Statistics and Computing (2023) 33:22
https://doi.org/10.1007/s11222-022-10188-x

ORIGINAL PAPER

Direct sampling with a step function

Andrew M. Raim[1]

Received: 17 March 2022 / Accepted: 27 November 2022 / Published online: 21 December 2022

Abstract

The direct sampling method proposed by Walker et al. (JCGS 2011) can generate draws from weighted distributions possibly having intractable normalizing constants. The method may be of interest as a tool in situations which require drawing from an unfamiliar distribution. However, the original algorithm can have difficulty producing draws in some situations. The present work restricts attention to a univariate setting where the weight function and base distribution of the weighted target density meet certain criteria. Here, a variant of the direct sampler is proposed which uses a step function to approximate the density of a particular augmented random variable on which the method is based. Knots for the step function can be placed strategically to ensure the approximation is close to the underlying density. Variates may then be generated reliably while largely avoiding the need for manual tuning or rejections. A rejection sampler based on the step function allows exact draws to be generated from the target with lower rejection probability in exchange for increased computation. Several applications of the proposed sampler illustrate the method: generating draws from the Conway-Maxwell Poisson distribution, a Gibbs sampler which draws the dependence parameter in a random effects model with conditional autoregression structure, and a Gibbs sampler which draws the degrees-of-freedom parameter in a regression with t-distributed errors.

Keywords Weighted distribution · Intractable normalizing constant · Inverse CDF sampling · Rejection sampling · Gibbs sampling

1 Introduction

This paper revisits the direct sampling method proposed by Walker et al. (2011). Consider drawing a random variable X with support $\Omega \subseteq \mathbb{R}$ whose density takes the form

$$f(x) = w(x)g(x)/\psi, \quad x \in \Omega, \quad \psi = \int_{\Omega} w(x)g(x)d\nu(x), \tag{1}$$

where $\nu(\cdot)$ is a dominating measure. The distribution of X may be discrete, continuous, or continuous with point masses. Density f can be recognized as a weighted distribution (e.g. Patil and Rao 1978) with a weight function $w : \mathbb{R} \to [0, \infty)$ which adjusts the base density g in some prescribed way. Direct sampling augments a random variable U so that the joint distribution of $[X, U]$ is easier to draw than X itself. Let $\mathrm{I}(\cdot)$ be the indicator function and suppose $c = \sup_{x \in \Omega} w(x)$ is finite. Assume that $[U \mid X = x] \sim \text{Uniform}(0, w(x)/c)$, so that

$$f(u \mid x) = \frac{c}{w(x)} \mathrm{I}(0 < u < w(x)/c).$$

Define the event $A_u = \{x \in \Omega : w(x) > uc\}$. The joint density of $[X, U]$ is then

$$f(x, u) = \frac{c}{\psi} g(x)\, \mathrm{I}(x \in A_u). \tag{2}$$

From (2), the marginal density of U may be obtained as

$$p(u) = \frac{c}{\psi} \mathrm{P}(A_u), \quad u \in [0, 1], \tag{3}$$

Disclaimer: This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the author and not those of the U.S. Census Bureau.

✉ Andrew M. Raim
andrew.raim@census.gov

[1] Center for Statistical Research and Methodology,
U.S. Census Bureau, Washington, DC 20233, U.S.A.

with $\mathrm{P}(A_u) = \int \mathrm{I}(x \in A_u) g(x) d\nu(x)$. The distribution of $[X \mid U = u]$ is then

$$f(x \mid u) = \frac{g(x)}{\mathrm{P}(A_u)} \mathrm{I}(x \in A_u). \tag{4}$$

Now U is bounded in $[0, 1]$, with $A_s \supseteq A_u$ if $s \leq u$ so that $\mathrm{P}(A_u)$ is monotonically nonincreasing in u. Evaluated at the endpoints $\{0, 1\}$, A_0 is equivalent to the support of w with $\mathrm{P}(A_0) = \int_\Omega \mathrm{I}(w(x) > 0) g(x) d\nu(x)$ and A_1 is an empty set with $\mathrm{P}(A_1) = 0$.

A draw from $f(x)$ may be approximately obtained by drawing U from $p(u)$ then X from $f(x \mid u)$ in the following way. For a predefined positive integer N, compute

$$q(k/N) = \frac{\mathrm{P}(A_{k/N})}{\sum_{\ell=0}^{N} \mathrm{P}(A_{\ell/N})}, \quad k = 0, 1, \ldots, N. \tag{5}$$

Sample discrete random variable K from the values $0, 1, \ldots, N$ with respective probabilities $q(0/N), \ldots, q(N/N)$, then draw from $[U \mid K = k] \sim \text{Beta}(k + 1, N - k + 1)$. The marginal density of U is then proportional to

$$\begin{aligned} &\sum_{k=0}^{N} \frac{u^k (1-u)^{N-k}}{B(k+1, N-k+1)} q(k/N) \\ &\propto \sum_{k=0}^{N} \binom{N}{k} u^k (1-u)^{N-k} q(k/N), \quad u \in (0, 1) \end{aligned} \tag{6}$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ is the beta function. Expression (6) is an approximation of $p(u)$ by Bernstein polynomials (e.g. Rivlin 1981). A variate x from the truncated distribution (4) may be obtained by repeating draws of candidate x^* from g, which is straightforward in many applications, until $x^* \in A_u$ where x is taken to be x^*. This algorithm was described by Walker et al. (2011) as a basic implementation of their direct sampling approach.

The sampler described thus far may encounter challenges in practice which prevent it from successfully drawing from the target distribution. First, the basic rejection sampling method described to draw from (4) may require a very large number of candidates when the set A_u has small probability under $g(x)$. Second, the function $\mathrm{P}(A_u)$ may take large sudden steps not efficiently captured by polynomial approximation. Third, the support of $\mathrm{P}(A_u)$ may be concentrated on an interval $[u_L, u_H]$ with u_H a very small positive number. For example, the second and third issues are seen in the bottom row of Figure 2.

It is possible to focus the Bernstein approximation (6) to the interval $[u_L, u_H]$ or consider other functional bases from the literature. However, we propose the use of step functions which are relatively simple with low computational burden. Expressions for the density, cumulative distribution function (CDF), and quantile function are available and exact draws may be taken directly via the quantile function. Through appropriate placement of knot points, a step function can effectively capture any jumps encountered in $p(u)$. Simple bounds on the accuracy of the approximation can be obtained in our setting, and such bounds can be improved by placing additional knot points until a desired tolerance is achieved. A step function can serve as an envelope in rejection sampling if exact draws from $f(x)$ are required. In addition to assuming univariate X, we restrict ourselves to weight functions w where A_u is an interval for each $u \in [0, 1]$. Ideally, endpoints of A_u and the CDF and quantile function of base distribution g are readily computed.

Walker et al. (2011) consider the direct sampling approach as an alternative to Markov chain Monte Carlo (MCMC) in Bayesian computing. Posterior distributions of interest are not available in a closed-form in many real world applications; in lieu of this, MCMC methods are commonly used to produce chains which are utilized as draws from the posterior. When it can be applied, direct sampling is appealing in this setting because it removes the need to assess stationarity and adequate mixing in chains, produces draws which are independent rather than serially correlated, and is not subject to the same serial computing limitations that are inherent within a chain. Braun and Damien (2011) and Braun and Damien (2016) extend the methodology to include a proposal distribution reminiscent of rejection sampling, and replace Bernstein polynomials with another approximation which is constructed by sampling. Direct sampling does not yet appear to be widely adopted in the literature; however, it is interesting to consider in the context of MCMC methods and rejection sampling methods with adaptation.

Metropolis-Hastings (Metropolis et al. 1953; Hastings 1970) and Gibbs sampling (Geman and Geman 1984) are two fundamental MCMC methods which are still heavily used in practice. Implementation of a Gibbs sampler depends upon the ability to draw from conditionals of a multivariate target distribution; this may be challenging when unfamiliar forms are encountered. Metropolis-Hastings can also be applied to multivariate targets or used within a Gibbs sampler; its performance relies upon the selection of a proposal distribution and may require extensive tuning to produce usable chains. Data augmentation (Tanner and Wong 1987) may help to yield familiar conditionals at the cost of having more variables in the chain. Slice sampling (Neal 2003) is a particular data augmentation method with similarities to the direct sampling approach; see Remark 1. Hamiltonian Monte Carlo methods (Duane et al. 1987) have become popular for their ability to traverse multivariate target distributions and produce MCMC chains with good mixing. Software platforms such as Stan (Carpenter et al. 2017) and PyMC (Salvatier et al. 2016) offer the attractive prospect of sampling from a target—especially in Bayesian modeling—by simply declaring it. Sampling methodology can be largely automated using algorithms such as the No-U-Turn variant of Hamiltonian Monte Carlo (Hoffman and Gelman 2014). In some settings, theoretical tools may be used to determine whether a Markov chain has adequately mixed to the target distribution or to use Markov chains to obtain exact draws from the target distribution (Levin and Peres 2017).

The term "direct sampling" is often used informally in the literature to refer to exact sampling using straightforward steps based on the properties of the target distribution. Methods of transformation—from one distribution to another—have been established to generate variates exactly from many standard distributions. The inverse CDF transformation (Devroye 1986, e.g. Section II) may be used to transform Uniform draws to a univariate target distribution of interest, provided that the quantile function of the target can be computed. Rejection sampling is a fundamental computational approach to generate exact draws; candidate draws are generated from a proposal distribution and accepted or rejected using a simple criterion involving the target and proposal. An appropriate proposal must be supplied so that the rejection probability is low enough for the algorithm to be of practical use. Adaptive Rejection Sampling (ARS) is a variant of rejection sampling that sequentially constructs a proposal distribution based on rejected draws (Gilks and Wild 1992). The probability of rejection becomes small as more candidate draws are rejected; however, the original ARS algorithm is specific to log-concave distributions. Martino et al. (2018, Section 3.6) provide background on step functions in the context of rejection sampling and attribute the idea to Ahrens (1993, 1995). Adaptive Rejection Metropolis Sampling (ARMS) drops the log-concavity requirement of ARS but is an MCMC method rather than an exact sampling method (Gilks et al. 1995). Evans and Swartz (1998) extend ARS to densities which meet a T-concavity criterion. Görür and Teh (2011) extend ARS to target log-densities which can be decomposed into sums of concave and convex functions; Martino and Míguez (2011) further support a convex transformation for each function in the summation. Independent Doubly Adaptive Rejection Metropolis Sampling (IA^{2}RMS) extends ARMS to ensure that the proposal approaches the target and variate generation becomes similar to ARS as the algorithm iterates (Martino et al. 2015). Erraqabi et al. (2016) and Achddou et al. (2019) consider adaptive rejection sampling using nonparametric methods to construct proposals for ARS—using kernel and nearest-neighbor estimation respectively—and establish theoretical guarantees on the rejection probability. Holden et al. (2009) study adaptive Metropolis-Hastings (without rejection sampling steps); under certain conditions, the history of proposed draws may be used to adapt the proposal so that it converges to the target while the stationary distribution of the chain is preserved. Martino et al. (2018) develop a computational framework for adaptive Metropolis-Hastings with univariate targets.

The rejection method proposed in this paper is an exact sampling technique—like ARS—suited to univariate weighted distributions. Such situations may be encountered as a step in multivariate sampling such as within a Gibbs sampler where a variety of techniques are used to handle the conditionals. An appealing aspect of the direct sampling approach, which is a notable departure from ARS and related work mentioned earlier, is the focus on approximating and adapting $p(u)$ rather than the target distribution directly. Because $p(u)$ is monotonically non-increasing for any choice of target density, the approximation problem is more limited in scope. This also facilitates straightforward theoretical bounds for accuracy and knot selection which may be guided by the bounds.

