(\prod_{i=1}^n \hat{x}_i\right)^{1-1/\hat{\alpha}} \left(\prod_{i=1}^n f_X(\hat{x}_i)\right) \pi(\hat{\alpha}, \hat{\beta})}{|\det \partial_\theta \tau(\mathbf{u}, \theta)|_{\theta=\hat{\theta}(\mathbf{u},\mathbf{t})}}, \end{aligned} \tag{19}$$

where

$$\hat{x}_i = \left(\frac{u_i}{\hat{\beta}}\right)^{\hat{\alpha}} \tag{20}$$

and

$$\begin{aligned} \det \partial_\theta \tau(\mathbf{u}, \theta)|_{\theta=\hat{\theta}(\mathbf{u},\mathbf{t})} = \frac{1}{\hat{\beta}(\mathbf{u}, \mathbf{t})} \Bigg[&\left(\sum_{i=1}^n s_1'(\hat{x}_i)\hat{x}_i\right)\left(\sum_{i=1}^n s_2'(\hat{x}_i)\hat{x}_i \log(\hat{x}_i)\right) \\ &- \left(\sum_{i=1}^n s_2'(\hat{x}_i)\hat{x}_i\right)\left(\sum_{i=1}^n s_1'(\hat{x}_i)\hat{x}_i \log(\hat{x}_i)\right)\Bigg]. \end{aligned}$$

When sampling from (19) by the Metropolis–Hastings algorithm (Section 2.3.3) it seems to be a good idea to let the proposal distribution $g(\mathbf{u})$ correspond to an i.i.d. sample from the original density (16) with parameter values equal to the maximum likelihood estimates based on the observed $\mathbf{t}$. Then the calculated $\hat{\alpha}, \hat{\beta}$ are expected to be around 1, and we therefore suggest to choose the density of Θ as

$$\pi(\alpha,\beta) = I(a_1 \le \alpha \le a_2, b_1 \le \beta \le b_2)/[(a_2-a_1)(b_2-b_1)]$$

for suitably chosen $0 < a_1 < 1 < a_2, 0 < b_1 < 1 < b_2$, see examples in Section 3.3.3.

In a practical application one would usually also have the original data $\mathbf{x} = (x_1, \dots, x_n)$ which led to the values $t_1 = T_1(\mathbf{x}), t_2 = T_2(\mathbf{x})$. The vector $\mathbf{x}$ may then be used as the initial sample of the Metropolis–Hastings simulation, and will give $\hat{\alpha} = \hat{\beta} = 1$. In this case, the successively simulated accepted conditional samples $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n)$ defined by (20) will have the correct distribution, so there is no need for a burn-in period in the Metropolis–Hastings simulations.

3.3.1 | Gamma distribution

The gamma distribution with shape parameter $k > 0$ and scale parameter $\theta > 0$ has density

$$f(x;k,\theta) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/\theta} \quad \text{for } x > 0, \tag{21}$$

which is of the form (16) with $s_1(x) = x, s_2(x) = \log x$. Hence we need to solve the equations

$$\sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{\alpha} = t_1,$$
$$\sum_{i=1}^{n} \log\left(\frac{u_i}{\beta}\right)^{\alpha} = t_2.$$

with respect to α and β. It is shown in Lemma 2 in the Appendix that there is a unique solution $(\hat{\alpha}, \hat{\beta})$ for (α, β).

We suggest using $k = \theta = 1$ in (21) to get $f_X(x) = e^{-x}$. Thus (19) gives

$$h(\mathbf{u},\mathbf{t}) = \frac{(\hat{\alpha}/\hat{\beta})^n e^{(1-1/\hat{\alpha})t_2} e^{-t_1} \pi(\hat{\alpha},\hat{\beta})}{(1/\hat{\beta})\left(t_1 t_2 - n\sum_{i=1}^{n} \hat{x}_i \log \hat{x}_i\right)},$$

which is the basis for simulation of conditional samples as already outlined.

Note that the assumption $f_X(x) = e^{-x}$, that is, the standard exponential density, means that the $(X_i/\beta)^\alpha$ are Weibull distributed with shape parameter α and scale parameter β. Thus the transformation (17), together with Lemma 1, leads to a natural pivot in Assumption 1 defined from a Weibull model.

3.3.2 | Inverse Gaussian distribution

The inverse Gaussian distribution has density which can be written as

$$f(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda}{2x} - \frac{\lambda x}{2\mu^2} + \frac{\lambda}{\mu}\right) \quad \text{for } x > 0, \tag{22}$$

TABLE 1 Values used for simulation of conditional samples

Case	Distribution	n	t_1	t_2	Sample sizes	π	ϵ_1, ϵ_2
1	Gamma	3	3	−1.7	10^4	$I_{[0.5,1.5]^2}$	1/4, 1/4
2	Gamma	10	20	1	10^4	$I_{[0.5,1.5]^2}$	1, 1
3	Inverse Gaussian	3	3	6	10^4	$I_{[0.5,1.5]^2}$	1/4, 1/4
4	Inverse Gaussian	10	20	6	10^4	$I_{[0.5,1.5]^2}$	1, 1

where $\lambda, \mu > 0$ are parameters, and which is hence of the form (16) with $s_1(x) = x, s_2(x) = 1/x$. The equations to solve are then

$$\sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{\alpha} = t_1,$$
$$\sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{-\alpha} = t_2.$$

As for the gamma case, there is a unique solution $(\hat{\alpha}, \hat{\beta})$ for (α, β), see Lemma 3 in the Appendix.

