*Festschrift for Kjell Doksum*

# CONDITIONAL MONTE CARLO BASED ON SUFFICIENT STATISTICS WITH APPLICATIONS

By Bo Henry Lindqvist and Gunnar Taraldsen

*Norwegian University of Science and Technology and SINTEF Information and Communication Technology*

We review and complement a general approach for Monte Carlo computations of conditional expectations given a sufficient statistic. The problem of direct sampling from the conditional distribution is considered in particular. This can be done by a simple parameter adjustment of the original statistical model if certain conditions are satisfied, but in general one needs to use a weighted sampling scheme. Several examples are given in order to demonstrate how the general method can be used under different distributions and observation plans. In particular we consider cases with, respectively, truncated and type I censored samples from the exponential distribution, and also conditional sampling for the inverse Gaussian distribution. Some new theoretical results are presented.

**1. Introduction.** We consider a pair $(X, T)$ of random vectors with joint distribution indexed by a parameter vector $\theta$. Throughout the paper we assume that $T$ is sufficient for $\theta$ compared to $X$, meaning that the conditional distribution of $X$ given $T = t$ can be specified independent of $\theta$ [Bickel and Doksum (2001), Ch. 1.5, Lehmann and Casella (1998), Ch. 1.6]. Statistical inference is often concerned with conditional expectations of the form $E\{\phi(X)|T = t\}$, which will hence not depend on the value of $\theta$. Applications include construction of optimal estimators, nuisance parameter elimination and goodness-of-fit testing.

Only in exceptional cases is one able to compute $E\{\phi(X)|T = t\}$ analytically. Typically this is not possible, thus leading to the need for approximations or simulation algorithms. Apparently because of the computational difficulties involved, methods based on conditional distributions given sufficient statistics are often not exploited in statistical applications. In fact, the literature is scarce even for the normal and multinormal distributions. Cheng

*AMS 2000 subject classifications:* Primary 62B05, 65C05; secondary 62H10, 62N05, 65C60

*Keywords and phrases:* sufficiency, conditional distribution, Monte Carlo simulation, pivotal statistic, truncated exponential distribution, type I censoring, inverse Gaussian distribution

(1984) used a result for Gamma-distributions to simulate conditional normal samples with given sample mean and sample variance, and then showed how to modify the idea to sample conditionally given the sufficient statistic for the inverse Gaussian distribution. Subsequently he extended the idea from his 1984 paper to derive a corresponding algorithm for the multivariate normal case [Cheng (1985)]. A related approach based on random rotations was recently suggested by Langsrud (2005). LT (2005) derived a method for the multinormal distribution that satisfies the pivotal condition and is based on a parametrization via Cholesky-decompositions. Diaconis and Sturmfels (1998) derived algorithms for sampling from discrete exponential families conditional on a sufficient statistic.

Engen and Lillegård (1997) considered the general problem of Monte Carlo computation of conditional expectations given a sufficient statistic. Their ideas were further developed and generalized in Lindqvist and Taraldsen (2005) [in the following referred to as LT (2005)] and in the technical report Lindqvist and Taraldsen (2001) where a more detailed measure theoretic approach was employed.

The present paper reviews basic ideas and results from LT (2005). The main purpose is to complement LT (2005) regarding computational aspects, examples and theoretical results. In particular we consider some new examples from lifetime data analysis with connections to work by Kjell Doksum [Bickel and Doksum (1969), exponential distributions; Doksum and Høyland (1992), inverse Gaussian distributions].

**2. Setup and basic algorithm.** Following LT (2005) we assume that there is given a random vector $U$ with a known distribution, such that $(X, T)$ for given $\theta$ can be simulated by means of $U$. More precisely we assume the existence of functions $\chi$ and $\tau$ such that, for each $\theta$, the joint distribution of $(\chi(U,\theta), \tau(U,\theta))$ equals the joint distribution of $(X, T)$ under $\theta$. Let in the following $f(u)$ be the probability density of $U$.

EXAMPLE 1 (*Exponential distribution*). Suppose $X = (X_1, \ldots, X_n)$ are i.i.d. from the exponential distribution with hazard rate $\theta$, denoted $\mathrm{Exp}(\theta)$. Then $T = \sum_{i=1}^n X_i$ is sufficient for $\theta$. Letting $U = (U_1, \ldots, U_n)$ be i.i.d. Exp(1) variables we can put

$$\begin{aligned} \chi(U,\theta) &= (U_1/\theta, \ldots, U_n/\theta), \\ \tau(U,\theta) &= \sum_{i=1}^n U_i/\theta. \end{aligned}$$

Consider again the general case and suppose that a sample from the conditional distribution of $X$ given $T = t$ is wanted. Since the conditional distribution by sufficiency does not depend on $\theta$, it is reasonable to believe that it can be described in some simple way in terms of the distribution of $U$, and thus enabling Monte Carlo simulation based on $U$. A suggestive method for this would be to first draw $U$ from its known distribution, then to determine a parameter value $\hat{\theta}$ such that $\tau(U, \hat{\theta}) = t$ and finally to use $X_t(U) = \chi(U, \hat{\theta})$ as the desired sample. In this way we indeed get a sample of $X$ with the corresponding $T$ having the correct value $t$. The question remains, however, whether or not $X_t(U)$ is a sample from the conditional distribution of $X$ given $T = t$.

EXAMPLE 1 *(continued).* For given $t$ and $U$ there is a unique $\hat{\theta} \equiv \hat{\theta}(U, t)$ with $\tau(U, \hat{\theta}) = t$, namely

$$\hat{\theta}(U, t) = \frac{\sum_{i=1}^{n} U_i}{t}.$$

This leads to the sample

$$X_t(U) = \chi\{U, \hat{\theta}(U, t)\} = \left( \frac{tU_1}{\sum_{i=1}^{n} U_i}, \ldots, \frac{tU_n}{\sum_{i=1}^{n} U_i} \right), \tag{2.1}$$

and it is well known [Aitchison (1963)] that the distribution of $X_t(U)$ indeed coincides with the conditional distribution of $X$ given $T = t$.