The remainder of the paper proceeds as follows. Section 2 discusses generating draws from (4) in this setting without rejections. Section 3 presents use of the step function in direct sampling. Section 4 considers three illustrative applications using this formulation of the direct sampler: drawing from the Conway-Maxwell Poisson distribution, a Gibbs sampler for a conditional autoregression random effects model including inference on the dependence parameter, and a Gibbs sampler for a regression model with errors following a Student's t-distribution including inference on the degrees of freedom. Finally, Section 5 concludes the paper. Supporting code is provided as an electronic supplement, including materials to replicate the examples, implemented in both pure R (R Core Team 2022) and with integrated C++ via the Rcpp framework (Eddelbuettel 2013).

Remark 1 It is clarifying to briefly review slice sampling (Neal 2003) and compare it with direct sampling. Consider drawing X from target distribution $f(x) = \bar{f}(x)/\psi$ where $\bar{f}(x)$ is an unnormalized density, $x \in \Omega$, and $\psi = \int_{\Omega} \bar{f}(x) d\nu(x)$. Slice sampling augments a random variable U so that the joint density of $[X, U]$ is $f(x, u) = \mathrm{I}(0 < u \le \bar{f}(x))/\psi$. The joint density yields conditionals $f(u \mid x) \propto \mathrm{I}(0 < u \le \bar{f}(x))$ and $f(x \mid u) \propto \mathrm{I}(u \le \bar{f}(x))$. Sampling from $f(x, u)$ may be carried out via Gibbs sampler using these conditionals; however, it may be nontrivial to draw uniformly from the set $S_u = \{x \in \Omega : \bar{f}(x) \ge u\}$. Both slice sampling and direct sampling are based on an augmented uniform random variable and circumvent the need to compute ψ. Both methods depend upon "slice" sets, with S_u based on $\bar{f}$ and A_u based on the weight function only. Slice sampling does not assume a weighted density, but produces an MCMC chain while direct sampling yields independent draws.

2 Drawing from the truncated base distribution

We first consider efficiently drawing from (4). Suppose density g is associated with CDF and quantile functions

$$G(x) = \int_{-\infty}^{x} g(s) d\nu(s), \quad G^{-}(\varphi) = \inf\{x \in \Omega : G(x) \geq \varphi\},$$

respectively. With A_u assumed to be an interval $(x_1(u), x_2(u))$, whose endpoints are identified by the roots of the equation $w(x) = cu$, (4) represents the base distribution g truncated to the interval $(x_1(u), x_2(u))$ with

$$f(x \mid u) = \frac{g(x)}{t - s} \cdot \mathrm{I}(x_1(u) < x < x_2(u)),$$

where $G(x-) = \lim_{t \uparrow x} G(t)$, $s = G(x_1(u))$ and $t = G(x_2(u)-)$ are CDF values evaluated at the endpoints, and $\lceil x \rceil$ and $\lfloor x \rfloor$ represent the ceiling and floor functions of x, respectively. The associated CDF of $[X \mid U = u]$ is

$$F(x \mid u) = \frac{G(x) - s}{t - s}, \quad x_1(u) < x < x_2(u), \tag{7}$$

with $F(x \mid u) = 0$ for $x < x_1(u)$ and $F(x \mid u) = 1$ for $x > x_2(u)$. We may invert $F(x \mid u)$ to obtain the $\varphi \in (0, 1)$ quantile of $[X \mid U = u]$ as

$$F^{-}(\varphi \mid u) = G^{-}((t - s)\varphi + s) \tag{8}$$

(e.g. Devroye 1986, p. 38), and therefore obtain exact draws via the inverse CDF method using $X = F^{-}(V \mid u)$ with $V \sim \text{Uniform}(0, 1)$.

3 Step function

To approximate the density $p(u)$, we first identify an interval $[u_L, u_H] \subseteq [0, 1]$ which contains the "descent" from its maximum value to a value of zero; any further effort should be focused within this interval. Define u_L as the smallest number such that $\mathrm{P}(A_{u_L}) < \mathrm{P}(A_0)$ for the unnormalized density and u_H as the smallest number such that $\mathrm{P}(A_{u_H}) > 0$. A bisection method may be used to locate u_L and u_H; see Remark 3 at the end of this section.

To approximate the unnormalized $\mathrm{P}(A_u)$, let $u_0 < \cdots < u_N$ be knot points with $u_0 = u_L$ and $u_N = u_H$ and consider the function

$$h^*(u) = \mathrm{P}(A_{u_0}) \cdot \mathrm{I}(0 \leq u < u_0) + \sum_{j=0}^{N-1} \mathrm{P}(A_{u_j}) \cdot \mathrm{I}(u_j \leq u < u_{j+1}).$$

A density is obtained using $h(u) = h^*(u)/a$ with

$$\begin{aligned} a &= \int_0^1 h^*(u) du \\ &= \mathrm{P}(A_{u_0}) \cdot u_0 + \sum_{j=0}^{N-1} \mathrm{P}(A_{u_j}) \cdot (u_{j+1} - u_j). \end{aligned}$$

The corresponding CDF is the piecewise linear function

$$H(u) = a^{-1} \mathrm{P}(A_{u_0}) u, \quad \text{if } 0 \leq u < u_0,$$

and

$$\begin{aligned} H(u) = &a^{-1} \mathrm{P}(A_{u_0}) u_0 + a^{-1} \sum_{j=0}^{\ell-1} \mathrm{P}(A_{u_j}) \cdot (u_{j+1} - u_j) \\ &+ a^{-1} \mathrm{P}(A_{u_\ell}) \cdot (u - u_\ell), \end{aligned}$$

if $u_\ell \leq u < u_{\ell+1}$ for $\ell \in \{0, \ldots, N-1\}$, $H(u) = 0$ if $u \leq 0$ and $H(u) = 1$ if $u \geq u_N$. The quantile function is also a piecewise linear function,

$$H^{-1}(\varphi) = u_\ell + (u_{\ell+1} - u_\ell) \frac{\varphi - H(u_\ell)}{H(u_{\ell+1}) - H(u_\ell)}$$

for $H(u_\ell) \leq \varphi < H(u_{\ell+1})$, $\ell \in \{0, \ldots, N-1\}$. A draw from h can now be generated by $U = H^{-1}(V)$ where $V \sim \text{Uniform}(0, 1)$.

The following proposition shows that the closeness of $h(u)$ to $p(u)$ can be characterized by total variation distance. Let $\mathscr{R}_j$ represent the rectangle in $\mathbb{R}^2$ whose upper-left point is $(u_{j-1}, \mathrm{P}(A_{u_{j-1}}))$ and lower-right point is $(u_j, \mathrm{P}(A_{u_j}))$, for $j = 1, \ldots, N$. The area of $\mathscr{R}_j$ is $|\mathscr{R}_j| = \left[\mathrm{P}(A_{u_{j-1}}) - \mathrm{P}(A_{u_j})\right](u_j - u_{j-1})$.

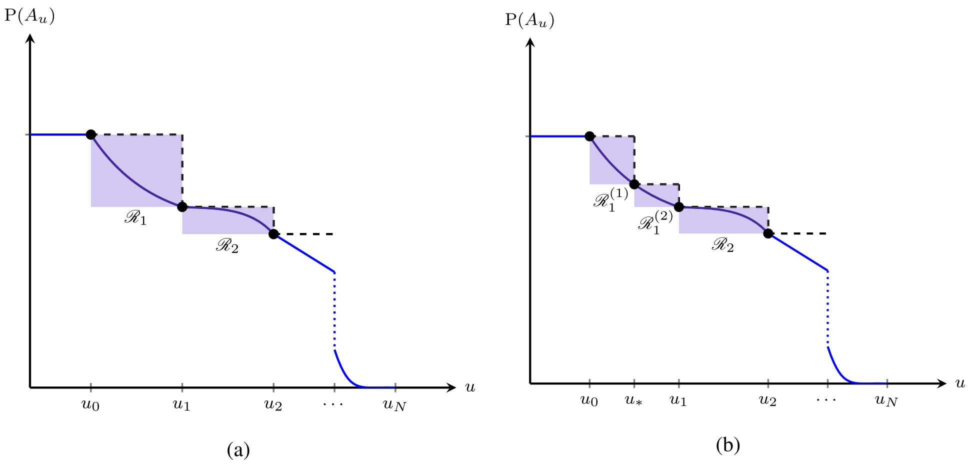
Proposition 1 *Let $\mathscr{B}$ denote the collection of measurable subsets of* $[0, 1]$*; then*

$$\sup_{B \in \mathscr{B}} \left| \int_B h(u) du - \int_B p(u) du \right| \leq \frac{c}{\psi} \sum_{j=1}^{N} |\mathscr{R}_j|. \tag{9}$$

Proof First considering the unnormalized densities,

$$\begin{aligned} &\sup_{B \in \mathscr{B}} \left| \int_B h^*(u) du - \int_B \mathrm{P}(A_u) du \right| \qquad (10) \\ &= \sup_{B \in \mathscr{B}} \int_B [h^*(u) - \mathrm{P}(A_u)] du \\ &= \int_0^1 [h^*(u) - \mathrm{P}(A_u)] du \\ &= \sum_{j=1}^{N} \int_{u_{j-1}}^{u_j} [h^*(u) - \mathrm{P}(A_u)] du \end{aligned}$$

Fig. 1 (a) Step function $h^*(u)$ (dashed black lines) to approximate a concocted $\mathrm{P}(A_u)$ (solid blue curve) based on knots $u_0, \ldots, u_N$. Blue shaded areas represent rectangles $\mathscr{R}_j$. (b) Updated step function with u_* inserted at midpoint of u_0 and u_1; $\mathscr{R}_1$ is replaced by $\mathscr{R}_1^{(1)}$ and $\mathscr{R}_1^{(2)}$



(a)

(b)

$$\leq \sum_{j=1}^{N} \int_{u_{j-1}}^{u_j} [\mathrm{P}(A_{u_{j-1}}) - \mathrm{P}(A_{u_j})]du$$
$$= \sum_{j=1}^{N} [\mathrm{P}(A_{u_{j-1}}) - \mathrm{P}(A_{u_j})](u_j - u_{j-1}) = \sum_{j=1}^{N} |\mathscr{R}_j|. \qquad (11)$$

We have used the fact that $h^*(u) = \mathrm{P}(A_u)$ for $u \in [0, u_0]$ and $h^*(u) \geq \mathrm{P}(A_u)$ otherwise. Integrating each term of the inequality $\mathrm{P}(A_u) \leq h^*(u) \leq 1$ over $u \in [0, 1]$ gives

$$\psi/c \leq a \leq 1 \quad \iff \quad 1 \leq 1/a \leq c/\psi. \qquad (12)$$

Combining this with (11) yields inequalities for the normalized densities

$$\int_B h(u)du - \int_B p(u)du$$
$$\leq \frac{c}{\psi} \left[\int_B h^*(u)du - \int_B \mathrm{P}(A_u)du \right]$$
$$\leq \frac{c}{\psi} \sum_{j=1}^{N} |\mathscr{R}_j|. \qquad (13)$$

and

$$\int_B p(u)du - \int_B h(u)du$$
$$\leq \frac{c}{\psi} \int_B h^*(u)du - \frac{1}{a} \int_B h^*(u)du$$
$$= \left[\frac{a - \psi/c}{a\psi/c} \right] \int_B h^*(u)du$$
$$\leq \left[\frac{a - \psi/c}{a\psi/c} \right] a$$
$$= \frac{c}{\psi} \left[\int_0^1 h^*(u)du - \int_0^1 \mathrm{P}(A_u)du \right]$$
$$\leq \frac{c}{\psi} \sum_{j=1}^{N} |\mathscr{R}_j|. \qquad (14)$$

The result follows from (13) and (14). □

The upper bound in (9) is seen as a product of two factors: the inverse of the normalizing constant of density (3) and $\sum_{j=1}^{N} |\mathscr{R}_j|$ which can be influenced by the selection of knots.