Let now $f_X(x)$ be the density obtained when $\mu = \lambda = 1$, that is,

$$f_X(x) = \sqrt{\frac{1}{2\pi x^3}} \exp\left(-\frac{1}{2x} - \frac{x}{2} + 1\right), x > 0.$$

Then $f(\mathbf{u}|\theta)$ is found from (18) and we get from (19),

$$h(\mathbf{u}, \mathbf{t}) = \frac{(\hat{\alpha}/\hat{\beta})^n \left(\prod_{i=1}^n \hat{x}_i\right)^{-1/2-1/\hat{\alpha}} e^{-(1/2)(t_1+t_2)+n} \pi(\hat{\alpha}, \hat{\beta})}{(1/\hat{\beta}) \left(t_2 \sum_{i=1}^n \hat{x}_i \log \hat{x}_i - t_1 \sum_{i=1}^n \log \hat{x}_i / \hat{x}_i\right)}.$$

It was suggested for the general case above to use the parametric model itself as a proposal distribution in Metropolis–Hastings simulations, with parameters given by the maximum likelihood estimates from the original data. Following Seshadri (2012, p. 7), the maximum likelihood estimates of the parameters in (22) are found from

$$\hat{\mu} = \overline{x}, \quad \hat{\lambda}^{-1} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{x_i} - \frac{1}{\overline{x}}\right),$$

where $\overline{x} = (1/n) \sum_{i=1}^n x_i$.

3.3.3 | A simulation study

A simulation study was performed in order to illustrate the above constructed algorithms for the gamma and inverse Gaussian distributions. The setup of the study is summarized in Table 1.

For example, in case 1, a sample $\mathbf{x}$ with $n = 3$ was given, resulting in the observed sufficient statistic $(t_1, t_2) = (3, -1.7)$ for the gamma distribution. Conditional samples were then simulated, as suggested above, using the Metropolis–Hastings algorithm with proposal distribution chosen

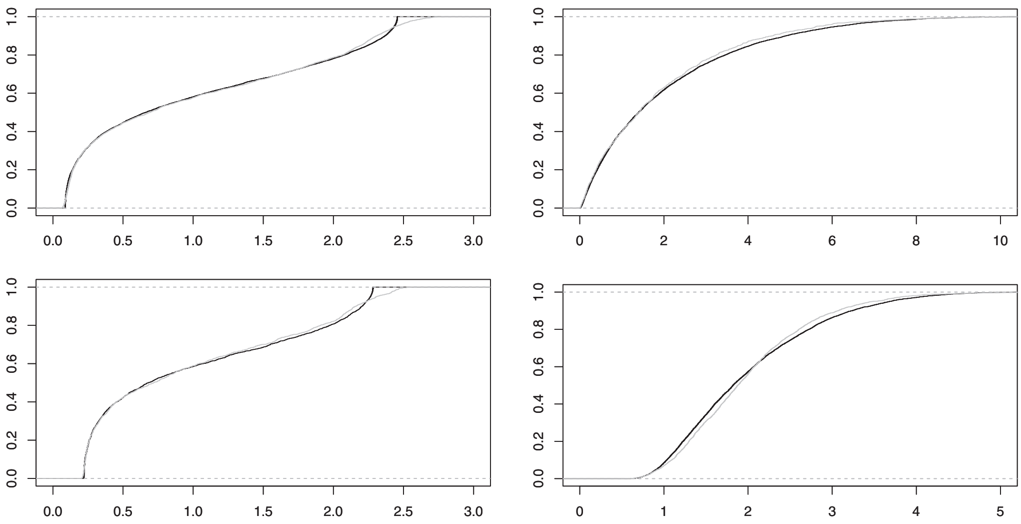

FIGURE 1 Simulated marginal cumulative distribution functions for the sampled $\hat{x}_1$ from the conditional samples for the cases of Table 1 using the Metropolis–Hastings algorithm as described in Section 3.3 (black), and the naive sampler of Section 2.3.4 (gray). Case 1: upper left. Case 2: upper right. Case 3: lower left. Case 4: lower right

as the gamma density (21) with parameters equal to the maximum likelihood estimates computed from the sufficient statistic (t_1, t_2). The density $\pi(\alpha, \beta)$ was chosen to be uniform over $(\alpha, \beta) \in [0.5, 1.5] \times [0.5, 1.5]$. In addition was applied the naive sampling method described in Section 2.3.4. Values ϵ_1, ϵ_2 (see Table 1) were chosen so that the sampler accepts an i.i.d. sample $\mathbf{x}' = (x_1', x_2', \dots, x_n')$ from the proposal distribution if and only if

$$|T_1(\mathbf{x}') - t_1| \leq \epsilon_1 \quad \text{and} \quad |T_2(\mathbf{x}') - t_2| \leq \epsilon_2.$$

In case 1 were used $\epsilon_1 = \epsilon_2 = 1/4$. Both the Metropolis–Hastings algorithm and the naive sampler were ran for enough iterations to produce at least 10^4 samples.

The description is similar for cases 2–4. Figure 1 shows, for each of the four cases in Table 1, the simulated cumulative distribution functions for the sampled $\hat{x}_1$. Although we have for illustrative purposes chosen relatively large values of ϵ_1, ϵ_2, the closeness of the curves for each case clearly indicate that the algorithms derived in the paper produce samples from the correct conditional distributions.

3.3.4 | Application to goodness-of-fit testing

As already noted, a typical use of conditional samples given sufficient statistics is in goodness-of-fit testing.