The algorithm used in Example 1 can more generally be described as follows:

ALGORITHM 1. *Conditional sampling of $X$ given $T = t$.*

1. Generate $U$ from the density $f(u)$.
2. Solve $\tau(U, \theta) = t$ for $\theta$. The (unique) solution is $\hat{\theta}(U, t)$.
3. Return $X_t(U) = \chi\{U, \hat{\theta}(U, t)\}$.

The following so called pivotal condition, discussed and verified in LT (2005), ensures that Algorithm 1 produces a sample $X_t(U)$ from the conditional distribution of $X$ given $T = t$. Note that uniqueness of $\hat{\theta}(U, t)$ in Step 2 is required.

*The pivotal condition.* Assume that $\tau(u, \theta)$ depends on $u$ only through a function $r(u)$, where the value of $r(u)$ can be uniquely recovered from the equation $\tau(u, \theta) = t$ for given $\theta$ and $t$. This means that there is a function

$\tilde{\tau}$ such that $\tau(u,\theta) = \tilde{\tau}\{r(u),\theta\}$ for all $(u,\theta)$, and a function $\tilde{v}$ such that $\tilde{\tau}\{r(u),\theta\} = t$ implies $r(u) = \tilde{v}(\theta,t)$. Note that in this case $\tilde{v}(\theta,T)$ is a pivotal quantity in the classical meaning that its distribution does not depend on $\theta$.

EXAMPLE 1 *(continued).* The pivotal condition is satisfied here with $r(U) = \sum_{i=1}^{n} U_i$. Thus Algorithm 1 is valid, as verified earlier by a direct method.

**3. General algorithm for unique $\hat{\theta}(u,t)$.** Algorithm 1 will in general not produce samples from the correct conditional distribution, even if the solution $\hat{\theta}(u,t)$ of $\tau(u,\theta) = t$ is unique. This was demonstrated by a counterexample in Lindqvist, Taraldsen, Lillegård and Engen (2003). A modified algorithm can, however, be constructed. The main idea [LT (2005)] is to consider the parameter $\theta$ as a random variable $\Theta$, independent of $U$, and with some conveniently chosen distribution $\pi$. Such an approach is similar to the one of Trotter and Tukey (1956), and this idea is also inherent in the approach of Engen and Lillegård (1997).

The key result is that the conditional distribution of $X$ given $T = t$ equals the conditional distribution of $\chi(U,\Theta)$ given $\tau(U,\Theta) = t$. This is intuitively obvious from the definition of sufficiency, which implies that this holds when $\Theta$ is replaced by any fixed value $\theta$. Note, however, that independence of $U$ and $\Theta$ is needed for this to hold. It follows that conditional expectations $E\{\phi(X)|T = t\}$ can be computed from the formula

$$E\{\phi(X)|T = t\} = E[\phi\{\chi(U,\Theta)\}|\tau(U,\Theta) = t]. \tag{3.1}$$

Assume in the rest of the section that the equation $\tau(u,\theta) = t$ has the unique solution $\hat{\theta}(u,t)$ for $\theta$. Then $\theta = \hat{\theta}\{u,\tau(u,\theta)\}$ is an identity in $\theta$ and $u$, and this fact together with (3.1) imply that

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

Thus we need only the conditional distribution of $U$ given $\tau(U,\Theta) = t$. Assuming this is given by a density $f(u|t)$, Bayes' formula implies that $f(u|t) \propto f(t|u)f(u)$, where $f(t|u)$ is the conditional density of $\tau(U,\Theta)$ given $U = u$ and $f(u)$ is the density of $U$. Now since $U$ and $\Theta$ are independent, $f(t|u)$ is simply the density of $\tau(u,\Theta)$ which we in the following denote by $W_t(u)$. It should be stressed that $W_t(u)$ is the density of $\tau(u,\Theta)$ as a function of $t$, for each fixed $u$, while in the following it will usually be considered as

a function of $u$. From this we get

$$E\{\phi(X)|T=t\} = \frac{E[\phi\{X_t(U)\}W_t(U)]}{E\{W_t(U)\}}, \tag{3.2}$$

where the denominator $E\{W_t(U)\} = \int W_t(u)f(u)du$ is merely the normalization of the conditional density $f(u|t)$. The formula shows that $W_t(u)$ acts as a weight function for a sample $u$ from $f(u)$.

It follows from (3.2) that sampling from the conditional distribution in principle can be done by the following scheme:

ALGORITHM 2. *Weighted conditional sampling of $X$ given $T=t$.*
Let $\Theta$ be a random variable and let $t \mapsto W_t(u)$ be the density of $\tau(u,\Theta)$.

1. Generate V from a density proportional to $W_t(u)f(u)$.
2. Solve $\tau(V,\theta) = t$ for $\theta$. The (unique) solution is $\hat{\theta}(V,t)$.
3. Return $X_t(V) = \chi\{V, \hat{\theta}(V,t)\}$.