There are a number of possible choices for the knots $u_1, \ldots, u_{N-1}$. Equally-spaced knots $u_j = u_L + (j/N)(u_H - u_L)$ provide simplicity but can fail to capture regions of $[u_L, u_H]$ with sudden changes in $\mathrm{P}(A_u)$. Proposition 1 motivates placement of knots to ensure that no $\mathscr{R}_j$ is too large. Namely, given $u_0, u_1, \ldots, u_k$ with associated rectangles $\mathscr{R}_1, \ldots, \mathscr{R}_k$. we consider placing a new knot u_* at the midpoint of $[u_{j-1}, u_j]$ which has the largest $|\mathscr{R}_j|$. This replaces $\mathscr{R}_j$ with new rectangles $\mathscr{R}_j^{(1)}$ and $\mathscr{R}_j^{(2)}$, yielding an improvement $|\mathscr{R}_j^{(1)}| + |\mathscr{R}_j^{(2)}| < |\mathscr{R}_j|$ in regions where $\mathrm{P}(A_u)$ is decreasing; otherwise, $|\mathscr{R}_j^{(1)}| + |\mathscr{R}_j^{(2)}| = |\mathscr{R}_j|$ so that the bound in (9) is no worse. Stated as Algorithm 2, this method often provides a better selection of knots under a fixed N than equally-spaced points, at the cost of increased computation. Use of a data structure such as a priority queue (Cormen et al. 2009, Section 6.5) can help to avoid repeated sorting of $|\mathscr{R}_1|, \ldots, |\mathscr{R}_k|$. An illustration of one step of Algorithm 2 is shown in Figure 1.

Remark 2 (Midpoint) The midpoint in Algorithm 2 is specified by a function $\mathrm{mid}(x, y)$, with typical choices being the arithmetic mean $\mathrm{mid}(x, y) = (x + y)/2$ or the geometric mean $\mathrm{mid}(x, y) = (xy)^{1/2}$. The arithmetic mean may yield a better approximation when $\mathrm{P}(A_u) \gg 0$ on a large potion of $[0, 1]$. However, the geometric mean may be preferred when some knots are extremely small. For example, if $u_L = 10^{-100}$ and $u_H = 10^{-10}$ and a large descent occurs near u_L, the geometric mean 10^{-55} is much closer to the descent than the arithmetic mean $\frac{1}{2}10^{-100} + \frac{1}{2}10^{-10} \approx \frac{1}{2}10^{-10}$. An example where knots are needed very close to zero is given in Sect. 4.1. The geometric mean is assumed for the remainder of the paper unless otherwise noted.

The step function $h^*(u)$ can be used to formulate a rejection sampler to take exact draws from $p(u)$. By construction, $h^*(u) \geq \mathrm{P}(A_u)$ for all $u \in [0, 1]$, so that rejection sampling can be carried out by Algorithm 1. The bound in (9) also bounds the probability of a rejection, which occurs when $[\text{Accept} = 0]$ in Line 6 of Algorithm 1.

Proposition 2 *The probability of rejection in Line 6 of Algorithm 1 is no greater than* $\frac{c}{\psi} \sum_{j=1}^{N} |\mathscr{R}_j|$.

Proof Suppose $V \sim \text{Uniform}(0, 1)$ and $U \sim h(u)$ are independent random variables and let $M = a/(\psi/c)$. The probability of rejecting a candidate is

$$\begin{aligned}
&\mathrm{P}\left(V > \frac{\mathrm{P}(A_U)}{h^*(U)}\right) \\
&\quad = \mathrm{P}\left(V > \frac{p(U)}{Mh(U)}\right) \\
&\quad = 1 - \mathrm{E}_U\left[\mathrm{P}\left(V \leq \frac{p(U)}{Mh(U)} \,\middle|\, U\right)\right] \\
&\quad = 1 - \mathrm{E}_U\left[\frac{p(U)}{Mh(U)}\right] = 1 - \frac{1}{M} = \frac{a - \psi/c}{a}.
\end{aligned}$$

Applying the inequality $a - \psi/c = \int_0^1 [h^*(u) - \mathrm{P}(A_u)]du \leq \sum_{j=1}^{N} |\mathscr{R}_j|$ from (11) to the numerator and $a \geq \psi/c$ from (12) to the denominator gives the result. □

As anticipated, rejection is assured to be less likely when h^* and $\mathrm{P}(A_u)$ are closer. A rejected u may be added to the set of knot points to decrease the probability of a rejection in subsequent proposals, as shown in Line 5 of Algorithm 1.

Algorithm 1 Rejection sampler based on direct sampling with step function h.

```
1: do
2:     Draw candidate $u$ from step density $h$.
3:     Draw $v$ from Uniform(0, 1).
4:     Accept ← $\mathrm{I}\{v \leq \mathrm{P}(A_u)/h^*(u)\}$.
5:     Update $h^*$ with $u$ as additional knot if adaptive rejection is
       desired.
6: while Accept = 0
7: Draw $x$ from $f(x \mid u)$.
8: return $x$.
```

Remark 3 [Bisection method] A bisection search method (e.g. Lange 2010, Section 5) is useful in several computations in this section. Suppose $\mathscr{S} \subseteq \mathbb{R}$ and $\zeta(x) : \mathscr{S} \to \{0, 1\}$ is a step function which increases from 0 to 1 at a point x^*. The objective of Algorithm 3 is to identify x^* by supplying lower and upper bounds $x_L < x_H$ such that $\zeta(x_L) = 0$ and $\zeta(x_H) = 1$, a function $\text{mid}(x, y) : \mathscr{S}^2 \to \mathscr{S}$ which returns a point in $[x, y]$, and a distance function $\text{dist}(x, y)$. We may therefore write $x^* = \min\{x \in [x_L, x_H] : \zeta(x) = 1\}$. Algorithm 3 is useful in the following computations.

Algorithm 2 Select knots $u_1, \ldots, u_{N-1}$ to reduce $\sum_{j=1}^{N} |\mathscr{R}_j|$.

```
Let $u^{(0)} = u_L$, and $u^{(1)} = u_H$.
for $i = 1, \ldots, N-1$ do
    Let $u_0 < \ldots < u_i$ be $u^{(0)}, \ldots, u^{(i)}$ in sorted order.
    Let $|\mathscr{R}_j| = \{\mathrm{P}(A_{u_{j-1}}) - \mathrm{P}(A_{u_j})\}(u_j - u_{j-1})$ for $j = 1, \ldots, i$.
    Let $j^* = \operatorname{argmax}_{j=1,\ldots,i} |\mathscr{R}_j|$.
    Let $u^{(i+1)} = \text{mid}(u_{j^*-1}, u_{j^*})$.
end for
Let $u_0 < \ldots < u_N$ be $u^{(0)}, \ldots, u^{(N)}$ in sorted order.
return $(u_0, \ldots, u_N)$.
```

1. To find u_L, the smallest $u \in [0, 1]$ such that $\mathrm{P}(A_u) < \mathrm{P}(A_0)$, we first locate a sufficiently small $j^* \in \{0, 1, 2, 4, 8, \ldots\}$ until $\mathrm{P}(A_{\exp(-j)}) = \mathrm{P}(A_0)$. Algorithm 3 may be used with $\zeta(u) = I\{\mathrm{P}(A_u) < \mathrm{P}(A_0)\}$ with $x_L = e^{-j^*}$, and $x_H = 1$.
2. To find u_H, the smallest $u \in [0, 1]$ such that $\mathrm{P}(A_u) = 0$, Algorithm 3 may be used with $\zeta(u) = I\{\mathrm{P}(A_u) > 0\}$, $x_L = u_L$, and $x_H = 1$.
3. The quantile function $H^{-1}(\varphi)$ may be evaluated by Algorithm 3. Given precomputed values $H(u_0)$, …, $H(u_N)$ of the associated CDF, the index ℓ of the interval containing φ can be identified using $\mathscr{S} = \{0, 1, \ldots, N\}$, $x_L = 0$, $x_H = N$, $\text{mid}(\ell_1, \ell_2) = \lfloor(\ell_1 + \ell_2)/2\rfloor$, and $\zeta(\ell) = I\{H(u_\ell) \geq \varphi\}$. From here, linearity between $H(u_\ell)$ and $H(u_{\ell+1})$ yields

$$\begin{aligned}
H^{-1}(\varphi) = &u_\ell + (u_{\ell+1} - u_\ell)\{\varphi - H(u_\ell)\}/\{H(u_{\ell+1}) \\
&- H(u_\ell)\}.
\end{aligned}$$

Algorithm 3 Bisection search for $x^* = \min\{x \in [x_L, x_H] : \zeta(x) = 1\}$. Inputs are bounds $x_L < x_H$, a step function $\zeta(x)$ with $\zeta(x_L) = 0$ and $\zeta(x_H) = 1$, a midpoint function $\text{mid}(x, y)$, a distance function $\text{dist}(x, y)$, and a tolerance $\delta > 0$.

```
$x = \text{mid}(x_L, x_H)$
while $\text{dist}(x_L, x_H) > \delta$ do
    $x_L = \zeta(x) \cdot x_L + [1 - \zeta(x)] \cdot x$
    $x_H = \zeta(x) \cdot x + [1 - \zeta(x)] \cdot x_H$
    $x = \text{mid}(x_L, x_H)$
end while
return $x$
```

4 Illustrative examples

We now demonstrate the direct sampler with step function through three examples. Algorithm 1 is used throughout with a prespecified number N of initial knots selected by Algorithm 2, and subsequent knots added through adaptive rejection. The direct sampler requires more computation than alternative methods which are mentioned in the three examples, but also generates an exact sample with relatively very few rejections. All reported run times were measured on an Intel Core i7–2600 3.40 GHz workstation with four CPU cores.

4.1 Sampling from Conway-Maxwell Poisson

The Conway-Maxwell Poisson distribution has become popular in recent years as a count model which can express either over- or underdispersion (Shmueli et al. 2005). The Conway-Maxwell Poisson distribution CMP(λ, ν) has probability mass function (pmf)

$$f(x \mid \boldsymbol{\theta}) = \frac{\lambda^x}{(x!)^\nu Z(\lambda, \nu)}, \quad x = 0, 1, \ldots, \quad \lambda > 0, \nu > 0, \tag{15}$$

a weighted density in the form of (1), with normalizing constant $Z(\lambda, \nu) = \sum_{x=0}^{\infty} \lambda^x/(x!)^\nu$. The over- or underdispersion of CMP(λ, ν) is most readily compared to Poisson(λ): CMP is overdispersed when $\nu < 1$, underdispersed when $\nu > 1$, and is equivalent when $\nu = 1$. CMP also has several other well-known special cases. When $\lambda \in (0, 1)$ and $\nu = 0$, CMP(λ, ν) becomes a Geometric distribution with density $f(x \mid \lambda) = (1 - \lambda)\lambda^x$ for $x = 0, 1, \ldots$. When $\nu \to \infty$, the CMP(λ, ν) density converges to that of Bernoulli$(\lambda/(1+\lambda))$.

The normalizing constant $Z(\lambda, \nu)$ is a series which does not appear to have a closed form. Approximating the normalizing constant has been a topic of interest (e.g. Gaunt et al. 2019). The expansion

$$Z(\lambda, \nu) = \frac{\exp(\nu\lambda^{1/\nu})}{\lambda^{(\nu-1)/2\nu}(2\pi)^{(\nu-1)/2}\nu^{1/2}} \left\{1 + O(\lambda^{-1/\nu})\right\}$$

given by Shmueli et al. (2005) illustrates in particular that the magnitude of $Z(\lambda, \nu)$ can vary wildly with λ and ν. For example, $Z(\lambda, 1) = e^\lambda$ for any $\lambda > 0$ but $Z(2, 0.075) \approx e^{780.515}$. Given this volatility, an exact method of generating variates which avoids computation of the normalizing constant is desirable.