Consider the null hypothesis H_0 that an observation vector $\mathbf{X}$ comes from a particular distribution indexed by an unknown parameter θ and such that $T(\mathbf{X})$ is sufficient for θ. For a test statistic $W(\mathbf{X})$ for which large values indicate departures from the null hypothesis, we define the conditional p-value by

$$p_{\text{cond}}^{W} = P_{H_0}(W(\mathbf{X}) \geq w_* | T(\mathbf{X}) = \mathbf{t}),$$

where w_* is the observed value of the test statistic and $\mathbf{t}$ is the observed value of the sufficient statistic. A conditional goodness-of-fit test based on W rejects H_0 at significance level α if $p^W_{\text{cond}} \leq \alpha$. Let now $\hat{\mathbf{x}}_j$ for $j = 1, 2, \dots, k$ be a sample from the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$. Then the observed p-value is approximated by

$$p^W_{\text{cond}} \approx \frac{1}{k}\sum_{j=1}^{k} I(W(\hat{\mathbf{x}}_j) \geq w_*). \tag{23}$$

Consider now data from Best et al. (2012), giving the precipitation from storms in inches at the Jug Bridge in Maryland, USA. The observed data are

$$\begin{aligned}&1.01,\ 1.11,\ 1.13,\ 1.15,\ 1.16,\ 1.17,\ 1.2,\ 1.52,\ 1.54,\ 1.54,\ 1.57,\ 1.64,\\ &1.73,\ 1.79,\ 2.09,\ 2.09,\ 2.57,\ 2.75,\ 2.93,\ 3.19,\ 3.54,\ 3.57,\ 5.11,\ 5.62\end{aligned}$$

comprising the data vector $\mathbf{x} = (x_1, x_2, \dots, x_n)$, where $n = 24$. The question is whether the gamma or inverse Gaussian distributions fit the data. Using the setup and notation from Section 3.3 we calculate the sufficient statistics as

$$t_1 = \sum_{i=1}^{n} x_i = 52.72, \quad t_2 = \sum_{i=1}^{n} \log x_i = 15.7815$$

for the gamma distribution and

$$t_1 = \sum_{i=1}^{n} x_i = 52.72, \quad t_2 = \sum_{i=1}^{n} \frac{1}{x_i} = 13.8363$$

for the inverse Gaussian distribution.

Some common test statistics for goodness-of-fit testing are constructed as follows. Let $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$ be the order statistic of $\mathbf{x}$. Then define the transformed values $z_i = F(x_{(i)}\ ;\ \hat{\theta}_1, \hat{\theta}_2)$, where $F(\cdot\ ; \theta_1, \theta_2)$ is the cumulative distribution function of the gamma or inverse Gaussian distributions with parameters θ_1, θ_2, while $\hat{\theta}_1, \hat{\theta}_2$ are the maximum likelihood estimates which can be found from the corresponding t_1 and t_2.

From this setup we can write down the following test statistics:

Kolmogorov–Smirnov test (Razali & Wah, 2011)

$$D = \max_{1 \leq i \leq n}\left(z_i - \frac{i-1}{n}, \frac{i}{n} - z_i\right).$$

The Cramér–von Mises test (Stephens, 1970)

$$\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(z_i - \frac{2j-1}{2n}\right)^2.$$

Anderson–Darling test (Stephens, 1970)

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\ln z_i + \ln(1 - z_{n-i+1})\right).$$

TABLE 2 Conditional p-values

Test	Inverse Gaussian distribution	Gamma distribution
D	0.217	0.061
ω^2	0.102	0.031
A^2	0.094	0.024

Now let D_*, ω_*, A_*^2 denote the observed values of the test statistics as calculated from the observed data $\mathbf{x}$. The approximated conditional p-values p_{cond}^{D}, $p_{\text{cond}}^{\omega^2}$, $p_{\text{cond}}^{A^2}$ can now be calculated from (23) for the null hypotheses of gamma distribution and inverse Gaussian distribution, respectively.

We simulated $k = 10^5$ samples from the conditional distributions and obtained the results of Table 2. The calculated conditional p-values indicate that the fit of the inverse Gaussian distribution is marginal, which agrees with the results of Best et al. (2012). Using significance level $\alpha = 0.05$, the tests based on ω^2 and A^2 suggest that the gamma distribution does not fit the data.

4 | CONDITIONAL MONTE CARLO WITH MULTIPLE SOLUTIONS OF THE EQUATION $\tau(\mathbf{u}, \theta) = \mathbf{t}$

In general it might be difficult or impossible to find a suitable statistical model and a pivot $\chi(\mathbf{U}^\theta, \theta)$ satisfying Assumption 1, for which the uniqueness requirement of Assumption 2 is satisfied as well. Suppose therefore that Assumption 1 holds, while the equation $\tau(\mathbf{u}, \theta) = \mathbf{t}$ does not have a unique solution for θ. Below we consider first the case when $\mathbf{X}$ has a discrete distribution, and then the case of continuous $\mathbf{X}$. In the former case, the set of solutions of the equation is typically an interval in Ω, while in the latter case there are usually a finite number of distinct solutions, where the number may depend on the values of $\mathbf{u}$ and $\mathbf{t}$.

4.1 | Conditional Monte Carlo with discrete X

Let the situation be as in Section 2.1, and assume that $\mathbf{X}$ has a discrete distribution. Then the function $T(\mathbf{X})$ also has a discrete distribution. Suppose we want to calculate $\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}]$ for a given $\mathbf{t}$ with $P(T(\mathbf{X}) = \mathbf{t}) > 0$. Assume that Assumption 1 and hence (4) and (5) hold for appropriately defined $\mathbf{U}$, Θ, and functions $\chi(\mathbf{u}, \theta)$ and $\tau(\mathbf{u}, \theta)$. Then

$$\begin{aligned}\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] &= \mathrm{E}[\phi(\chi(\mathbf{U}, \Theta))|\tau(\mathbf{U}, \Theta) = \mathbf{t}] \\ &= \frac{\mathrm{E}[\phi(\chi(\mathbf{U}, \Theta))I(\tau(\mathbf{U}, \Theta) = \mathbf{t})]}{P(\tau(\mathbf{U}, \Theta) = \mathbf{t})},\end{aligned} \tag{24}$$

which makes sense since $P(\tau(\mathbf{U}, \Theta) = \mathbf{t}) = P(T(\mathbf{X}) = \mathbf{t}) > 0$.