*The weight function $W_t(u)$ in the Euclidean case.* Suppose that the vector $X$ has a distribution depending on a $k$-dimensional parameter $\theta$ and that $T(X)$ is a $k$-dimensional sufficient statistic. Choose a density $\pi(\theta)$ for $\Theta$ and let $W_t(u)$ be the density of $\tau(u,\Theta)$. Since $\tau(u,\theta) = t$ if and only if $\theta = \hat{\theta}(u,t)$ it follows under standard assumptions that

$$W_t(u) = \pi\{\hat{\theta}(u,t)\}|\det \partial_t \hat{\theta}(u,t)| = |\frac{\pi(\theta)}{\det \partial_\theta \tau(u,\theta)}|_{\theta=\hat{\theta}(u,t)}. \tag{3.3}$$

The formula (3.2) can thus be written

$$E\{\phi(X)|T=t\} = \frac{\int \phi[\chi\{u,\hat{\theta}(u,t)\}]|\frac{\pi(\theta)}{\det \partial_\theta \tau(u,\theta)}|_{\theta=\hat{\theta}(u,t)} f(u)du}{\int |\frac{\pi(\theta)}{\det \partial_\theta \tau(u,\theta)}|_{\theta=\hat{\theta}(u,t)} f(u)du}, \tag{3.4}$$

and can be computed by simulation using a pseudo-sample from the distribution of $U$ as will be explained in Section 6.1.

EXAMPLE 2 (*Truncated exponential lifetimes*). Let $X = (X_1,\ldots,X_n)$ be a sample from the exponential distribution with hazard rate $\theta$, but assume now that $X_i$ is an observation truncated at $\tau_i$ $(i=1,\ldots,n)$, where the $\tau_i > 0$ are known numbers. This means that the distribution function of $X_i$ is

$$F_i(x_i,\theta) = \frac{1-e^{-\theta x_i}}{1-e^{-\theta\tau_i}},\ 0 \le x_i \le \tau_i,\ i=1,\ldots,n. \tag{3.5}$$

As for the non-truncated exponential case in Example 1, the statistic $T = \sum_{i=1}^n X_i$ is sufficient for $\theta$. Suppose we wish to consider the conditional distribution of $X$ given $T = t$. It turns out to be convenient to extend the parameter set to allow $\theta$ to be any real number. Indeed, $F_i$ defined in (3.5) is a c.d.f. for all real $\theta$ if we define $F_i(x_i, 0) = x_i/\tau_i,\ 0 \le x_i \le \tau_i$, obtained by taking the limit as $\theta \to 0$ in (3.5).

Now a sample $X$ can be simulated by ordinary inversion based on (3.5) using a sample $U = (U_1, U_2, \ldots, U_n)$ from the standard uniform distribution, denoted Un$[0, 1]$. This gives
$\chi(U, \theta) = (\eta_1(U_1, \theta), \ldots, \eta_n(U_n, \theta))$, $\tau(U, \theta) = \sum_{i=1}^n \eta_i(U_i, \theta)$ where

$$\eta_i(u_i, \theta) = \begin{cases} -\log\{1 - (1 - e^{-\theta\tau_i})u_i\}/\theta & \text{if } \theta \neq 0 \\ \tau_i u_i & \text{if } \theta = 0 \end{cases}.$$

The function $\eta_i(u_i, \theta)$ is strictly decreasing in $\theta$, which follows since $F_i(x_i, \theta)$ is strictly increasing in $\theta$. Consequently the solution $\hat\theta(u, t)$ of $\tau(u, \theta) = t$ is unique.

It turns out that the pivotal condition of Section 2 is not satisfied in the present case. Indeed, Lindqvist et al. (2003) studied the case $n = 2$ and found that Algorithm 1 does not produce the correct distribution. Thus we use instead Algorithm 2 and (3.4), for which we need to compute $|\partial_\theta \tau(u, \theta)|_{\theta=\hat\theta(u,t)}$. We obtain

$$|\partial_\theta \tau(u, \theta)|_{\theta=\hat\theta(u,t)} = \frac{1}{\hat\theta(u, t)} \left( t - \sum_{i=1}^n \frac{\tau_i u_i e^{-\hat\theta(u,t)\tau_i}}{1 - (1 - e^{-\hat\theta(u,t)\tau_i})u_i} \right).$$

In principle we can then use (3.4) with any choice of the density $\pi(\theta)$ for which the integrals exist. The simple choice of $\pi(\theta) = 1/|\theta|$ turns out to work well in this example and is in accordance with the discussion in Section 6.3 regarding the use of noninformative priors.

We close the example by noting that since $\theta = 0$ corresponds to the $X_i$ being uniform, the target conditional distribution is that of $n$ independent Un$[0, \tau_i]$ random variables given their sum. There seems to be no simple expression for this distribution, not even when the $\tau_i$ are equal.

**4. The general case.** Recall the basic idea described in Section 3 that conditional expectations $E\{\phi(X)|T = t\}$ can be computed from the formula (3.1) where we have introduced the random parameter $\Theta$. In the general case, where there may not be a unique solution of $\tau(u, \theta) = t$, we compute (3.1) by conditioning on $U$ in addition to $\tau(U, \Theta) = t$. This leads to the most general result of LT (2005) which states that

$$E\{\phi(X)|T = t\} = \frac{\int Z_t(u) W_t(u) f(u) du}{\int W_t(u) f(u) du}, \tag{4.1}$$

where $Z_t(u)$ is the conditional expectation of $\phi\{\chi(u,\Theta)\}$ given $\tau(u,\Theta)=t$ for fixed $u$, $W_t(u)$ is the density of the variable $\tau(u,\Theta)$ at $t$, for fixed $u$, and $f(u)$ is the density of $U$.

Thus our method essentially amounts to changing computations of conditional expectations of $\phi\{\chi(U,\theta)\}$ given $\tau(U,\theta)=t$ for fixed $\theta$ into the often much simpler problem of computing conditional expectations of $\phi\{\chi(u,\Theta)\}$ given $\tau(u,\Theta)=t$ for fixed $u$. Note the freedom to choose a suitable distribution $\pi$ for $\Theta$.

The formula (4.1) implies the following principal scheme for simulation of $X$ given $T=t$.