Several recent papers have considered Bayesian analysis with CMP using the exchange algorithm (Møller et al. 2006; Murray et al. 2006). The exchange algorithm utilizes a data augmentation step in Metropolis-Hastings sampling; an exact draw from the data-generating model is used to avoid computing the normalizing constant $Z(\lambda, \nu)$ in the acceptance ratio and therefore obtain an MCMC sampler for the unknown parameters. Chanialidis et al. (2018) and Benson and Friel (2021) take rejection sampling approaches to generate exact CMP draws and implement the exchange algorithm: Chanialidis et al. (2018) creates an envelope based on a piecewise Geometric distribution, while Benson and Friel (2021) use an envelope based on the Geometric distribution when $\nu < 1$ and Poisson otherwise.

The direct sampler described in Algorithm 1 can be used to obtain exact draws from CMP with low probability of rejection and avoid explicit computation of the normalizing constant. To do this, the cases $\nu \geq 1$ and $\nu < 1$, corresponding to under- and overdispersion, are now addressed individually.

Case $\nu \geq 1$. Rewrite the unnormalized density (15) as

$$f(x \mid \boldsymbol{\theta}) \propto \frac{\lambda^x}{(x!)^\nu} = \left(\frac{\lambda}{1+\lambda}\right)^x \frac{1}{1+\lambda}(1+\lambda)^{x+1}\frac{\lambda^x}{(x!)^\nu} \tag{16}$$

and let base density $g(x \mid \lambda) = [\lambda(1+\lambda)^{-1}]^x(1+\lambda)^{-1}, x = 0, 1, \ldots$, be the pmf of a Geometric$(1/\{1+\lambda\})$ distribution. The weight function, on the log-scale, and its first and second derivative are respectively

$$\log w(x \mid \lambda, \nu) = (x+1)\log(1+\lambda) - \nu \log \Gamma(x+1),$$
$$\frac{\partial}{\partial x}\log w(x \mid \lambda, \nu) = \log(1+\lambda) - \nu\psi(x+1), \tag{17}$$
$$\frac{\partial^2}{\partial x^2}\log w(x \mid \lambda, \nu) = -\nu\psi'(x+1). \tag{18}$$

Note that $\psi(x+1)$ is increasing for $x \geq 0$, where $\psi(1) = -0.57721566\ldots$ is the Euler-Mascheroni constant. For any $\lambda > 0$, $\log(1+\lambda) > 0$ so that we may locate an x_L and x_H that yield negative and positive values of (17), respectively. A root x^* of (17) exists in $[x_L, x_H]$, and may be identified using a root-finding method such as Algorithm 3. The function (18) is negative for all x, so that $\log w(x \mid \lambda, \nu)$ is concave and x^* is a maximizer with $c = \log w(x^* \mid \lambda, \nu)$. To find the endpoints $x_1(u) < x_2(u)$ of the interval $A_u = \{x > 0 : \log w(x \mid \lambda, \nu) > \log(uc)\}$, root-finding may be applied twice to the function $\log w(x \mid \lambda, \nu) - \log(uc)$: once to obtain $x_1(u)$ from the interval $[0, c]$, and again to obtain $x_2(u)$ from the interval $[c, x_H^*]$, where x_H^* is a number large enough that $\log w(x_H^* \mid \lambda, \nu) - \log(uc)$ is negative.

Case $\nu < 1$. Variates from CMP(λ, ν) can become very large as ν is taken closer to zero, especially when $\lambda > 1$. Here, the support of a Geometric$(1/\{1+\lambda\})$ base distribution may be practically disjoint from the target CMP, leading to extremely small probabilities in computations such as (7) and (8). To illustrate, suppose $X \sim$ CMP(λ, ν) with $\lambda = 2$ and $\nu = 0.075$; here, P$(X \leq 7,306) \approx e^{-40}$ but P$(S >$

$7,086) \approx e^{-2873.531}$ for $S \sim \text{Geometric}(1/\{1+\lambda\})$ so that $[X \leq S]$ effectively never occurs. A more convenient base distribution is given by the reparameterization of CMP based on ν and $\mu = \lambda^{1/\nu}$ used by Guikema and Goffelt (2008) to formulate regression models. The unnormalized portion of density (15) may now be decomposed as

$$f(x \mid \boldsymbol{\theta}) \propto \frac{\mu^{\nu x}}{(x!)^\nu} = \left(\frac{\mu}{1+\mu}\right)^x \frac{1}{1+\mu}(1+\mu)^{x+1}\frac{\mu^{x(\nu-1)}}{(x!)^\nu} \tag{19}$$

so that the base density $g(x \mid \mu) = [\mu(1+\mu)^{-1}]^x(1+\mu)^{-1}$, $x = 0, 1, \ldots$, is the pmf of $T \sim \text{Geometric}(1/\{1+\mu\})$. The quantiles of T corresponding to probabilities 0.025 and 0.975 are 261 and 38,075, compared to 9,607 and 11,061 for X, suggesting that the distribution of T is more suitable than that of S as a base distribution. The log of the weight function and its first and second derivative are now, respectively,

$$\begin{aligned}\log w(x \mid \mu, \nu) &= (x+1)\log(1+\mu) - \nu\log(x+1)\\ &\quad +x(\nu-1)\log\mu,\\ \tfrac{\partial}{\partial x}\log w(x \mid \mu, \nu) &= \{\log(1+\mu) + (\nu-1)\log\mu\} \qquad (20)\\ &\quad -\nu\psi(x+1),\\ \tfrac{\partial^2}{\partial x^2}\log w(x \mid \mu, \nu) &= -\nu\psi'(x+1). \qquad (21)\end{aligned}$$

A root of (20) exists if the term $\log(1+\mu) + (\nu-1)\log\mu$ is positive. To verify positivity, if $\mu < 1$, then $\log(1+\mu) + (\nu-1)\log\mu \geq \log(1+\mu) \geq 0$; On the other hand, if $\mu \geq 1$ then $\log(1+\mu) + (\nu-1)\log\mu \geq \log(1+\mu) - \log\mu = \log(1/\mu+1) \geq 0$. Maximization and root-finding may then proceed similarly to the case where $\nu \geq 1$.

Figure 2 compares the unnormalized density $\text{P}(A_u)$ with unnormalized step function $h^*(u)$ in two CMP settings with $\lambda = 2$: one with $\nu = 0.5$ using $N = 13$ knots and one with $\nu = 0.2$ using $N = 20$ knots, corresponding to progressively higher levels of overdispersion. Three different knot selection methods are shown for comparison: equal spacing, Algorithm 2 using geometric midpoints, and Algorithm 2 using arithmetic midpoints. Although $\nu < 1$ in this example, the step function has been constructed from decomposition (16) to illustrate the effect of using a base distribution which differs either moderately and greatly from the target. The case $\nu = 0.5$ is handled relatively well by all three methods, with arithmetic midpoint providing the best approximation followed by equal spacing, then geometric midpoint. A decrease in ν to 0.2 is seen to create a much more difficult situation, with much of the density occurring in a small subinterval of $[u_L, u_H]$. Here, equal spacing will require a large N to obtain a useful approximation. Algorithm 2 with arithmetic midpoint produces a better approximation than equal spacing, but has not yet located the steep descent shown on the left of the display. On the other hand, the geometric midpoint is able to capture this feature; this is due to its suitability with very small magnitude numbers as discussed in Remark 2.

Remark 4 [Weighted rectangles] The approximation in Figures 2e and 2f can be further improved, without increasing N, by using a weighted priority $\omega\log\{\text{P}(A_{u_{j-1}}) - \text{P}(A_{u_j})\} + (1-\omega)\log(u_j - u_{j-1})$, $\omega \in (0, 1)$, in place of $\log|\mathscr{R}_j|$ to order the rectangles in Algorithm 2. In particular, $\omega > 1/2$ prioritizes taller rectangles over wider ones having equal area which encourages knot placement at sudden descents occurring on very short intervals.

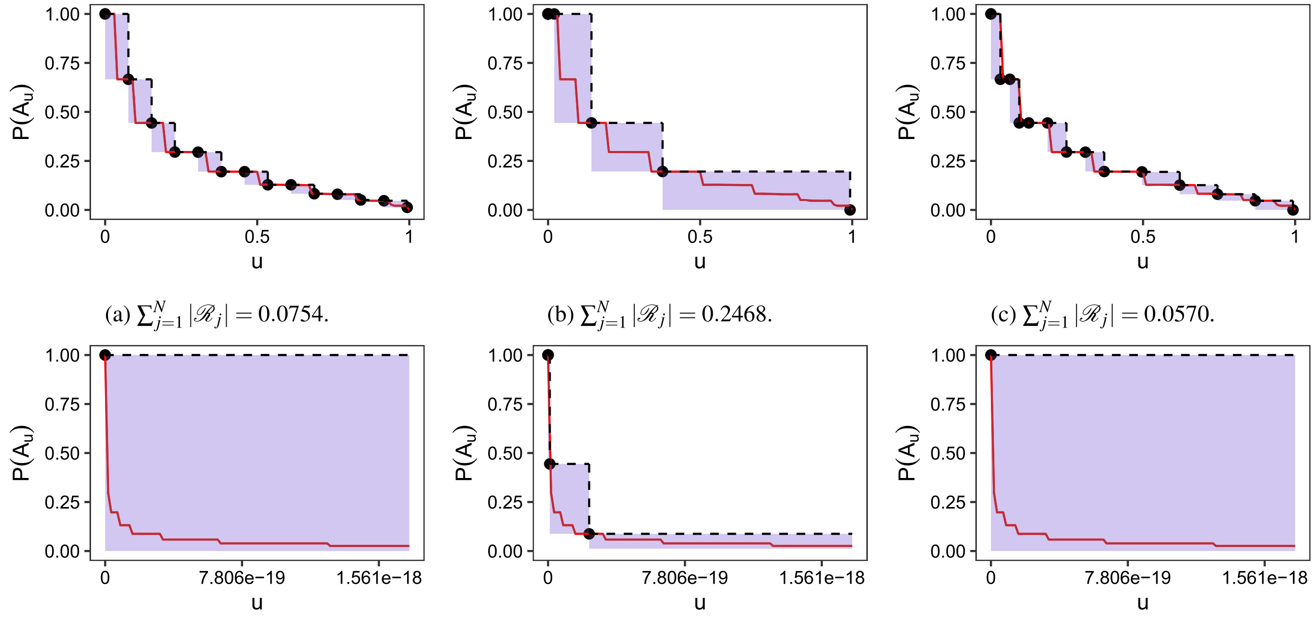
Figure 3 displays draws of $\text{CMP}(\lambda, \nu)$ from Algorithm 1 with $\lambda = 2$ and $\nu \in \{0.05, 0.5, 2, 5\}$. Here, decomposition (16) is used with $\nu \in \{2, 5\}$ and (19) is used with $\nu \in \{0.05, 0.5\}$. As anticipated, the empirical pmf of 20,000 draws matches closely to the exact pmf (15). With $N = 10$ knots initially selected in each case, the number of rejections was 279, 86, 40, and 27 in Figures 3a, 3b, 3c, and 3d, respectively. This demonstrates the ability of the samplers obtained in this section to generate CMP variates with small probability of rejection. This may be contrasted to acceptance rates as low as about 20% reported by Benson and Friel (2021); however, their rejection sampler requires less computation and therefore may be faster in practice.