For given $\mathbf{t}$ and $\mathbf{u}$, let $\Gamma(\mathbf{u}, \mathbf{t}) = \{\theta : \tau(\mathbf{u}, \theta) = \mathbf{t}\}$. In the discrete case, these sets are usually sets with positive Lebesgue measure, typically intervals or unions of intervals. If $\pi(\theta)$ is chosen as the density of Θ, then we can continue from (24) to get

$$\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X})=\mathbf{t}] = \frac{\int\int_{(\mathbf{u},\theta):\tau(\mathbf{u},\theta)=\mathbf{t}} \phi(\chi(\mathbf{u},\theta))f(\mathbf{u}|\theta)\pi(\theta)d\theta d\mathbf{u}}{\int\int_{(\mathbf{u},\theta):\tau(\mathbf{u},\theta)=\mathbf{t}} f(\mathbf{u}|\theta)\pi(\theta)d\theta d\mathbf{u}}$$

$$= \frac{\int\left[\int_{\theta\in\Gamma(\mathbf{u},\mathbf{t})} \phi(\chi(\mathbf{u},\theta))f(\mathbf{u}|\theta)\pi(\theta)d\theta\right]d\mathbf{u}}{\int\left[\int_{\theta\in\Gamma(\mathbf{u},\mathbf{t})} f(\mathbf{u}|\theta)\pi(\theta)d\theta\right]d\mathbf{u}}. \tag{25}$$

In many cases, including the example below, $\chi(\mathbf{u},\theta)$ is the same for all θ satisfying $\tau(\mathbf{u},\theta)=\mathbf{t}$. In such a case there is a function $x(\mathbf{u},\mathbf{t})$ such that $\chi(\mathbf{u},\theta)=x(\mathbf{u},\mathbf{t})$ for all $\theta\in\Gamma(\mathbf{u},\mathbf{t})$, and if we introduce in addition

$$h(\mathbf{u},t) = \int_{\theta\in\Gamma(\mathbf{u},t)} f(\mathbf{u}|\theta)\pi(\theta)d\theta, \tag{26}$$

we can write (25) as

$$\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X})=t] = \frac{\int \phi(x(\mathbf{u},\mathbf{t}))h(\mathbf{u},\mathbf{t})d\mathbf{u}}{\int h(\mathbf{u},\mathbf{t})d\mathbf{u}}. \tag{27}$$

This is of the same form as (6) in Theorem 1 and can hence be seen as the discrete version of (6) and hence also a version of (1). Similar results are found in Lindqvist and Taraldsen (2005) for the case where $T(\mathbf{X})$ is sufficient.

4.1.1 | Example with Bernoulli variables

Let $\mathbf{X}=(X_1,X_2,\ldots,X_n)$ be a vector of independent Bernoulli random variables where $P(X_i=1)=p_i$ for $i=1,2,\ldots,n$. Suppose we want to sample $\mathbf{X}$ conditional on $\sum_{i=1}^n X_i=t$ where $0<t<n$. A typical application is in logistic regression, see example 4 of Lindqvist and Taraldsen (2005).

Let $\mathbf{U}^\theta=(U_1^\theta,U_2^\theta,\ldots,U_n^\theta)$ be independent variables in $U(0,\theta)$, where $\theta\in\Omega=(0,\infty)$, say, and let

$$\chi(\mathbf{u},\theta)=(I(u_1<p_1\theta),\ldots,I(u_n<p_n\theta)).$$

Then $\chi(\mathbf{U}^\theta,\theta)\sim\mathbf{X}$, so Assumption 1 holds. Define

$$\tau(\mathbf{u},\theta)=\sum_{i=1}^n I(u_i<p_i\theta).$$

Let then $\psi_i(\mathbf{u})=u_i/p_i$ for $i=1,2,\ldots,n$ and let the corresponding ordered values be

$$\psi_{(1)}(\mathbf{u})<\psi_{(2)}(\mathbf{u})<\cdots<\psi_{(n)}(\mathbf{u})$$

(these are different with probability 1). Then

$$\tau(\mathbf{u},\theta)=t \Leftrightarrow \psi_{(t)}(\mathbf{u})<\theta<\psi_{(t+1)}(\mathbf{u}),$$

so the sets $\Gamma(\mathbf{u}, t) = \{\theta : \tau(\mathbf{u}, \theta) = t\}$ are intervals on Ω. It is, moreover, seen that $\chi(\mathbf{u}, \theta)$ is the same for all $\theta \in \Gamma(\mathbf{u}, t)$. Hence we may use (27) with $h(\mathbf{u}, t)$ given by (26) which here becomes

$$h(\mathbf{u}, t) = \int_{\psi_{(t)}(\mathbf{u})}^{\psi_{(t+1)}(\mathbf{u})} \theta^{-n} I(\max u_i \leq \theta)\pi(\theta)d\theta. \tag{28}$$

A sampling algorithm may hence consist in drawing $\mathbf{u}$ from a density proportional to $h(\mathbf{u}, t)$ and then returning the sample $x(\mathbf{u}, t)$. Various choices may be made for $\pi(\theta)$, for example, $\pi(\theta) \propto \theta^n$ on some bounded interval, which simplifies (28).