ALGORITHM 3. *General weighted conditional sampling of $X$ given $T=t$.* Let $\Theta$ be a random variable and let $t\mapsto W_t(u)$ be the density of $\tau(u,\Theta)$.

1. Generate V from a density proportional to $W_t(u)f(u)$ and let the result be $V=v$.
2. Generate $\Theta_t$ from the conditional distribution of $\Theta$ given $\tau(v,\Theta)=t$.
3. Return $X_t(V)=\chi(V,\Theta_t)$.

4.1. *The general Euclidean case.* As in Section 3, suppose that the vector $X$ has a distribution depending on a $k$-dimensional parameter $\theta$ and that $T(X)$ is a $k$-dimensional sufficient statistic. In this case, the equation $\tau(u,\theta)=t$ will typically have a finite number of solutions, where this number may vary as $u$ varies. Define

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

and note that the density $t\mapsto W_t(u)$ of $\tau(u,\Theta)$ is now given by

$$W_t(u)=\sum_{\hat\theta\in\Gamma(u,t)}\frac{\pi(\hat\theta)}{|\det\partial_\theta\tau(u,\theta)|_{\theta=\hat\theta}|}. \tag{4.2}$$

Furthermore, the conditional distribution of $\Theta$ given $\tau(u,\Theta)=t$ is concentrated on $\Gamma(u,t)$ and is given by

$$Pr\{\Theta=\hat\theta\mid\tau(u,\Theta)=t\}=\frac{\pi(\hat\theta)}{|\det\partial_\theta\tau(u,\theta)|_{\theta=\hat\theta}|W_t(u)},\quad\hat\theta\in\Gamma(u,t). \tag{4.3}$$

The following formula generalizes the result (3.4):

$$\mathrm{E}\{\phi(X)|T=t\}=\frac{\int\sum_{\hat\theta\in\Gamma(u,t)}\phi(\chi(u,\hat\theta))\frac{\pi(\hat\theta)}{|\det\partial_\theta\tau(u,\theta)|_{\theta=\hat\theta}|}f(u)du}{\int\sum_{\hat\theta\in\Gamma(u,t)}\frac{\pi(\hat\theta)}{|\det\partial_\theta\tau(u,\theta)|_{\theta=\hat\theta}|}f(u)du}. \tag{4.4}$$

We note that the treatment of multiple roots of the equation $\tau(u,\theta) = t$ in the present context is similar to the treatment in Michael et al. (1976) in connection with generation of random variates from transformations with multiple roots. Formulas (4.2) and (4.3) can in fact together be considered as a multivariate generalization of equation 3 in Michael et al. (1976) [see also Taraldsen and Lindqvist (2005)].

The following two examples illustrate the use of Algorithm 3 and equation (4.4). In the first example $\Gamma(u,t)$ contains at most one value of $\theta$, but may be empty. In the second example we may have an arbitrary number of elements in $\Gamma(u,t)$.

EXAMPLE 3 (*Type I censored exponential lifetimes*). Let $n$ units with potential lifetimes $Y_1, Y_2, \ldots, Y_n$ be observed from time 0, but assume that the observation of the $i$th unit is censored at a given time $c_i > 0$ $(i = 1, \ldots, n)$. This means that we observe only $X_i = \min(Y_i, c_i)$. In the reliability terminology this is called Type I censoring. Suppose $Y_1, \ldots, Y_n$ are i.i.d. with distribution $\mathrm{Exp}(\theta)$. Then the likelihood of $X_1, \ldots, X_n$ can be written $\theta^R \exp(-\theta S)$ where $R = \sum_i I(X_i < c_i)$ is the number of noncensored observations and $S = \sum_i X_i$ is the sum of all observations. Here $I(A)$ is the indicator function of the event $A$. Now $T = (R, S)$ is sufficient for $\theta$, but note that a two-dimensional statistic is here sufficient for a one-dimensional parameter.

It should be remarked that the potential censoring times $c_i$ are assumed known also for the units where $X_i < c_i$. For example this is the case if $n$ machines, or patients in a medical study, are observed from possibly different starting points in time, and until a common terminal point. Let $c_1, \ldots, c_n$ be fixed, known numbers in the following.

As in Example 1, let $U = (U_1, \ldots, U_n)$ be a vector of $n$ i.i.d. $\mathrm{Exp}(1)$ variables. We then simulate $X$ for a given value of $\theta$ by means of $\chi(U,\theta) = (\eta_1(U_1,\theta), \ldots, \eta_n(U_n,\theta))$ where

$$\eta_i(u_i,\theta) = \min(u_i/\theta, c_i),\ i = 1, \ldots, n.$$

Thus $T = (R, S)$ is simulated by $\tau(U,\theta) = (\gamma(U,\theta), \psi(U,\theta))$ where $\gamma(U,\theta) = \sum_i I(U_i/\theta < c_i)$ and $\psi(U,\theta) = \sum_i \eta_i(U_i,\theta)$.

We now show how to find the functions $W_t(u)$ and $Z_t(u)$ needed in (4.1). First we show that the equation $\tau(u,\theta) = t$ has at most one solution for $\theta$ for fixed $u, t$, but may have none. Let the observed value of the sufficient statistic, $t = (r, s)$, be fixed with $0 < r \le n$, $0 < s < \sum_i c_i$. Then consider the equations $\gamma(u,\theta) = r$, $\psi(u,\theta) = s$ for a given $u$. Since $\psi(u,\theta)$ is strictly decreasing in $\theta$, from $\sum_i c_i$ to 0, there is a unique $\hat{\theta}$ which satisfies $\psi(u,\hat{\theta}) =$