4.2 Sampling the dependence parameter in conditional autoregression

In a conditional autoregression (CAR) setting (e.g. Cressie 1991, Section 6.6), the joint distribution of a random vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ implies a certain regression for each conditional distribution $[\eta_i \mid \boldsymbol{\eta}_{-i}]$, $i = 1, \ldots, k$, where $\boldsymbol{\eta}_{-i} = (\eta_1, \ldots, \eta_{i-1}, \eta_{i+1}, \ldots \eta_k)$. Let us consider a particular mixed effects CAR model which is useful for data observed on areal units in a spatial domain. Suppose there are k distinct areas and let $\boldsymbol{A} = (a_{ij})$ be a $k \times k$ adjacency matrix; $a_{ij} = 1$ if areas i and j are adjacent and $i \neq j$, otherwise $a_{ij} = 0$. Let $\boldsymbol{D} = \text{Diag}(a_{1+}, \ldots, a_{k+})$ with $a_{i+} = \sum_{j=1}^k a_{ij}$ be a diagonal matrix containing the row sums of $\boldsymbol{A}$. Suppose $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a vector of observed outcomes, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{S} \in \mathbb{R}^{n \times k}$ are fixed design matrices, and

$$\begin{aligned}&\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{S\eta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}),\\ &\boldsymbol{\eta} \sim \text{N}(\boldsymbol{0}, \tau^2(\boldsymbol{D} - \rho\boldsymbol{A})^{-1}).\end{aligned} \tag{22}$$

Here it can be shown that the conditionals have distribution $[\eta_i \mid \boldsymbol{\eta}_{-i}] \sim \text{N}((\rho/a_{i+})\sum_{j=1}^k \eta_j a_{ij}, \tau^2/a_{i+})$. The parameter ρ must be in the interval $[0, 1]$; the matrix $\boldsymbol{D} - \rho\boldsymbol{A}$ is nonsingular provided that $\rho < 1$, while the inverse does not exist when $\rho = 1$ and a pseudo-inverse may be instead considered. To complete a Bayesian specification of the model,



(a) $\sum_{j=1}^{N} |\mathscr{R}_j| = 0.0754$.

(b) $\sum_{j=1}^{N} |\mathscr{R}_j| = 0.2468$.

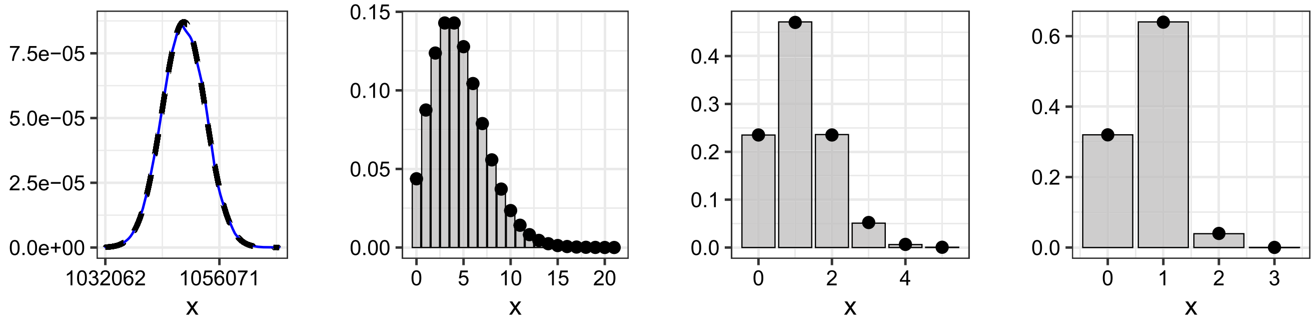
(c) $\sum_{j=1}^{N} |\mathscr{R}_j| = 0.0570$.

(d) $\sum_{j=1}^{N} |\mathscr{R}_j| = 4.561 \times 10^{-11}$.

(e) $\sum_{j=1}^{N} |\mathscr{R}_j| = 1.007 \times 10^{-17}$.

(f) $\sum_{j=1}^{N} |\mathscr{R}_j| = 1.743 \times 10^{-15}$.

Fig. 2 Realizations of $\mathrm{P}(A_u)$ (solid red line) and $h^*(u)$ (dashed line) for Conway-Maxwell Poisson with $\lambda = 2$. Shaded rectangles represent the rectangles $\mathscr{R}_j$. The top row (a), (b), and (c) correspond to $\nu = 0.5$ with $N = 13$ while the bottom row (d), (e), and (f) correspond to $\nu = 0.2$ with $N = 20$. Knots for (a) and (d) were selected by equal spacing, (b) and (e) were selected using Algorithm 2 with geometric mean to compute midpoints, and (c) and (f) were selected using Algorithm 2 with arithmetic mean to compute midpoints. Most knots are excluded from the displays in (d), (e), and (f). Subcaptions display total rectangle area $\sum_{j=1}^{N} |\mathscr{R}_j|$ achieved by the approximation



(a) $\nu = 0.05$.

(b) $\nu = 0.5$.

(c) $\nu = 2$.

(d) $\nu = 5$.

Fig. 3 Empirical density of 20,000 draws versus pmf of CMP from with $\lambda = 2$ and ν as specified in the subcaption. In (a), solid blue line and dashed black line represent empirical density and CMP pmf, respectively. In (b), (c), and (d), gray bars and black dots represent empirical density and CMP pmf, respectively

consider the prior

$$\begin{aligned} &\boldsymbol{\beta} \sim \mathrm{N}(0, \sigma_\beta^2 \boldsymbol{I}) \quad \sigma^2 \sim \mathrm{Uniform}(0, M_\sigma) \\ &\tau^2 \sim \mathrm{Uniform}(0, M_\tau) \quad \rho \sim \mathrm{Uniform}(0, 1), \end{aligned} \tag{23}$$

following Lee (2013). From (22) and (23), and regarding $\boldsymbol{\eta}$ as augmented data to be drawn with the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2)$ and ρ, the following conditionals with familiar distributions are obtained for a Gibbs sampler:

1. $[\boldsymbol{\beta} \mid -] \sim \mathrm{N}(\boldsymbol{\vartheta}_\beta, \boldsymbol{\Omega}_\beta^{-1})$ where $\boldsymbol{\Omega}_\beta = \sigma^{-2}\boldsymbol{X}^\top \boldsymbol{X} + \sigma_\beta^{-2}\boldsymbol{I}$ and $\boldsymbol{\vartheta}_\beta = \sigma^{-2}\boldsymbol{\Omega}_\beta^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{\eta})$,
2. $[\boldsymbol{\eta} \mid -] \sim \mathrm{N}(\boldsymbol{\vartheta}_\eta, \boldsymbol{\Omega}_\eta^{-1})$ where $\boldsymbol{\Omega}_\eta = \sigma^{-2}\boldsymbol{S}^\top \boldsymbol{S} + \tau^{-2}(\boldsymbol{D} - \rho \boldsymbol{A})$ and $\boldsymbol{\vartheta}_\eta = \sigma^{-2}\boldsymbol{\Omega}_\eta^{-1}\boldsymbol{S}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$,
3. $[\sigma^2 \mid -] \sim \mathrm{IG}_{[0, M_\sigma]}(a_\sigma, b_\sigma)$, an Inverse Gamma distribution with shape $a_\sigma = n/2$ and rate $b_\sigma = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{S}\boldsymbol{\eta}\|^2$,
4. $[\tau^2 \mid -] \sim \mathrm{IG}_{[0, M_\tau]}(a_\tau, b_\tau)$ with $a_\tau = k/2$ and $b_\tau = \frac{1}{2}\boldsymbol{\eta}^\top(\boldsymbol{D} - \rho\boldsymbol{A})\boldsymbol{\eta}$.

Here, $[\boldsymbol{T} \mid -]$ denotes the distribution of $\boldsymbol{T}$ based on all other random variables and a distribution with subscript $[a, b]$ denotes that it is truncated to that interval. The conditional for ρ takes the more unfamiliar form

$$f(\rho \mid -) \propto |\boldsymbol{D} - \rho \boldsymbol{A}|^{1/2} \exp\left\{-\frac{\rho}{2\tau^2}\boldsymbol{\eta}^\top \boldsymbol{A}\boldsymbol{\eta}\right\} \mathrm{I}\{\rho \in [0, 1]\}. \tag{24}$$

Lee (2013) uses a Metropolis-Hastings approach to sample from (24). At the rth iteration, a candidate ρ^* is drawn from truncated Normal proposal distribution $\mathrm{N}_{[0,1]}(\rho^{(r-1)}, \sigma^2_{\text{prop}})$ so that $\rho^{(r)}$ is assigned to ρ^* with probability $\min\{1, f(\rho^* \mid -)/f(\rho^{(r-1)} \mid -)\}$ and to $\rho^{(r-1)}$ otherwise. This requires selecting—or adaptively tuning—the proposal variance σ^2_{prop} to be large enough that the chain is not restricted to very small moves, but not too large that too many proposals are rejected. Let us now consider a direct sampler to generate exact draws from (24). First, suppose $\boldsymbol{U\Phi U}^\top$ is the spectral decomposition of $\boldsymbol{A}$ and let $\boldsymbol{Q} = \boldsymbol{\Phi}^{1/2}\boldsymbol{U}^\top \boldsymbol{D}^{-1}\boldsymbol{U\Phi}^{1/2}$. Using a well-known property of determinants (e.g. Banerjee and Roy 2014, Theorem 10.11),

$$\begin{aligned} |\boldsymbol{D} - \rho \boldsymbol{A}| &= |\boldsymbol{D} - \rho \boldsymbol{U\Phi U}^\top| \\ &= |\boldsymbol{D}| \cdot |\boldsymbol{I} - \rho \boldsymbol{Q}|. \end{aligned} \tag{25}$$

Let $\lambda_1 \geq \cdots \geq \lambda_k$ be the eigenvalues of $\boldsymbol{Q}$ with corresponding eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. Then $\boldsymbol{v}_i$ is also an eigenvector of $\boldsymbol{I} - \rho\boldsymbol{Q}$ with corresponding eigenvalue $1 - \rho\lambda_i$. Therefore, $|\boldsymbol{D} - \rho\boldsymbol{A}| = \prod_{i=1}^k a_{i+} \cdot \prod_{i=1}^k (1 - \rho\lambda_i)$. Note that the elements of $\boldsymbol{\Phi}^{1/2}$ may be complex numbers but $\lambda_1, \ldots, \lambda_k$ are real. From (24), we may write

$$\begin{aligned} f(\rho \mid -) \propto &\left[\prod_{i=1}^k (1 - \rho\lambda_i)\right]^{1/2} \\ &\times \exp\left\{\frac{\rho}{2\tau^2}\boldsymbol{\eta}^\top \boldsymbol{A}\boldsymbol{\eta}\right\} \mathrm{I}\{\rho \in [0, 1]\}. \end{aligned}$$

Let us take $g(\rho) = \mathrm{I}\{\rho \in [0, 1]\}$ so that the base distribution is Uniform(0, 1) and the weight function w is specified on the log-scale by

$$\begin{aligned} \log w(\rho) =& \frac{1}{2}\sum_{i=1}^k \log(1 - \rho\lambda_i) + \frac{\rho}{2\tau^2}\boldsymbol{\eta}^\top \boldsymbol{A}\boldsymbol{\eta} \\ &+ \log \mathrm{I}\{\rho \in [0, 1]\}. \end{aligned} \tag{26}$$

For $\rho \in (0, 1)$,

$$\frac{\partial}{\partial\rho} \log w(\rho) = -\frac{1}{2}\sum_{i=1}^k \frac{\lambda_i}{1 - \rho\lambda_i} + \frac{1}{2\tau^2}\boldsymbol{\eta}^\top \boldsymbol{A}\boldsymbol{\eta}, \tag{27}$$

$$\frac{\partial^2}{\partial\rho^2} \log w(\rho) = -\frac{1}{2}\sum_{i=1}^k \frac{\lambda_i^2}{(1 - \rho\lambda_i)^2}. \tag{28}$$

Now, for $\rho \in [0, 1)$ and assuming all areas have at least one adjacent neighbor so that $\boldsymbol{D} - \rho\boldsymbol{A}$ is positive definite, both $|\boldsymbol{D}| = \prod_{i=1}^k a_{k+} > 0$ and $|\boldsymbol{D} - \rho\boldsymbol{A}| > 0$. Therefore, (25) implies $0 < |\boldsymbol{I} - \rho\boldsymbol{Q}| = \prod_{i=1}^k (1 - \rho\lambda_i)$ so that $\rho\lambda_i < 1$ for each $i = 1, \ldots, k$. Now it can be seen that (28) is negative and a root of (27) is a maximum of (26). Furthermore, $\frac{1}{2}\sum_{i=1}^k \frac{\lambda_i}{1-\rho\lambda_i}$ is an increasing function of ρ so that (27) has at most one root. Therefore, the maximizer ρ^* of (26) occurs at the root if it exists; otherwise, it occurs at one of the endpoints $\{0, 1\}$ of the domain. To find the roots $\{\rho_1(u), \rho_2(u)\}$ of the interval $A(u) = \{\rho \in [0, 1] : w(\rho) > uc\}$, note that $\rho_1(u) = 0$ if $\log w(0) > \log(uc)$; otherwise, a solution $\rho_1(u)$ to $\log w(\rho) = \log(uc)$ may be found in $[0, \rho^*]$ numerically. Similarly, $\rho_2(u) = 1$ if $\log w(1) > \log(uc)$; otherwise, a solution $\rho_2(u)$ to $\log w(\rho) = \log(uc)$ may be found in $[\rho^*, 1]$ numerically. Operations involving the Uniform(0, 1) base distribution outlined in Section 2 are simple, using expressions for the CDF $G(\rho) = \rho$ for $\rho \in [0, 1]$ and quantile function $G^-(\varphi) = \varphi$ for $\varphi \in [0, 1]$.