4.2 | Multiple solutions of the equation $\tau(\mathbf{u}, \theta) = t$ in the continuous case

Let $\mathbf{X}$ be an absolutely continuous random vector and let the situation be as considered in Section 2.2. Let

$$\Gamma(\mathbf{u}, \mathbf{t}) = \{\hat{\theta} : \tau(\mathbf{u}, \hat{\theta}) = \mathbf{t}\}$$

and assume that this is a finite set for any $\mathbf{u}, \mathbf{t}$. In this case the transformation $(\mathbf{u}, \theta) \to (\mathbf{u}, \tau(\mathbf{u}, \theta))$ is many-to-one, so there is no inverse as in the proof of Theorem 1. Instead, for a fixed $\mathbf{u}$ there is a contribution to the probability element of $(\mathbf{u}, \hat{\theta})$ for each $\hat{\theta} \in \Gamma(\mathbf{u}, \mathbf{t})$, given by

$$f(\mathbf{u}|\hat{\theta})\pi(\hat{\theta})|\det \partial_\theta \tau(\mathbf{u}, \theta)|^{-1}_{\theta=\hat{\theta}},$$

where this expression can be deduced from the proof of Theorem 1. The following formula then extends (6) to the situation of the present subsection,

$$\begin{aligned} \mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}] &= \mathrm{E}[\phi(\chi(\mathbf{U}, \Theta))|\tau(\mathbf{U}, \Theta) = t] \\ &= \frac{\int \sum_{\hat{\theta}\in\Gamma(\mathbf{u},\mathbf{t})} \phi(\chi(\mathbf{u}, \hat{\theta}))f(\mathbf{u}|\hat{\theta})\pi(\hat{\theta})|\det \partial_\theta \tau(\mathbf{u}, \theta)|^{-1}_{\theta=\hat{\theta}}d\mathbf{u}}{\int \sum_{\hat{\theta}\in\Gamma(\mathbf{u},\mathbf{t})} f(\mathbf{u}|\hat{\theta})\pi(\hat{\theta})|\det \partial_\theta \tau(\mathbf{u}, \theta)|^{-1}_{\theta=\hat{\theta}}d\mathbf{u}}. \end{aligned}$$

A corresponding result for the case when $T(\mathbf{X})$ is sufficient can be found in Lindqvist and Taraldsen (2007).

5 | CONSTRUCTION OF PIVOTS

As is clear from the construction developed in Section 2, the parameter θ of the artificial model should always have the same dimension as the statistic $T(\mathbf{X})$. This ensures that the number of equations to solve for obtaining the $\hat{\theta}(\mathbf{u}, \mathbf{t})$ in Assumption 2 is the same as the number of unknowns. This is also the underlying assumption in Section 4 in the case of multiple solutions to the equation $\tau(\mathbf{u}, \theta) = \mathbf{t}$.

In the examples of Section 3, there were natural pivots, for example, based on the scaling transformation u/θ for one-dimensional $T(\mathbf{x})$, and $(u/\beta)^\alpha$ or $(u - \beta)/\alpha$ for two-dimensional $T(\mathbf{x})$.

When there are no natural parametric families to satisfy Assumption 1, one may instead start by considering flexible functions $\chi(\mathbf{u}, \theta)$ with appropriate dimension of θ. Having chosen the function $\chi(\mathbf{u}, \theta)$, the statistical model that ensures Assumption 1 to hold, is then given by Lemma 1. Equation (18) is an example of such a construction.

While the presented examples have $T(\mathbf{X})$ with dimension one or two, a natural question is what to do in order to condition on $T(\mathbf{X})$ with dimension $k > 2$. An obvious choice for continuous $\mathbf{X}$ might be to let $\theta = (\eta_0, \eta_1, \dots, \eta_{k-1})$ and consider

$$\chi(\mathbf{u}, \theta) = \left(\sum_{j=0}^{k-1} \eta_j u_1^j, \dots, \sum_{j=0}^{k-1} \eta_j u_n^j \right). \tag{29}$$

If we put $k = 2$ in (29), then this is in fact equivalent to the transformation based on $(u - \alpha)/\beta$. For the case of positive X_i, a general suggestion might be to use

$$\chi(\mathbf{u}, \theta) = \left(\exp\left\{ \sum_{j=0}^{k-1} \eta_j u_1^j \right\}, \dots, \exp\left\{ \sum_{j=0}^{k-1} \eta_j u_n^j \right\} \right),$$

which for $k = 2$ is equivalent to the transformation based on $(u/\beta)^\alpha$.

The crucial next question is, however, whether $\tau(\mathbf{u}, \theta) = T(\chi(\mathbf{u}, \theta))$ satisfies Assumption 2. It is clear that this will not generally be the case. Appropriate modifications of $\chi(\mathbf{u}, \theta)$ may then be tried, but as considered in Section 4, one may still get around the problem when there are multiple solutions of the given equations.

We will not pursue this latter issue here. It is notable, however, that the problem of determining the proper function $\chi(\mathbf{u}, \theta)$ and checking Assumption 2 is completely connected to the function $T(\mathbf{x})$, and not to the distribution $f_{\mathbf{X}}(\mathbf{x})$ of $\mathbf{X}$. Thus, having chosen a suitable function $\chi(\mathbf{u}, \theta)$ and checked that Assumption 2 holds for our $T(\mathbf{x})$, we can via Lemma 1 and Theorem 1 perform conditional sampling of $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$ for in principle any joint distribution $f_{\mathbf{X}}(\mathbf{x})$ for $\mathbf{X}$, only noting the requirement that the range of $\chi(\mathbf{u}, \theta)$ contains that of $\mathbf{X}$. Thus, in the example of sampling normals in Section 3.2, only a slight modification of the algorithm would be necessary in order to sample from, for example, a Student distribution or a Cauchy distribution conditionally on the statistics $T(\mathbf{X})$ of that example.

6 | CONCLUDING REMARKS

In this work we have presented a new method for dealing with conditional distributions of a random vector $\mathbf{X}$ given $T(\mathbf{X}) = \mathbf{t}$ for a given function $T(\mathbf{X})$. The method is motivated by the classical notion of conditional Monte Carlo introduced by Trotter and Tukey (1956). Moreover, our approach is strongly motivated by the corresponding problem when $T(\mathbf{X})$ is a sufficient statistic, as studied by Engen and Lillegård (1997) and later by two of the present authors in Lindqvist and Taraldsen (2005).

It has been indicated in the Introduction that, although the latter approach and the one of the present paper share several features, they still lead to differing algorithms. In the new approach, an artificial parametric model is introduced, which does not need to have any connection to how the data were obtained. By contrast, the approach of Lindqvist and Taraldsen (2005) is tailored for analyzing a particular statistical model and the associated data. As discussed in the latter paper, the particular methods for conditioning may then also suggest efficient algorithms for other inference tasks such as calculation of Bayesian posteriors or fiducial distributions.