$s$. However, this $\hat{\theta}$ may not solve $\gamma(u,\theta) = r$. In the cases where indeed $\gamma(u,\hat{\theta}) = r$, put $K(u,t) = 1$ and put $K(u,t) = 0$ otherwise. If $K(u,t) = 1$ then define $I(u,t) = \{i_1, \ldots, i_r\}$ to be the set of indices $i$ for which $u_i/\hat{\theta} < c_i$. With this notation we can express the solution $\hat{\theta}$ when $K(u,t) = 1$ as

$$\hat{\theta}(u,t) = \frac{\sum_{i\in I(u,t)} u_i}{s - \sum_{i\notin I(u,t)} c_i}.$$

Next, choose a density $\pi(\theta)$ for $\theta > 0$, for example $\pi(\theta) = 1/\theta$ in accordance with Example 2. We then find the density $W_t(u) \equiv W_{(r,s)}(u)$ of $\tau(u,\Theta)$ to be

$$\begin{aligned} W_t(u)\, ds &= \pi\{\theta : \gamma(u,\theta) = r, s \le \psi(u,\theta) \le s + ds\} \\ &= \begin{cases} 0 & \text{if } K(u,t) = 0 \\ \hat{\theta}(u,t)^2 \pi(\hat{\theta}(u,t)) ds / \sum_{i\in I(u,t)} u_i & \text{if } K(u,t) = 1. \end{cases} \end{aligned}$$

Further, $Z_t(u)$ is the conditional expectation of $\phi\{\chi(u,\Theta)\}$ given $\tau(u,\Theta) = t$. This is easily found since the conditional distribution of $\Theta$ given $\tau(u,\Theta) = t$ is a one-point mass at $\hat{\theta}(u,t)$ if $K(u,t) = 1$ and can be arbitrarily chosen otherwise. Formula (4.4) therefore gives

$$E\{\phi(X)|T = t\} = \frac{E\{K(U,t)\phi[\chi\{U,\hat{\theta}(U,t)\}]W_t(U)\}}{E\{K(U,t)W_t(U)\}}.$$

The choice $\pi(\theta) = 1/\theta$ yields the simple weight function

$$W_t(u) = (s - \sum_{i\notin I(u,t)} c_i)^{-1}, \tag{4.5}$$

valid when $K(u,t) = 1$.

An important special case is when the $c_i$ are all equal. In this case $W_t(u)$ in (4.5) does not depend on $u$ and we can sample directly from the conditional distribution of $X$ given $T = t$ for fixed $t$ by drawing $u$ until $K(u,t) = 1$ and then using $X_t(u) = \chi\{u,\hat{\theta}(u,t)\}$.

EXAMPLE 4 (*Inverse Gaussian distributed lifetimes*). Let $X = (X_1, \ldots, X_n)$ be a sample from the inverse Gaussian distribution with density

$$f(x;\mu,\phi) = \sqrt{\frac{\mu\phi}{2\pi x^3}} \exp\left(-\frac{\mu\phi}{2x} - \frac{\phi x}{2\mu} + \phi\right), \quad x > 0 \tag{4.6}$$

[Seshadri (1999), p. 2] where $\mu, \phi > 0$ are parameters. Denote this distribution by IG$(\mu,\phi)$. Note that a more common parametrization uses $\mu$ together

with $\lambda = \mu\phi$, but the one used in (4.6) is more convenient for our purposes as will become clear below. Doksum and Høyland (1992) considered models for accelerated life testing experiments which were based on the inverse Gaussian distribution. In the present example we shall consider conditional sampling given the sufficient statistic, which may have several interesting applications in this connection.

A sufficient statistic is given by [Seshadri (1999), p. 7]

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

Since $\mu$ is a scale parameter in (4.6) we can simulate from $\mathrm{IG}(\mu, \phi)$ by first simulating from $\mathrm{IG}(1, \phi)$ and then multiplying the result by $\mu$. We shall use the method suggested by Michael et al. (1976) which seems to be easier than ordinary inversion since there is no closed form expression for the inverse cumulative distribution function.

Let $U_i$ be $\mathrm{Un}[0, 1]$ and $V_i$ be $\chi^2_1$ for $i = 1, \ldots, n$, where all variables are independent. Here $\chi^2_1$ means the chi-square distribution with 1 degree of freedom. Let

$$\begin{aligned} W_i &= 1 - (2\phi)^{-1} \left( \sqrt{V_i^2 + 4\phi V_i} - V_i \right), \\ Z_i &= (1 + W_i)^{-1}. \end{aligned}$$

Then [Michael et al. (1976)] the variables

$$\eta(U_i, V_i, \phi) = I(U_i \le Z_i)\, W_i + I(U_i > Z_i)\, (1/W_i)$$

are distributed as $\mathrm{IG}(1, \phi)$ and hence

$$\chi(U, V, \mu, \phi) = (\mu\eta(U_1, V_1, \phi), \ldots, \mu\eta(U_n, V_n, \phi))$$

is a simulated sample of size $n$ from $\mathrm{IG}(\mu, \phi)$. Here $U = (U_1, \ldots, U_n)$, $V = (V_1, \ldots, V_n)$. Moreover, we simulate $T = (T_1, T_2)$ by

$$\begin{aligned} \tau(U, V, \mu, \phi) &= (\tau_1(U, V, \mu, \phi), \tau_2(U, V, \mu, \phi)) \\ &= \left( \sum_{i=1}^{n} \mu\eta(U_i, V_i, \phi), \sum_{i=1}^{n} (1/\mu)(1/\eta(U_i, V_i, \phi)) \right). \end{aligned}$$

In order to compute conditional expectations or to sample from the conditional distribution of $X$ given $T = t$ we need to solve the equations