Now we have a complete Gibbs sampler based on conjugate steps to draw $\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2$, and τ^2, and a direct sampling step to draw ρ. To illustrate the sampler, we revisit the analysis from Lee (2013) on property prices in Glasgow, Scotland. The data are available in the CARBayesdata package Lee (2020). There are $k = 270$ areal units with one observation per area so that $n = k$. The response $\boldsymbol{y}$ is taken to be log of median housing price (in thousands) of properties sold in 2008. Columns of design matrix $\boldsymbol{X}$ include an intercept (corresponding to β_0), log of number of recorded crimes per 10,000 residents (β_1), median number of rooms in a property (β_2), percentage of properties which sold in a year (β_3), and log of average driving time to the nearest shopping center (β_7). The remaining columns are based on a categorical variable indicating the most prevalent property type in the area with levels: “flat” (β_4), “semi-detached” (β_5), “terraced” (β_6), and “detached” (baseline). Because there is one observation per area, $\boldsymbol{S}$ is taken to be a $k \times k$ identity matrix.

Following Lee (2013), hyperparameter values are taken to be $\sigma^2_\beta = 1000$, $M_\sigma = 1000$, and $M_\tau = 1000$. A direct sampler is used to draw ρ following Algorithm 1. Initially, $N = 30$ knots are selected in each iteration of the Gibbs sampler via Algorithm 2. Table 1 compares summaries of the draws from this Gibbs sampler to those from CARBayes version 1.6 (Lee 2013), utilizing a Gibbs sampler with Metropolis-Hastings step for ρ. Figure 4 displays draws of ρ from the two samplers. Following Lee (2013), results for both samplers are based on a chain of 100,000 draws with 20,000 discarded as burn-in and keeping one of every remaining 10 to yield 8,000 draws from each. The two results

Table 1 Summary of draws for each parameter based on 8,000 saved draws: 100,000 total draws with 20,000 discarded as burn-in and 9 of every remaining 10 discarded for thinning

	Mean	SD	2.5%	97.5%
(a) Gibbs with direct sampling step				
β_0	4.7745	0.2537	4.2767	5.2608
β_1	−0.1129	0.0305	−0.1721	−0.0531
β_2	0.2218	0.0253	0.1727	0.2720
β_3	0.0023	0.0003	0.0017	0.0029
β_4	−0.2533	0.0578	−0.3677	−0.1417
β_5	−0.1624	0.0500	−0.2602	−0.0647
β_6	−0.2901	0.0627	−0.4153	−0.1671
β_7	−0.0017	0.0289	−0.0581	0.0552
σ^2	0.0244	0.0047	0.0151	0.0334
τ^2	0.0479	0.0180	0.0208	0.0903
ρ	0.9885	0.0109	0.9591	0.9992
(b) Gibbs with Metropolis-Hastings step				
β_0	4.7576	0.2459	4.2734	5.2394
β_1	−0.1116	0.0307	−0.1731	−0.0525
β_2	0.2222	0.0261	0.1710	0.2733
β_3	0.0023	0.0003	0.0016	0.0029
β_4	−0.2558	0.0581	−0.3687	−0.1419
β_5	−0.1637	0.0512	−0.2638	−0.0635
β_6	−0.2927	0.0628	−0.4157	−0.1727
β_7	−0.0017	0.0294	−0.0594	0.0556
σ^2	0.0237	0.0048	0.0144	0.0333
τ^2	0.0540	0.0192	0.0241	0.0995
ρ	0.9815	0.0146	0.9429	0.9979

are quite similar, the most notable difference being that the posterior of ρ is somewhat more right-skewed under the direct sampler. To obtain 100,000 draws of ρ, Metropolis-Hastings step rejected 40,232 proposals while the direct sampler rejected a total of only 458 proposals. Recall that each draw of Metropolis-Hastings samples approximately from conditional (24) while direct sampling with rejection is exact. However, the direct sampler required substantially more computation in this setting, taking on the order of 1.7 hours compared to 10 minutes for CARBayes. Performance improvements for the former may be possible; in particular, the weight function (26) changes only by an additive constant $\frac{\rho}{2\tau^2}\boldsymbol{\eta}^\top \boldsymbol{A}\boldsymbol{\eta}$ within each step of the Gibbs sampler so that repetition of some computations may be avoided.

4.3 Sampling the degrees-of-freedom in robust regression

The Student t-distribution may be considered as an alternative to the Normal distribution when additional variability is needed in a linear model. Let t_ν denote a t-distribution with degrees of freedom ν and density function

$$f(x \mid \nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

Suppose outcomes $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \sigma\epsilon_i$ are observed for $i = 1, \ldots, n$, where $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} t_\nu$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ are given covariates. Using a particular data augmentation and taking ν to be fixed, it is possible to formulate a Gibbs sampler whose steps consist of drawing from standard distributions (Gelman et al. 2013). With ν not fixed and inference desired on $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \nu)$, Geweke (1994) proposes a rejection sampler for the conditional of ν. Let us illustrate how a direct sampler can also be used to effectively generate draws from this nonstandard conditional distribution.

An augmented version of the model assumes variables $s_1, \ldots, s_n$ with $\boldsymbol{D}_s = \text{Diag}(s_1, \ldots, s_n)$ such that

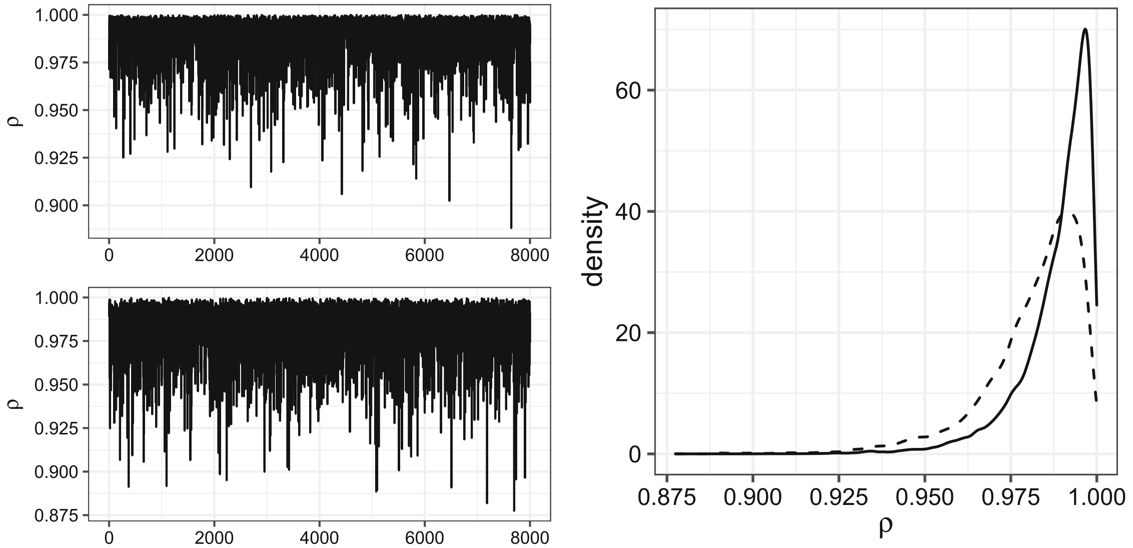
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \sim \text{N}(\boldsymbol{0}, \boldsymbol{D}_s), \quad s_i \overset{\text{iid}}{\sim} \text{IG}(\nu/2, \nu\sigma^2/2), \tag{29}$$

for $i = 1, \ldots n$ with prior distributions $\boldsymbol{\beta} \sim \text{N}(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I})$ and $\sigma^2 \sim \text{Gamma}(a_\sigma, b_\sigma)$. Furthermore, let ν have a $\text{Uniform}(a_\nu, b_\nu)$ prior. The joint distribution of all random variables is

$$\begin{aligned}
&f(\boldsymbol{y}, \boldsymbol{s}, \mu, \sigma^2, \nu) \\
&= (2\pi)^{-n/2} \left[\prod_{i=1}^n s_i^{-1/2}\right] \exp\left\{-\frac{1}{2}\sum_{i=1}^n \frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{s_i}\right\} \\
&\times \frac{(\nu\sigma^2/2)^{n\nu/2}}{\Gamma(\nu/2)^n} \left[\prod_{i=1}^n s_i\right]^{-\frac{\nu}{2}-1} \exp\left\{-\frac{\nu\sigma^2}{2}\sum_{i=1}^n \frac{1}{s_i}\right\} \\
&\times \left[2\pi\sigma_\beta^2\right]^{-d/2} \exp\left\{-\frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\} \\
&\times \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)}(\sigma^2)^{a_\sigma - 1} e^{-b_\sigma \sigma^2}\, \text{I}(a_\nu \le \nu \le b_\nu).
\end{aligned}$$

This distribution yields conditionals

$$\begin{aligned}
&[\boldsymbol{\beta} \mid \boldsymbol{s}, \sigma^2, \nu, \boldsymbol{y}] \sim \text{N}\left(\boldsymbol{\vartheta}_\beta, \boldsymbol{\Omega}_\beta^{-1}\right), \\
&[\sigma^2 \mid \boldsymbol{s}, \boldsymbol{\beta}, \nu, \boldsymbol{y}] \sim \text{Gamma}\left(a_\sigma + \frac{n\nu}{2}, b_\sigma + \frac{\nu}{2}\sum_{i=1}^n \frac{1}{s_i}\right), \\
&[\boldsymbol{s} \mid \sigma^2, \boldsymbol{\beta}, \nu, \boldsymbol{y}] = \prod_{i=1}^n \text{IG} \\
&\qquad \times \left(s_i \,\middle|\, \frac{\nu+1}{2}, \frac{\nu\sigma^2}{2} + \frac{(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2}{2}\right),
\end{aligned}$$



(a) Trace plots from direct sampler (top) and Metropolis-Hastings (bottom).

(b) Empirical density from direct sampler (solid line) and Metropolis-Hastings (dashed line).