An apparent advantage of the new method is, on the other hand, the flexible choice of distributions for $\mathbf{X}$, which can be done independently of the function $T(\mathbf{x})$ and the transformation $\chi(\mathbf{u}, \theta)$, as discussed at the end of the previous section. Suppose instead we will use the method of Lindqvist and Taraldsen (2005) for calculation of conditional expectations $\mathrm{E}[\phi(\mathbf{X})|T(\mathbf{X}) = \mathbf{t}]$ in a nonstatistical setting. Lindqvist and Taraldsen (2005) suggested to embed the problem into a statistical model where $\mathbf{X}$ is observed and $T(\mathbf{X})$ is a sufficient statistic. This can be done by considering the exponential family with densities

$$f(\mathbf{x}; \theta) = c(\theta)h(\mathbf{x})e^{\theta' T(\mathbf{x})}, \tag{30}$$

where $h(\mathbf{x})$ is the original density of $\mathbf{X}$, and θ is a parameter vector of the same dimension as $T(\mathbf{X})$. As indicated by Lindqvist and Taraldsen (2005), it may, however, be difficult to find appropriate data generating functions $\chi(\mathbf{U}, \theta)$ for simulation from densities of the form (30). Indeed, in contrast to the method of the present paper, such functions will have to depend on $T(\mathbf{x})$. The derived sampling algorithms may also otherwise be untractable.

The strength of the new method can thus be described as not having to condition on sufficient statistics. Methods based on conditioning on sufficient statistics may moreover not be used when appropriate sufficient statistics cannot be found. This is the case, for example, when the minimal sufficient statistic has a higher dimension than the parameter vector. This situation was considered by Lillegård (2001), who suggested to condition on the maximum likelihood estimator. The author considered the Behrens–Fisher problem in particular. In very recent research, Barber and Janson (2020) consider resampling of data conditional on an asymptotically efficient estimator. They name their method as approximate cosufficient sampling, where the clue is to condition on a perturbed maximum likelihood estimator which is asymptotically a minimal sufficient statistic. The method presented in our paper may have a potential in such approaches.

ACKNOWLEDGMENT

The authors would like to thank the Editor in Chief, the Associate Editor, and the reviewers for their valuable comments and suggestions which led to a much improved paper.

ORCID

Bo H. Lindqvist https://orcid.org/0000-0001-8952-9311
Rasmus Erlemann https://orcid.org/0000-0002-4120-2560
Gunnar Taraldsen https://orcid.org/0000-0003-4980-7019

REFERENCES

Bahadur, R., & Bickel, P. (1968). Substitution in conditional expectation. *Annals of Mathematical Statistics*, *39*(2), 377–378.

Barber, R. F., & Janson, L. (2020). Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *arXiv preprint arXiv:2007.09851*.

Best, D. J., Rayner, J. C., & Thas, O. (2012). Comparison of some tests of fit for the inverse Gaussian distribution. *Advances in Decision Sciences*, *2012*, 1–9. https://doi.org/10.1155/2012/150303

Bornn, L., Shephard, N., & Solgi, R. (2019). Moment conditions and Bayesian non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*(1), 5–43.

Brubaker, M., Salzmann, M., & Urtasun, R. (2012). *A family of MCMC methods on implicitly defined manifolds*. In *Artificial intelligence and statistics* (pp. 161–172). PMLR.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.

Cheng, R. C. (1984). Generation of inverse Gaussian variates with given sample mean and dispersion. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *33*(3), 309–316.

Diaconis, P., Holmes, S., & Shahshahani, M. (2013). *Sampling from a manifold*. In *Advances in modern statistical theory and applications: A festschrift in honor of Morris L. Eaton* (pp. 102–125). Institute of Mathematical Statistics.

Dubi, A., & Horowitz, Y. (1979). The interpretation of conditional Monte Carlo as a form of importance sampling. *SIAM Journal of Applied Mathematics*, *36*, 115–122.

Engen, S., & Lillegård, M. (1997). Stochastic simulations conditioned on sufficient statistics. *Biometrika*, *84*(1), 235–240.

Evans, M., & Swartz, T. (2000). *Approximating integrals via Monte Carlo and deterministic methods*. Oxford University Press.

Gracia-Medrano, L., & O'Reilly, F. (2005). Transformations for testing the fit of the inverse-Gaussian distribution. *Communications in Statistics – Theory and Methods*, *33*(4), 919–924.

Granovsky, B. (1981). Optimal formulae of the conditional Monte Carlo. *SIAM Journal on Algebraic Discrete Methods*, *2*, 289–294.

Hammersley, J. (1956). Conditional Monte Carlo. *Journal of the ACM (JACM)*, *3*(2), 73–76.

Hammersley, J., & Handscomb, D. (1964). *Monte Carlo methods. Methuen's monographs on applied probability and statistics*. Methuen.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation. Springer Texts in Statistics* (2nd ed.). Springer-Verlag.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). Springer Science & Business Media.

Lillegård, M. (2001). Tests based on Monte Carlo simulations conditioned on maximum likelihood estimates of nuisance parameters. *Journal of Statistical Computation and Simulation*, *71*(1), 1–10.

Lindqvist, B. H., & Rannestad, B. (2011). Monte Carlo exact goodness-of-fit tests for nonhomogeneous Poisson processes. *Applied Stochastic Models in Business and Industry*, *27*(3), 329–341.

Lindqvist, B. H., & Taraldsen, G. (2005). Monte Carlo conditioning on a sufficient statistic. *Biometrika*, *92*(2), 451–464.

Lindqvist, B. H., & Taraldsen, G. (2007). *Conditional Monte Carlo based on sufficient statistics with applications*. In V. Nair (Ed.), *Advances in statistical modeling and inference: Essays in honor of Kjell A Doksum* (pp. 545–561). World Scientific.