$\tau(u, v, \mu, \phi) = t = (t_1, t_2)$ with respect to $\mu$ and $\phi$. This can be done by first solving the equation

$$\tau_1(u, v, \mu, \phi) \cdot \tau_2(u, v, \mu, \phi) = t_1 t_2, \tag{4.7}$$

which is free of $\mu$. It turns out that the solution for $\phi$ is not necessarily unique. In fact, the number of roots is finite but may vary with $(u, v)$. However, for each root found for $\phi$ we can easily solve for $\mu$ using $\tau_1(u, v, \mu, \phi) = t_1$. It should be noted that the functions $\eta(u_i, v_i, \phi)$ are discontinuous in $\phi$ due to the indicator functions involved in their definition. However, the discontinuities are easy to calculate, and the functions behave smoothly as functions of $\phi$ between them. This simplifies the solution of the equation (4.7) and enables rather straightforward computation of $W_t(u)$ in (4.2). A possible choice of the density $\pi$ is to put $\pi(\mu, \phi) = 1/(\mu\phi)$ since Jeffreys' priors for, respectively, known $\phi$ and known $\mu$ are $1/\mu$ and $1/\phi$ (see Section 6.3 for the use of Jeffreys' priors in the present context). The desired simulations and computations can thus be performed by the methods of the present section.

As mentioned in the introduction, Cheng (1984) presented a method for simulation of conditional distributions in the case of inverse Gaussian distributed samples. His method is based on a subtle decomposition of chi-squared random variates and appears to be somewhat simpler than the method presented here.

4.2. *The discrete case.* Suppose that both $X$ and $T$ have discrete distributions, while the parameter space is a subset of the $k$-dimensional Euclidean space. In this case the sets $\Gamma(u, t)$ are usually sets with positive Lebesgue measure. These may in many cases be found explicitly, so that $W_t(u) = Pr\{\tau(u, \Theta) = t\}$ can be computed directly. In some instances, however, the set $\Gamma(u, t)$ is difficult to find. For such cases Engen and Lillegård (1997) suggest replacing $\pi$ by a discrete measure, such as the counting measure on a grid of points in the parameter space.

A thorough treatment of the discrete case is given in LT (2005), including an example with logistic regression.

**5. On the distribution of $\hat{\theta}(U, t)$.** Consider again the case when $\tau(u, \theta) = t$ has the unique solution $\hat{\theta}(u, t)$. For computational reasons it may be desirable to have some knowledge of the probability distribution of $\hat{\theta}(U, t)$ as a function of $U$.

Note first that for the case when $\theta$ is one-dimensional and $T$ is stochastically increasing in $\theta$, Lillegård and Engen (1999) used the variates $\hat{\theta}(U, t)$

to derive exact confidence intervals for $\theta$. More precisely they showed that one obtains an exact $(1-2k/(m+1))$-confidence interval for $\theta$ by sampling $m+1$ values of $\hat{\theta}(U,t)$ and then using the interval from the $k$th smallest to the $k$th largest of them. They called this method conditional parametric bootstrapping. Their result can be rephrased to say that the interval between the $\alpha/2$ and $1-\alpha/2$ percentiles of the distribution of $\hat{\theta}(U,t)$ is an exact $1-\alpha$ confidence interval for $\theta$. In fact, the distribution of $\hat{\theta}(U,t)$ is in this case a fiducial distribution in the sense of Fisher [Wilks (1962), p. 370]. This suggests that, under given standard conditions, the distribution of $\hat{\theta}(U,t)$ should at least asymptotically be comparable to that of a decent estimator of $\theta$, for example the maximum likelihood estimator.

This turns in fact out to be true under reasonable conditions. A rough argument for the extended case where $\theta$ and $T$ are $k$-dimensional can be given as follows. Suppose that we have $U=(U_1,U_2,\ldots,U_n)$ where we shall consider the case where $n\to\infty$. Furthermore, assume that the parametrization is such that $\theta=E\{T\}=E\{\tau(U,\theta)\}$. Lehmann and Casella (1998, p. 116) calls this the mean value parametrization. In this case $T$ is itself an unbiased estimator of $\theta$, and is the maximum likelihood estimator if the underlying model is an exponential family [Lehmann and Casella (1998), p. 470]. Our basic assumption for the following derivation is that

$$n^{1/2}(\tau(U,\theta)-\theta)\xrightarrow{d}N_k(0,\Sigma(\theta))$$

as $n\to\infty$ for some positive definite matrix $\Sigma(\theta)$. This is satisfied in the exponential family case, where $\Sigma(\theta)$ is the inverse Fisher information matrix.

Now we consider a fixed value of $t$ and define $\hat{\theta}(U,t)$ to be the unique solution of $\tau(U,\theta)=t$. Assume furthermore that we can show that $\hat{\theta}(U,t)\to t$ in probability as $n\to\infty$. In this case, for any $U$,

$$\begin{aligned} t &= \tau(U,\hat{\theta}(U,t)) \\ &= \tau(U,t)+\partial_\theta\tau(U,\theta)|_{\theta=\tilde{\theta}}(\hat{\theta}(U,t)-t), \end{aligned}$$

where $\tilde{\theta}$ is between $\hat{\theta}(U,t)$ and $t$ in the sense that each component of $\tilde{\theta}$ is a convex combination of the corresponding components of $\hat{\theta}(U,t)$ and $t$.

Hence

$$n^{1/2}(\hat{\theta}(U,t)-t)=(\partial_\theta\tau(U,\theta)|_{\theta=\tilde{\theta}})^{-1}n^{1/2}(t-\tau(U,t))$$

and provided $\partial_\theta\tau(U,\theta)|_{\theta=\tilde{\theta}}\xrightarrow{p}I$ (where $I$ is the identity matrix) we have

$$n^{1/2}(\hat{\theta}(U,t)-t)\to N_k(0,\Sigma(t)) \tag{5.1}$$

The requirement that $\partial_\theta \tau(U,\theta)|_{\theta=\tilde{\theta}} \xrightarrow{p} I$ is typical in asymptotic results related to estimating equations, see for example Welsh (1996, Section 4.2.4) and Sørensen (1999) for sufficient conditions. The reason for the limit $I$ above is that $E\{\tau(U,\theta)\} = \theta$. We will not pursue this further here, since the methods we derive are meant for use in non-asymptotic inference.