Fig. 4 Draws of ρ from Gibbs sampler with direct sampling step versus Metropolis-Hastings step

having familiar forms so that draws are straightforward, where

$$\boldsymbol{\Omega}_\beta = \boldsymbol{X}^\top \boldsymbol{D}_s^{-1} \boldsymbol{X} + \sigma_\beta^{-2} \boldsymbol{I}, \quad \boldsymbol{\vartheta}_\beta = \boldsymbol{\Omega}_\beta^{-1} \boldsymbol{X}^\top \boldsymbol{D}_s^{-1} \boldsymbol{y},$$

and $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is the matrix with rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. More interesting is the distribution of $[\nu \mid \boldsymbol{s}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y}]$, which has the form

$$f(\nu \mid \boldsymbol{s}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y}) \propto w(\nu) g(\nu) \tag{30}$$

with Uniform(a_ν, b_ν) base distribution $g(\nu) = (b_\nu - a_\nu)^{-1} \cdot \mathrm{I}(a_\nu \leq \nu \leq b_\nu)$ and weight function w such that

$$\begin{aligned} &\log w(\nu) \\ &\quad = \frac{n\nu}{2} \log(\nu/2) - n \log \Gamma(\nu/2) - \frac{\nu}{2} \sum_{i=1}^n \log \frac{s_i}{\sigma^2} \\ &\qquad - \frac{\nu}{2} \sum_{i=1}^n \frac{\sigma^2}{s_i} + \log \mathrm{I}(a_\nu \leq \nu \leq b_\nu) \\ &\quad = n \left[\frac{\nu}{2} \log \frac{\nu}{2} - \log \Gamma\left(\frac{\nu}{2}\right) \right] - A\nu + \log \mathrm{I}(a_\nu \leq \nu \leq b_\nu) \end{aligned}$$

where $A = \frac{1}{2} \sum_{i=1}^n \log(s_i/\sigma^2) + \frac{1}{2} \sum_{i=1}^n \sigma^2/s_i$. Temporarily disregarding the indicator $\mathrm{I}(a_\nu \leq \nu \leq b_\nu)$ and considering $\nu \in (0, \infty)$,

$$\frac{d}{d\nu} \log w(\nu) = \frac{n}{2} \left[\log\left(\frac{\nu}{2}\right) - \psi\left(\frac{\nu}{2}\right) \right] + n/2 - A. \tag{31}$$

It can be shown (e.g. Alzer 1997) that

$$\begin{aligned} &\frac{1}{2x} < \log x - \psi(x) < \frac{1}{x}, \quad x > 0, \\ &\lim_{x \to 0} \frac{\log x - \psi(x)}{1/x} = 1, \quad \lim_{x \to \infty} \frac{\log x - \psi(x)}{1/x} = \frac{1}{2}. \end{aligned}$$

Therefore $\log\left(\frac{\nu}{2}\right) - \psi(\frac{\nu}{2})$ is positive, decreases to 0 as ν increases, and increases as ν decreases to 0. Furthermore, the function $g(x) = \frac{1}{2} \log(x/\sigma^2) + \frac{1}{2}(\sigma^2/x)$ is minimized by $x = \sigma^2$ so that $A = \sum_{i=1}^n g(s_i) \geq n/2$. Notice that (31) has a root in $(0, \infty)$ when $A > n/2$, and has no root if $A = n/2$. We can gather some information about the behavior of $\log w(\nu)$ from (31).

1. When $\frac{d}{d\nu} \log w(\nu)$ has no root, it is always positive so that $\log w(\nu)$ is an increasing function. Here, $\log c = \log w(b_\nu)$.
2. When $\frac{d}{d\nu} \log w(\nu)$ has a root, $\log w(\nu)$ has a single maximizer ν^*. Therefore, $\log w(\nu)$ is unimodal on $[a_\nu, b_\nu]$ with $\log c = \log w(\nu^*)$ if $\nu^* \in [a_\nu, b_\nu]$. If $\nu^* > b_\nu$, $\log w(\nu)$ is an increasing function on $[a_\nu, b_\nu]$ with $\log c = \log w(b_\nu)$. Otherwise, $\nu^* < a_\nu$, and $\log w(\nu)$ is a decreasing function on $[a_\nu, b_\nu]$ with $\log c = \log w(a_\nu)$.

Numerical root finding such as Algorithm 3 may be used to compute the endpoints $\{\nu_1(u), \nu_2(u)\}$ of the interval $A_u(\nu) = \{\nu > 0 : \log w(x) > \log(uc)\}$. If there is a maximizer ν^* in the interval $[a_\nu, b_\nu]$, ν_1 will be found in $[a_\nu, \nu^*]$ and ν_2 will be found in $[\nu^*, b_\nu]$. If $\log w(\nu)$ is strictly increasing,

$\nu_2 = b_\nu$ and ν_1 is found in $[a_\nu, b_\nu]$. Otherwise, if $\log w(\nu)$ is strictly decreasing, $\nu_1 = a_\nu$ and ν_2 is found in $[a_\nu, b_\nu]$.

Notice that a bounded prior for ν is needed to obtain a finite maximum value c of the weight function; therefore, our choice of Uniform prior is a departure from the Exponential prior assumed by Geweke (1994). A rejection sampler similar to Geweke's can be obtained by following the original derivation with several minor differences. First, Geweke's constant A features an additional term with the Exponential hyperparameter which is now absent. Second, we take the proposal distribution to be a truncated Exponential distribution with density $q(x \mid \alpha, a_\nu, b_\nu) \propto \alpha e^{-x\alpha}\,\mathrm{I}(a_\nu \leq x \leq b_\nu)$ rather than an untruncated Exponential distribution $q(x \mid \alpha) \propto \alpha e^{-x\alpha}$. Constraining $\alpha = 1/\nu$, let ν^* be the value of ν which maximizes the ratio $f(\nu \mid \boldsymbol{s}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y})/q(\nu \mid \alpha, a_\nu, b_\nu)$; this ν satisfies

$$\frac{n}{2}\left[\log\left(\frac{\nu}{2}\right) + 1 - \psi\left(\frac{\nu}{2}\right)\right] + \frac{1}{\nu} - A = 0. \tag{32}$$

Algorithm 4 gives a rejection sampler based on q to generate candidates and the maximized ratio to determine when to accept.

Figure 5 compares the empirical density of 100,000 draws from the direct sampler with N knots initially selected using Algorithm 4. The values $n = 200$, $a_\nu = 0.01$, and $b_\nu = 200$ are fixed and A is varied to take on values 101, 120, 200, and 400. As expected, both samplers generate draws from the same target distribution. Table 2 shows the number of rejections to obtain 100,000 draws for both samplers, now including $N \in \{5, 20, 50, 100\}$ initially selected knots. Here it is apparent that Algorithm 4 rejects on the order of ten candidates for each saved variate while the direct sampler rejects less than 1% of draws on average. However, within a practical Gibbs sampling setting, Algorithm 4 may still be faster because each step requires very little computation.

Algorithm 4 Rejection sampling based on Geweke (1994).

```
1. Let $\nu^*$ be the value of $\nu$ which satisfies (32).
2. Draw candidate $\nu$ from the truncated Exponential distribution $q(x \mid 1/\nu^*, a_\nu, b_\nu)$.
3. Draw $\omega \sim \text{Uniform}(0, 1)$ and accept $\nu$ as a draw from (30) if
       $$\omega < \frac{\left(\frac{\nu}{2}\right)^{n\nu/2}\left[\Gamma\left(\frac{\nu}{2}\right)\right]^{-n}\exp(-\nu A + \nu/\nu^*)}{\left(\frac{\nu^*}{2}\right)^{n\nu^*/2}\left[\Gamma\left(\frac{\nu^*}{2}\right)\right]^{-n}\exp(-\nu^* A + 1)};$$
   otherwise, reject $\nu$ and go to step 2.
```

Table 2 Rejected candidates to obtain 100,000 draws using Algorithm 4 and the direct sampler with N initial knots

A	Direct Sampler				Algorithm 4
	$N = 5$	$N = 20$	$N = 50$	$N = 100$	
101	608	647	589	495	841,390
120	643	605	581	496	1,049,358
200	622	575	549	523	1,173,088
400	614	564	581	533	1,273,444

Lange et al. (1989) provide a number of interesting examples of regression analyses using t-distributed errors. In particular, their Example 3 studies the relationship between two measurements of blood flow in the canine myocardium. The variable r_i measures regional myocardial blood flow from an invasive procedure, while y_i is a measurement obtained using positron emission tomography within n cases indexed $i = 1, \ldots, n$. It is assumed that

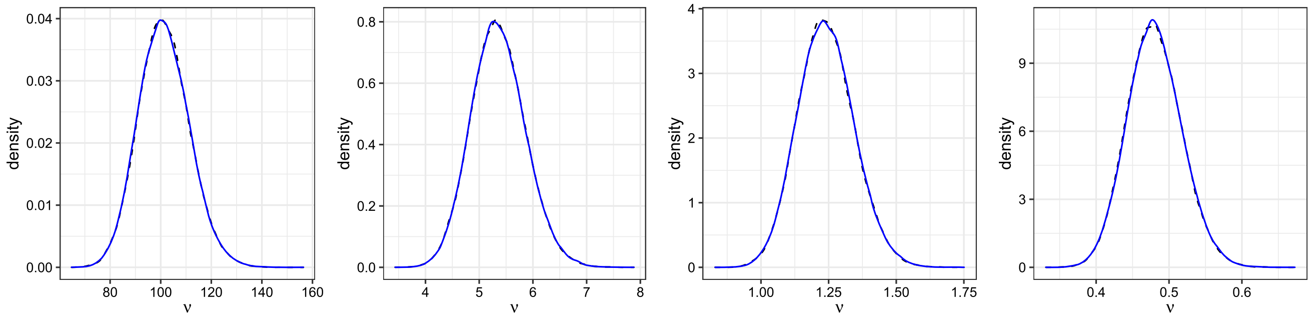
$$y_i = \mu(r_i) + \sigma\epsilon_i, \quad \epsilon_i \sim t_\nu,$$
$$\mu(r) = r\{1 - \phi_1 \exp(-\phi_2/r)\},$$

for parameters $\boldsymbol{\phi} = (\phi_1, \phi_2)$. We consider a simulated dataset based on this setting, with $n = 200$ and $r_i \overset{\text{iid}}{\sim} \text{Uniform}(0, 10)$. Data generating values of parameters $\boldsymbol{\phi}$ are to taken to be $\phi_1 = 0.746$ and $\phi_2 = 274.7$, based on estimates reported in Lange et al. (1989), while $\nu = 2$ and $\sigma = 1.25$ are taken to be the degrees of freedom and scale for the random errors. To fit linear model (29), the ith row $\boldsymbol{x}(r_i)$ of design matrix $\boldsymbol{X}$ is obtained from a cubic polynomial basis using the bs function in the R splines package (R Core Team 2022). We apply the Gibbs sampler with direct sampling to draw ν using $N = 30$ initial knots. A chain of 10,000 iterations is computed with first 5,000 discarded as as burn-in sample. Hyperparameters are taken to be $a_\nu = 0.01$, $b_\nu = 200$, $a_\sigma = 1$, $b_\sigma = 1$, and $\sigma_\beta^2 = 100$. Table 3 summarizes the saved draws of $\boldsymbol{\theta}$, while Figure 6 compares the fitted function to the true function $\mu(r)$. The model appears to be capturing the data-generating values of σ, ν, and $\mu(r)$ appropriately. We note that total sampling time was 12.2 seconds, of which 10.3 seconds was spent drawing ν with the direct sampler. There were 260 rejections in the 10,000 draws of ν.

Hosszejni (2021) presents a recent survey for Bayesian inference of ν. Here it is noted that the approach of Geweke (1994)—considered in the present section—works well for small ν but mixing tends to worsen for larger ν. Therefore, other sampling strategies are recommended when ν may be larger.

5 Conclusion

The density $p(u)$ which arises in direct sampling (Walker et al. 2011) is monotone, nonincreasing on [0, 1], and subject to sudden jumps. This motivated us to consider step functions to approximate $p(u)$. Useful samplers may be obtained for univariate target distributions where $A_u = \{x \in \Omega : w(x) >$



(a) $A = 101$.

(b) $A = 120$.

(c) $A = 200$.

(d) $A = 400$.