Lindqvist, B. H., Taraldsen, G., Lillegård, M., & Engen, S. (2003). A counterexample to a claim about stochastic simulations. *Biometrika*, *90*(2), 489–490.

Lockhart, R. A., O'Reilly, F., & Stephens, M. (2009). Exact conditional tests and approximate bootstrap tests for the von Mises distribution. *Journal of Statistical Theory and Practice*, *3*(3), 543–554.

Lockhart, R. A., O'Reilly, F. J., & Stephens, M. A. (2007). Use of the Gibbs sampler to obtain conditional tests, with applications. *Biometrika*, *94*(4), 992–998.

O'Reilly, F., & Gracia-Medrano, L. (2006). On the conditional distribution of goodness-of-fit tests. *Communications in Statistics - Theory and Methods*, *35*(3), 541–549.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Ripley, B. (1987). *Stochastic simulation*. Wiley.

Rudin, W. (1987). *Real and complex analysis* (3rd ed.). McGraw-Hill.

Santos, J. D., & Filho, N. L. S. (2019). A Metropolis algorithm to obtain co-sufficient samples with applications in conditional tests. *Communications in Statistics - Simulation and Computation*, *48*(9), 2655–2659.

Seshadri, V. (2012). *The inverse Gaussian distribution: Statistical theory and applications*. Springer Science & Business Media.

Stephens, M. A. (1970). Use of the Kolmogorov–Smirnov, Cramer–von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *32*(1), 115–122.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian computation. *PLoS Computational Biology*, *9*(1), e1002803.

Trotter, H., & Tukey, J. (1956). *Conditional Monte Carlo for normal samples*. In H. Meyer (Ed.), *Proceedings of the Symposium On Monte Carlo Methods* (pp. 64–79). John Wiley and Sons.

Wendel, J. (1957). Groups and conditional Monte Carlo. *Annals of Mathematical Statistics*, *28*(4), 1048–1052.

How to cite this article: Lindqvist, B. H., Erlemann, R., & Taraldsen, G. (2021). Conditional Monte Carlo revisited. *Scandinavian Journal of Statistics*, 1–26. https://doi.org/10.1111/sjos.12549

APPENDIX

Pseudocode for the algorithms of Section 2.3

Algorithm 1. Pseudo-code for importance sampling. $\boldsymbol{t}$ is the value of $T(\mathbf{X})$, $\phi(\mathbf{X})$ is the function of $\mathbf{X}$, and N is the number of simulations

```
Data: $\boldsymbol{t}, \phi, N$
Result: Calculates $\mathrm{E}[\phi(\mathbf{X})|\mathbf{T} = \mathbf{t}]$
for $i = 1$ to $N$ do
    Draw $\mathbf{u}^i \sim g$;
    Solve $\tau(\boldsymbol{u}^{\boldsymbol{i}}, \boldsymbol{\theta}) = \boldsymbol{t}$ for $\hat{\theta}(\boldsymbol{u}^{\boldsymbol{i}}, \boldsymbol{t})$;
    $\mathrm{num}^i = \phi(\chi(\mathbf{u}, \hat{\theta}(\mathbf{u}^i, \mathbf{t})))h(\mathbf{u}^i, \mathbf{t})/g(\mathbf{u}^i)$;
    $\mathrm{den}^i = \theta(\mathbf{u}^i, \mathbf{t})h(\mathbf{u}^i, \mathbf{t})/g(\mathbf{u}^i)$;
end
$\exp = \mathit{mean}(\mathrm{num})/\mathit{mean}(\mathrm{den})$
```

Algorithm 2. Pseudo-code for rejection sampling. $\boldsymbol{t}$ is the value of $T(\mathbf{X})$, M is the bound for h/g, and N is the number of samples

```
Data: $\boldsymbol{t}, n, M, N$
Result: Samples $\hat{x}^i$ for $i = 1, \dots, N$
i=0;
while $i \leq N$ do
    Draw $\mathbf{u} \sim g$;
    Solve $\tau(\boldsymbol{u}, \boldsymbol{\theta}) = \boldsymbol{t}$ for $\hat{\theta}(\boldsymbol{u}, \boldsymbol{t})$Draw $z \sim \mathrm{Unif}[0, 1]$;
    if $Mz \leq h(\mathbf{u}, \mathbf{t})/g(\mathbf{u})$ then
        $\hat{\mathbf{x}}^i = \chi(\mathbf{u}, \hat{\theta}(\mathbf{u}, \mathbf{t}))$
        $i = i + 1$
    end
end
```

Algorithm 3. Pseudo-code for Metropolis–Hastings algorithm. $\boldsymbol{t}$ is the value of $T(\mathbf{X})$, and N is the number of samples

```
Data: $\boldsymbol{t}, N$
Result: Samples $\hat{x}^i$ for $i = 1, \dots, N$
initialization;
Draw $\mathbf{u}^0 \sim g$;                                   // Require $h(\mathbf{u}^0, \mathbf{t}) > 0$
for $i = 1$ to $N$ do
    Draw $\mathbf{u}' \sim g$ Solve $\tau(\boldsymbol{u}', \boldsymbol{\theta}) = \boldsymbol{t}$ for $\hat{\theta}(\boldsymbol{u}', \boldsymbol{t})$;
    Draw $z \sim \text{Unif}[0, 1]$;
    if $z \leq \frac{h(\boldsymbol{u}',\boldsymbol{t})}{h(\boldsymbol{u}^{i-1},\boldsymbol{t})} \cdot \frac{g(\boldsymbol{u}^{i-1})}{g(\boldsymbol{u}')}$ then
        $\boldsymbol{u}^i = \boldsymbol{u}'$
    else
        $\mathbf{u}^i = \mathbf{u}^{i-1}$
    end
    $\hat{\mathbf{x}}^i = \chi(\mathbf{u}, \hat{\theta}(\mathbf{u}^i, \mathbf{t}))$;
end
```