The conclusion is that for a large class of models $\hat{\theta}(U,t)$ has the same asymptotic distribution as $T$ under the parameter value $\theta = t$. Thus in multiparameter exponential models we conclude that $\hat{\theta}(U,t)$ (under given conditions) has the same asymptotic distribution as the maximum likelihood estimator for $\theta$. Note that by the invariance property of the maximum likelihood estimator and of $\hat{\theta}(U,t)$ (see Section 6.3) this holds under any parametrization.

Finally we can reinterpret our result (5.1) to say that conditionally on $T$, $n^{1/2}\{\hat{\theta}(U,T) - T\}$ has the same limiting distribution as $n^{1/2}(T-\theta)$. This result is analogous to asymptotic results for bootstrapping (Bickel and Freedman, 1981), in which the $\hat{\theta}(U,T)$ are replaced by bootstrapped statistics.

## 6. Computational aspects.

6.1. *Monte Carlo computation of conditional expectations.* A basic idea of our approach is that expectations of functions of $U$, such as (3.4) and (4.4), can be evaluated by Monte Carlo simulation. Basically, we can estimate $E\{h(U)\}$ by $(1/m)\sum_{i=1}^m h(u_i)$ where $u_1,\ldots,u_m$ is a computer generated pseudo sample from the distribution of $U$. The literature on Monte Carlo simulation [for example Ripley (1987)] contains various methods for improving on this naive approach of estimating $E\{h(U)\}$.

6.2. *Choice of simulation method for* $(X,T)$. Our approach relies on the functions $(\chi(U,\theta),\tau(U,\theta))$ chosen for simulation of $(X,T)$ in the original model. There is usually no unique way of selecting a simulation method. In the simulation of inverse Gaussian variables in Example 4 it would be possible, for example, to use ordinary inversion based on the cumulative distribution function, or even to use simulation of Wiener processes as described in Chhikara and Folks (1989). Each simulation scheme would give a different solution technique for handling the conditional distributions.

6.3. *Choice of the density* $\pi$. *Jeffreys' prior.* For a given setup in terms of $(\chi(U,\theta),\tau(U,\theta))$ we need to specify a density $\pi(\theta)$, except when conditions for using Algorithm 1 are fulfilled. In practice the effectiveness of an algorithm is connected to variation in the $W_t(u)$ which should be small or at

best absent. For example, in order to minimize this variation in the case of formula (3.4), the density $\pi(\theta)$ should be chosen so that $\pi\{\hat{\theta}(u,t)\}$ is similar to $|\det\partial_\theta \tau(u,\theta)|_{\theta=\hat{\theta}(u,t)}$.

Under the pivotal condition (Section 2) we may always choose $\pi$ so that $W_t(u)$ does not depend on $u$. As a simple illustration, consider the simple pivotal case where $\theta$ is one-dimensional and $\tau(u,\theta) = r(u)\theta$. This means that $T/\theta$ is a pivotal quantity. Assume that the parametrization is such that $E\{r(U)\} = 1$ so that we have the mean value parametrization. In this case $\hat{\theta}(u,t) = t/r(u)$ so $\partial_t\hat{\theta}(u,t) = 1/r(u) = \hat{\theta}(u,t)/t$. Hence we get $W_t(u)$ in (3.3) constant in $u$ by choosing $\pi(\theta) = 1/\theta$. Assuming that $T$ is the maximum likelihood estimator of $\theta$ then under regularity conditions the Fisher-information is given by $1/\mathrm{Var}\{\tau(U,\theta)\} \propto 1/\theta^2$, so $1/\theta$ is Jeffreys' prior here. As another illustration it is shown in LT (2005) that in the case where $X$ is a sample from $N(\mu,\sigma)$ we obtain constant $W_t(u)$ by choosing $\pi(\mu,\sigma) = 1/\sigma$, which is the standard improper, noninformative prior for this case.

In fact there are reasons to choose improper, noninformative priors, such as Jeffreys' prior, also in general for the distribution $\pi$. Consider in particular a one-to-one reparametrization from a $k$-dimensional parameter $\theta$ to the $k$-dimensional $\xi$ defined by $\theta = h(\xi)$. We then define $\tau_h(u,\xi) = \tau(u,h(\xi))$ from which it follows that the $\hat{\xi}(u,t)$ which solves the equation $\tau_h(u,\xi) = t$ satisfies $\hat{\theta}(u,t) = h\{\hat{\xi}(u,t)\}$. Now let $J$ be the $k \times k$-matrix with elements $J_{ij} = \partial h_i(\xi)/\partial\xi_j$. Then we can write (3.3) as

$$W_t(u) = \pi[h\{\hat{\xi}(u,t)\}]\ |\det J\ \det\partial_t\hat{\xi}(u,t)|.$$

This shows that if we change the parametrization, then the weights $W_t(u)$ are unchanged provided we change $\pi$ by the ordinary change of variable formula for densities. Thus a consistent principle for choosing $\pi$ should have this property of invariance under reparametrizations. It is well known that Jeffreys' prior (Jeffreys, 1946) has this property, and there seems to be reasons why in fact Jeffreys' prior distribution is a reasonable candidate for general use.