Fig. 5 Empirical density of 100,000 draws from (30) with sample size $n = 200$ and A specified in the subcaption. Dashed black curve represents Algorithm 4 and solid blue curve represents direct sampler with $N = 5$ initial knots

Table 3 Summary of posterior distribution

	Mean	SD	2.5%	97.5%
β_1	-0.7677	0.5734	-1.8974	0.3453
β_2	3.9061	1.4799	1.0359	6.7897
β_3	8.4826	0.7902	6.8905	10.0353
β_4	10.4723	0.8227	8.8485	12.0764
σ^2	1.4777	0.3002	0.9868	2.1660
ν	2.3409	0.5045	1.5891	3.5602

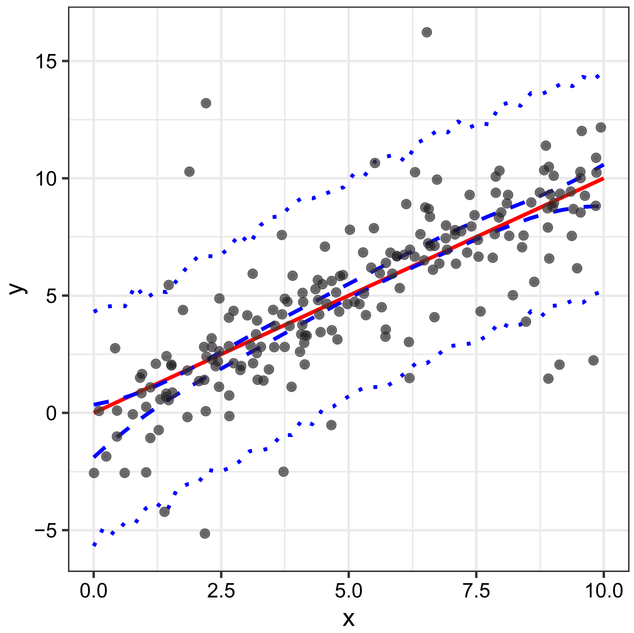


Fig. 6 The true function $\mu(r)$ (solid red curve), pointwise 95% credible interval for $\mu(r)$ (dashed blue curve), and pointwise 95% interval for posterior predictive distribution of $r \in [0, 10]$, $y_1, \ldots, y_n$ (dotted blue line), for $r \in [0, 10]$. Observed $y_1, \ldots, y_n$ are shown as block dots

$cu\}$ is an interval, and may further be combined with rejection sampling to generate exact draws with a small number of rejections. Care is required in the implementation of the sampler; e.g., the possibility of encountering very small magnitude floating point numbers motivates use of the geometric midpoint and carrying out many of the calculations on the log-scale. The idea may be extended to settings where A_u is a more complicated set such as a union of intervals, provided that the endpoints can be identified without too much computation.

The proposed sampler may be used with multivariate target distributions in principle, although underlying operations generally become challenging to implement. This includes characterization of the set $A_u = \{x \in \Omega : w(x) > uc\}$ for each $u \in [0, 1]$, computation of the probability $P(A_u)$, and a reliable method to generate draws from $f(x \mid u) \propto g(x)\,\mathrm{I}(x \in A_u)$. Walker et al. (2011) note that similar difficulties are encountered in multivariate slice sampling. As an example, taking $w(x)$ to be the density of k-dimensional $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the set A_u is seen to be an ellipse centered at $\boldsymbol{\mu}$. However, $P(A_u)$ represents the probability of the ellipse with respect to $g(x)$ while drawing from $f(x \mid u)$ amounts to drawing from $g(x)$ truncated to the ellipse; both operations may be nontrivial to compute and require consideration which is specific to $g(x)$. Further investigation into multivariate targets may be an interesting topic for future work. For now, we consider the method to be more suitable to univariate targets.

We note that a probabilistic programming platform such as Stan could be considered for the examples presented in Section 4. Especially in Sections 4.2 and 4.3, the use of Stan is compelling as it can be used in lieu of Gibbs sampling without the need to develop customized sampling code. An efficient coding of the CAR model may be needed to achieve good performance in Stan (e.g., Joseph 2016). To sample from CMP, as in Section 4.1, by a method based on Hamiltonian Monte Carlo requires the usual assumption of a continuous target distribution to be relaxed (Zhou 2020; Nishimura et al.

2020). Inference on CMP parameters in this setting could be considered in Stan, rather than the exchange algorithm, which would not require sampling from CMP itself. A Stan implementation of the CMP distribution may require care due to the volatility of the normalizing constant. Although methods such as Hamiltonian Monte Carlo are available for exploration of complicated multidimensional posteriors, Gibbs sampling is still heavily used in the literature and in practice. In a Gibbs sampler, one might encounter a non-standard conditional where there is not a clear data augmentation strategy, or where it is preferred to avoid Metropolis steps when tuning may be a concern to ensure adequate mixing of the overall MCMC chain. When such conditionals are suitably handled by the proposed direct sampler, it is possible to obtain exact draws with a low probability of rejection and little tuning needed by the analyst. The Bayesian examples in Section 4 are intended to show the proposed sampler in this light.

Acknowledgements The author is grateful to Drs. Scott Holan, Kyle Irimata, Ryan Janicki, and James Livsey at the U.S. Census Bureau for discussions which motivated this work, and to the associate editor and two anonymous referees for their attention to the manuscript.

References

Achddou, J., Lam-Weil, J., Carpentier, A., Blanchard, G.: A minimax near-optimal algorithm for adaptive rejection sampling. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory, volume 98 of Proceedings of Machine Learning Research*, pages 94–126. PMLR, 22–24 Mar 2019

Ahrens, J.H.: Sampling from general distributions by suboptimal division of domains. Grazer Math. Berichte **319**, 20 (1993)

Ahrens, J.H.: A one-table method for sampling from continuous and discrete distributions. Computing **54**(20), 127–146 (1995)

Alzer, H.: On some inequalities for the gamma and psi functions. Math. Computat. **66**(217), 373–389 (1997)

Banerjee, S., Roy, A.: Linear Algebra and Matrix Analysis for Statistics. CRC, Chapman and Hall (2014)

Benson, A., Friel, N.: Bayesian inference, model selection and likelihood estimation using fast rejection sampling: the Conway-Maxwell-Poisson distribution. Bayes. Anal. **16**(3), 905–931 (2021)

Braun, M. and Damien, P.: (2011) Generalized direct sampling for hierarchical Bayesian models. https://arxiv.org/abs/1108.2245

Braun, M., Damien, P.: Scalable rejection sampling for Bayesian hierarchical models. Market. Sci. **35**(3), 427–444 (2016)

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: a probabilistic programming language. J. Statist. Soft. **76**(1), 1–32 (2017)

Chanialidis, C., Evers, L., Neocleous, T., Nobile, A.: Efficient Bayesian inference for COM-Poisson regression models. Statist. Comput. **23**, 595–608 (2018)

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. The MIT Press: 3rd edition. (2009)

Cressie, N.: Statistics for Spatial Data. Wiley, New Jersey (1991)

Devroye, L.: Non-Uniform Random Variate Generation. Springer, London (1986)

Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Phys. Lett. B **195**(2), 216–222 (1987)

Erraqabi, A., Valko, M., Carpentier, A., Maillard, O.: Pliable rejection sampling. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages* 2121–2129, New York, USA, 20–22 Jun 2016. PMLR

Eddelbuettel, D.: Seamless R and C++ Integration with Rcpp. Springer, New York (2013)

Evans, M., Swartz, T.: Random variable generation using concavity properties of transformed densities. J. Computat. Graph. Statist. **7**(4), 514–528 (1998)

Gaunt, R.E., Iyengar, S., Olde Daalhuis, A.B., Simsek, B.: An asymptotic expansion for the normalizing constant of the Conway-Maxwell-Poisson distribution. Ann. Instit. Statist. Math. **71**, 163–180 (2019)

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, 3rd edn. CRC Press, Chapman and Hall (2013)

Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Patt. Analy. Mach. Intell. **6**(6), 721–741 (1984)

Geweke, J.: Priors for macroeconomic time series and their application. Econom. Theory **10**(3–4), 609–632 (1994)

Gilks, W.R., Wild, P.: Adaptive rejection sampling for Gibbs sampling. J. Royal Statist. Soc. Ser. C (Appl. Statist.) **41**(2), 337–348 (1992)

Gilks, W.R., Best, N.G., Tan, K.K.C.: Adaptive rejection Metropolis sampling within Gibbs sampling. J. Royal Statist. Soc. Ser. C (Appl. Statist.) **44**(4), 455–472 (1995)

Görür, D., Teh, Y.-W.: Concave-convex adaptive rejection sampling. J. Computat. Graph. Statist. **20**(3), 670–691 (2011)

Guikema, S.D., Goffelt, J.P.: A flexible count data regression model for risk analysis. Risk Analy. **28**(1), 213–223 (2008)

Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**(1), 97–109 (1970)

Hoffman, M.D., Gelman, A.: The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. **15**(47), 1593–1623 (2014)

Holden, L., Hauge, R., Holden, M.: Adaptive independent Metropolis-Hastings. Ann. Appl. Probab. **19**(1), 395–413 (2009)

Hosszejni, D.: Bayesian estimation of the degrees of freedom parameter of the Student-*t* distribution—a beneficial re-parameterization, (2021). https://arxiv.org/abs/2109.01726

Joseph, M. mbjoseph/carstan: First release, December (2016). https://doi.org/10.5281/zenodo.210407

Lange, K.L., Little, R.J.A., Taylor, J.M.G.: Robust statistical modeling using the t distribution. J. Am. Statist. Assoc. **84**(408), 881–896 (1989)

Lange, K.: Numerical Analysis for Statisticians, 2nd edn. Springer, London (2010)

Lee, D.: CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. J. Statist. Soft. **55**(13), 1–24 (2013)

Levin, D.A., Peres, Y.: Markov Chains and Mixing Times. American Mathematical Society, 2nd edn, (2017)

Lee, D.: *CARBayesdata: Data Used in the Vignettes Accompanying the CARBayes and CARBayesST Packages*, (2020). URL https://CRAN.R-project.org/package=CARBayesdata. R package version 2.2

Martino, L., Míguez, J.: A generalization of the adaptive rejection sampling algorithm. Statist. Comput. **21**(4), 633–647 (2011)

Murray, I., Ghahramani, Z., MacKay, D.J.C.: MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, pages 359–366, Arlington, Virginia, USA, 2006. AUAI Press. ISBN 0974903922

Martino, L., Read, J., Luengo, D.: Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling. IEEE Trans. Sign. Process. **63**(12), 3123–3138 (2015)

Martino, L., Luengo, D., Míguez, J.: Independent Random Sampling Methods. Springer International Publishing, Cham (2018)

Møller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K.: An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika **93**(2), 451–458 (2006)

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)

Neal, R.M.: Slice sampling. Ann. Statist. **31**(3), 705–767 (2003)

Nishimura, A., Dunson, D.B., Lu, J.: Discontinuous Hamiltonian Monte Carlo for discrete parameters and discontinuous likelihoods. Biometrika. **107**(2), 365–380 (2020)

R Core Team: *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,* Vienna, Austria, (2022). https://www.R-project.org/

Rivlin, T.J.: An Introduction to the Approximation of Functions. Dover, New York (1981)

Patil, G.P., Rao, C.R.: Weighted distributions and size-biased sampling with applications to wildlife populations and human families. Biometrics **34**(2), 179–189 (1978)

Salvatier, J., Wiecki, T.V., Fonnesbeck, C.: Probabilistic programming in Python using PyMC3. Peer J. Comput. Sci. **2**, e55 (2016)

Shmueli, G, Minka, TP, Kadane JB, Borle, S, Boatwright, P: A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. J. Royal Statist. Soc. Ser. C (Appl. Statist) 54(1), 127–142 (2005)

Tanner, M.A., Wong, W.H.: The calculation of posterior distributions by data augmentation. J. Amer. Statist. Associat. **82**(398), 528–540 (1987)

Walker, S.G., Laud, P.W., Zantedeschi, D., Damien, P.: Direct sampling. J. Computat. Graph. Statist. **20**(3), 692–713 (2011)

von Neumann, J.: Various techniques used in connection with random digits. In: Householder, A.S., Forsythe, G.E., Germond, H.H. (eds.) Monte Carlo Method, volume 12 of National Bureau of Standards Applied Mathematics Series, chapter 13, pp. 36–38. US Government Printing Office, Washington, DC. (1951)

Zhou, G.: Mixed Hamiltonian Monte Carlo for mixed discrete and continuous variables. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546