Algorithm 4. Pseudo-code for the naive sampler. $\boldsymbol{t}$ is the value of $T(\mathbf{X})$, ϵ is the error bound, and N is the number of samples

```
Data: $\boldsymbol{t}, \epsilon, N$
Result: Samples $\hat{x}^i$ for $i = 1, \dots, N$
i=0;
while $i \leq N$ do
    Draw $\mathbf{x} \sim f_{\mathbf{X}}$;
    if $||T(\mathbf{x}) - \mathbf{t}|| \leq \epsilon$ then
        $\hat{\mathbf{x}}^i = \mathbf{x}$
        $i = i + 1$
    end
end
```

Lemmas for Section 3.3

Lemma 2. *Let $n \in \mathbb{N}$ and $u_1, u_2, \dots, u_n \in \mathbb{R}^+$, and let for some $v_1, v_2, \dots, v_n \in \mathbb{R}^+$, $\sum_{i=1}^n v_i = t_1$, $\sum_{i=1}^n \ln v_i = t_2$. Then the system of equations*

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{\alpha} = t_1 \\ \displaystyle\sum_{i=1}^{n} \ln \left(\frac{u_i}{\beta}\right)^{\alpha} = t_2, \end{cases}$$

has a unique solution for $\alpha, \beta \in \mathbb{R}^+$.

Proof. We can transform the system into

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{\alpha} = t_1 \\ \displaystyle\frac{\sum_{i=1}^{n} u_i^{\alpha}}{\left(\prod_{i=1}^{n} u_i^{\alpha}\right)^{1/n}} = \frac{t_1}{\exp(t_2/n)} \end{cases}.$$

If the function

$$p(\alpha) = \frac{\sum_{i=1}^{n} u_i^{\alpha}}{\left(\prod_{i=1}^{n} u_i^{\alpha}\right)^{1/n}}$$

is monotone, then there is a unique solution. The derivative is

$$\begin{aligned} p'(\alpha) &= \left(\sum_{i=1}^{n} \left(\frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}}\right)^{\alpha}\right)' \\ &= \sum_{i=1}^{n} \left(\frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}}\right)^{\alpha} \ln \frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}}. \end{aligned}$$

We note that $\lim_{\alpha\to 0^+} p'(\alpha) = 0$. The second derivative is

$$p''(\alpha) = \sum_{i=1}^{n} \left(\frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}}\right)^{\alpha} \ln^2 \frac{u_i}{\left(\prod_{i=1}^{n} u_i\right)^{1/n}} \geq 0.$$

Since the second derivative is positive, the first derivative is increasing. Hence we can conclude that the first derivative is always positive and p is increasing. The solution exists, since

$$\lim_{\alpha\to 0^+} p(\alpha) = n$$

and

$$\frac{t_1}{\exp(t_2/n)} = \frac{\sum_{i=1}^{n} v_i}{\left(\prod_{i=1}^{n} v_i\right)^{1/n}} \geq n.$$

The last inequality holds because the arithmetic mean is always larger than or equal to the geometric mean. ■

Lemma 3. *Let $n \in \mathbb{N}$ and $u_1, u_2, \dots, u_n \in \mathbb{R}^+$, and let for some $v_1, v_2, \dots, v_n \in \mathbb{R}^+$, $\sum_{i=1}^{n} v_i = t_1$, $\sum_{i=1}^{n} v_i^{-1} = t_2$. Then the system of equations*

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{\alpha} = t_1 \\ \displaystyle\sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{-\alpha} = t_2, \end{cases}$$

has a unique solution for $\alpha, \beta \in \mathbb{R}^+$.

Proof. We can transform the system into

$$\begin{cases} \sum_{j=1}^{n} u_j^{\alpha} \sum_{i=1}^{n} u_i^{-\alpha} = t_1 t_2 \\ \sum_{i=1}^{n} \left(\frac{u_i}{\beta}\right)^{-\alpha} = t_2 \end{cases}.$$

If the function

$$p(\alpha) = \sum_{j=1}^{n} u_j^{\alpha} \sum_{i=1}^{n} u_i^{-\alpha}$$

is monotone, then there is a unique solution for α. In order to prove the monotonicity, let $y_{ij} = \frac{u_j}{u_i}$, where $i, j = 1, 2, \dots, n$, $i \neq j$. The derivative is

$$\begin{aligned} p'(\alpha) &= \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \left(\frac{u_j}{u_i}\right)^{\alpha}\right)' = \left(\sum_{j=1}^{n} \sum_{i=1}^{n} y_{ij}^{\alpha}\right)' \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} y_{ij}^{\alpha} \ln y_{ij} = \sum_{i<j} \ln y_{ij} \left(y_{ij}^{\alpha} - y_{ij}^{-\alpha}\right). \end{aligned} \tag{A1}$$

Now, if $y_{ij} > 1$, then $\ln y_{ij} > 0$ and $y_{ij}^{\alpha} > y_{ij}^{-\alpha}$, which means that

$$\ln y_{ij} \left(y_{ij}^{\alpha} - y_{ij}^{-\alpha}\right) > 0.$$

If $y_{ij} < 1$, then $\ln y_{ij} < 0$ and $y_{ij}^{\alpha} < y_{ij}^{-\alpha}$, which means that

$$\ln y_{ij} \left(y_{ij}^{\alpha} - y_{ij}^{-\alpha}\right) > 0.$$

Hence, we can conclude that (A1) is positive and the function p is increasing. Since

$$\lim_{\alpha \to 0^+} p(\alpha) = n^2$$

and

$$t_1 t_2 = \sum_{i=1}^{n} v_i \sum_{i=1}^{n} v_i^{-1} \geq n^2$$

the solution always exists. ■