6.4. *Direct sampling from the conditional distributions.* Algorithm 1 describes how to sample from the conditional distribution of $X$ given $T = t$ under special conditions. Sampling from the conditional distribution using Algorithms 2 or 3 may, however, in general be difficult since the normalizing constant of the density $W_t(u)f(u)$ may not be easily available. Rejection sampling can be used if we are able to bound $W_t(u)$ from above. Looking at (3.3) we find that a possible way of doing this is to seek a positive function

$\rho(\theta)$ such that for all $u$ we have

$$|\det \partial_\theta \tau(u,\theta)|_{\theta=\hat{\theta}(u,t)} \geq \rho\{\hat{\theta}(u,t)\}.$$

In this case we can put $\pi(\theta) = \rho(\theta)$ to get $W_t(u) \leq 1$ for all $u$. Then we may simulate $V$ in Step 1 by first drawing a $U = u$ and then accepting it with probability $W_t(u)$.

A possible method for sampling without bounding $W_t(u)$ is by means of the SIR-algorithm of Rubin (1988). In the case of Algorithm 2 this method can be described as follows:

First sample $u_1, \ldots, u_m$ independently from the density $f(u)$. Then define $w_i = W_t(u_i)$ for $i = 1, \ldots, m$ and let $F_m$ denote the discrete probability measure which assigns probability $w_i / \sum_{i'=1}^m w_{i'}$ to $u_i$. Then $F_m$ converges to the desired conditional distribution as $m \to \infty$. Hence for $m$ large enough we can obtain independent samples in Step 1 of Algorithms 2 and 3 by sampling from $F_m$.

Samples in Step 1 of Algorithms 2 and 3 can also be obtained by using the independence sampler based on the Metropolis-Hastings algorithm [Tierney (1994)], but this leads to dependent samples from the conditional distribution.

## References


Aitchison, J. (1963). Inverse distributions and independent gamma distributed products of random variables. *Biometrika* **50** 505-508.

Bahadur, R. R. and Bickel, P. J. (1968). Substitution in conditional expectation. *Ann. Math. Statist.* **39** 377-378.

Bickel, P. J. and Doksum, K. A. (1969). Tests for monotone failure rate based on normalized spacings. *Ann. Math. Statist.* **40** 1216-1235.

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics, 2nd Edition, Vol I.* Prentice-Hall, Upper Saddle River, NJ.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196-1217.

Cheng, R. C. H. (1984). Generation of inverse Gaussian variates with given sample mean and dispersion. *Appl. Stat.* **33** 309-316.

Cheng, R. C. H. (1985). Generation of multivariate normal saamples with given sample mean and covariance matrix. *J. Statist. Comput. Simul.* **21** 39-49.

Chhikara, R.S. and Folks, J. L. (1989). *The inverse Gaussian distribution. Theory, Methodology, and Applications.* Marcel Dekker, New York.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** 363-397.

Doksum, K. A. and Høyland, A. (1992). Models for variable-stress accelerated life testing experiments based on Wiener processes and the Inverse Gaussian distribution. *Technometrics* **34** 74-82.

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

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A* **186** 453-461.

LANGSRUD, Ø. (2005). Rotation tests. *Statistics and Computing* **15** 53-60.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation.* Springer-Verlag, New York.

LILLEGÅRD, M. and ENGEN, S. (1999). Exact confidence intervals generated by conditional parametric bootstrapping. *J. Appl. Stat.* **26** 447-459.

LINDQVIST, B. H. and TARALDSEN, G. (2001). Monte Carlo conditioning on a sufficient statistic. Statistics No. 9/2001, Dep. of Math. Sciences, Norwegian University of Science and Technology, Trondheim. Available as *http://www.math.ntnu.no/preprint/statistics/2001/S9-2001.ps.*

LINDQVIST, B. H. and TARALDSEN, G. (2005). Monte Carlo conditioning on a sufficient statistic. *Biometrika* **92** 451-464.

LINDQVIST, B. H., TARALDSEN, G., LILLEGÅRD, M., and ENGEN, S. (2003). A counterexample to a claim about stochastic simulations. *Biometrika* **90** 489-490.

MICHAEL, J. R., SCHUCANY, W. R. and HAAS, R. W. (1976). Generating random variates using transformations with multiple roots. *Am. Stat.* **30** 88-90.

RIPLEY, B. (1987). *Stochastic Simulation.* Wiley, New York.

RUBIN, D.B. (1988). Using the SIR algorithm to simulate posterior distributions (with discussion). In *Bayesian Statistics, Vol. 3* (Bernardo et al., eds.) 395-402. Oxford University Press, Oxford.

SESHADRI, V. (1999). *The Inverse Gaussian Distribution. Statistical Theory and Applications.* Lecture Notes in Statistics 137. Springer, New York.

SØRENSEN, M. (1999). On asymptotics of estimating functions. *Brazilian J. Prob. Statist.* **13** 111 - 136.

TARALDSEN, G. AND LINDQVIST, B. H. (2005). The multiple roots simulation algorithm, the inverse Gaussian distribution, and the sufficient conditional Monte Carlo method. Statistics No. 4/2005, Dep. of Math. Sciences, Norwegian University of Science and Technology, Trondheim. Available as *http://www.math.ntnu.no/preprint/statistics/2005/S4-2005.pdf.*

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701-1762.

TROTTER and TUKEY (1956). Conditional Monte Carlo for normal samples. In *Symposium on Monte Carlo Methods* (H. A. Meyer, ed.) 64-79. Wiley, New York.

WELSH, A. H. (1996). *Aspects of Statistical Inference.* Wiley, New York.

WILKS, S. S. (1962). *Mathematical Statistics.* Wiley, New York.

DEPARTMENT OF MATHEMATICAL SCIENCES
NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY
NO-7491 TRONDHEIM, NORWAY
E-MAIL: bo@math.ntnu.no

SINTEF ICT
O. S. BRAGSTADS PLASS
NO-7465 TRONDHEIM, NORWAY
E-MAIL: gunnar.taraldsen@sintef.no