Rejection Sampling for Weighted Densities by Majorization

Andrew M. Raim*, **James A. Livsey**, **Kyle M. Irimata**
Center for Statistical Research and Methodology, U.S. Census Bureau
andrew.raim@census.gov andrewraim.github.io

Summary

- Rejection sampling is a classical algorithm to generate exact draws from a target distribution (von Neumann, 1951). When presented with a new target, it may be nontrivial to formulate a rejection sampler that achieves a good enough acceptance rate to be practically useful.
- We present a method for weighted target densities which operates by majorizing the weight function. This creates a natural envelope for rejection sampling, and does not necessarily require properties such as log-concavity.
- This presentation focuses on univariate targets which partition the support into intervals. The partition can be adapted to reduce rejection rates.
- Some related work is discussed on the last page.

Target Density

Objective: Generate draws from the **weighted** target density

$$f(x) = f_0(x)/\psi, \quad f_0(x) = w(x)g(x), \quad \psi = \int_\Omega w(x)g(x)d\nu(x), \quad \text{where}$$

1. Ω is the support of f,
2. $w(x) \geq 0$ is a weight function,
3. $g(x)$ is the "base distribution", a density function with $\Omega \subseteq \text{supp}\, g$,
4. ψ is a normalizing constant which may not have a convenient form,
5. ν is a dominating measure.

Proposal Distribution

- Partition Ω into N disjoint regions $\mathcal{D}_1, \ldots, \mathcal{D}_N$ and suppose $\overline{w}_j = \max_{x \in \mathcal{D}_j} w(x)$ and $\underline{w}_j = \min_{x \in \mathcal{D}_j} w(x)$ for each region $j = 1, \ldots, N$.
- This suggests an (unnormalized) density as the proposal:

$$h_0(x) = \begin{cases} \overline{w}_1 g(x) & \text{if } x \in \mathcal{D}_1, \\ \vdots & \\ \overline{w}_N g(x) & \text{if } x \in \mathcal{D}_N, \end{cases} \quad \implies \quad f_0(x) \leq h_0(x).$$

- Normalizing yields a finite mixture of truncated densities

$$h(x) = \sum_{j=1}^N \pi_j g_j(x), \quad \pi_j = \bar{\xi}_j \Big/ \sum_{\ell=1}^N \bar{\xi}_\ell, \quad \bar{\xi}_j = \overline{w}_j \, \mathrm{P}(T \in \mathcal{D}_j),$$

$$g_j(x) = g(x)\,\mathrm{I}(x \in \mathcal{D}_j) \Big/ \mathrm{P}(T \in \mathcal{D}_j), \quad T \sim g.$$

Rejection Sampling

We consider a standard rejection sampling algorithm with h as the proposal.

1. Draw u from Uniform$(0, 1)$.
2. Draw x from proposal h.
3. If $u \leq f_0(x)/h_0(x)$, accept x as a draw from f; otherwise return to Step 1.

Some properties of the sampler.

- The probability of rejecting each proposed draw is $1 - \psi/a$, where ψ and $a = \sum_{j=1}^N \bar{\xi}_j$ are normalizing constants for f_0 and h_0, respectively.
- A useful upper bound for the probability of rejection is

$$1 - \psi/a \leq \frac{1}{\psi} \sum_{j=1}^N \text{vol}_j, \quad \text{vol}_j = (\overline{w}_j - \underline{w}_j)\, \mathrm{P}(T \in \mathcal{D}_j).$$

- We will refer to vol$_j$ as the "volume" for the jth region, and $\sum_{j=1}^N \text{vol}_j$ as the volume for the proposal. An efficient proposal will have small volumes without N too large.

Drawing from Proposal

Draws from h can be obtained from the finite mixture form:

1. Draw index j from $1, \ldots, N$ with probabilities $\pi_1, \ldots, \pi_N$.
2. Draw x from the distribution g truncated to $\mathcal{D}_j$.

Step 2 is straightforward for the univariate distributions in this presentation.

- Suppose regions are intervals of the form $\mathcal{D}_j = (\alpha_{j-1}, \alpha_j]$.
- Let u be a draw from Uniform(α_{j-1}, α_j).
- Let G and G^- be the cumulative distribution and quantile functions for the base distribution.
- Draw x is taken to be $G^-\big(\{G(\alpha_j) - G(\alpha_{j-1})\}u + G(\alpha_{j-1})\big)$.

Adapting the Proposal

- We decompose Ω into $\mathcal{D}_1, \ldots, \mathcal{D}_N$ before sampling, with N prescpecified.
- Select the next region j to split with probability proportional to their volumes; then bifurcate region j at its midpoint.
- Special handling is needed for intervals where one or both limits are infinite (to find a suitable bifrucation point), or where support is discrete (so that each region contains at least one support point).

Computational Details

- All computations are done in R (R Core Team, 2023).
- Calculations are kept on the log-scale, as much as possible, to accommodate numbers with very large and very small magnitudes.
- For example, Step 1 of "Drawing from Proposal" is carried out on the unnormalized log-probabilities $\log \bar{\xi}_1, \ldots, \log \bar{\xi}_N$ using the Gumbel softmax trick (e.g., Maddison et al., 2014).
- Numerical optimization of $w(x)$ is used on each $\mathcal{D}_j$ to obtain $\overline{w}_j$ and $\underline{w}_j$.

A First Example

- Consider drawing from the polynomial-normal distribution

$$f_0(x) = \underbrace{\frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}}_{g(x)} \cdot \underbrace{\prod_{\ell=1}^m (x - \lambda_\ell)(x - \bar{\lambda}_\ell)}_{w(x)}, \quad x \in \mathbb{R},$$

  from Evans and Swartz (1998), where $w(x)$ is a non-negative polynomial of degree $2m$, each λ_ℓ is a root, and $\bar{\lambda}_\ell$ is its complex conjugate.
- Let $m = 2$ with $\lambda_1 = 1 + 0.5i$ and $\lambda_2 = -3 + 0.5i$, and suppose knots $(\alpha_1, \ldots, \alpha_{N-1})$ are taken to be -4.5, -3.5, -2.5, -1.5, 0, 1.5, and 2.5.
- Figure 1 displays $w(x)$ (solid black) and the associated $\overline{w}_j(x)$ and $\underline{w}_j(x)$ (top and bottom of the blue rectangle, respectively) for the proposal. Solid blue lines are knot locations.
- Figure 2 displays $f_0(x)$ (solid black) and the proposal $h_0(x)$ (dashed blue). Volumes of each region are displayed at the top.

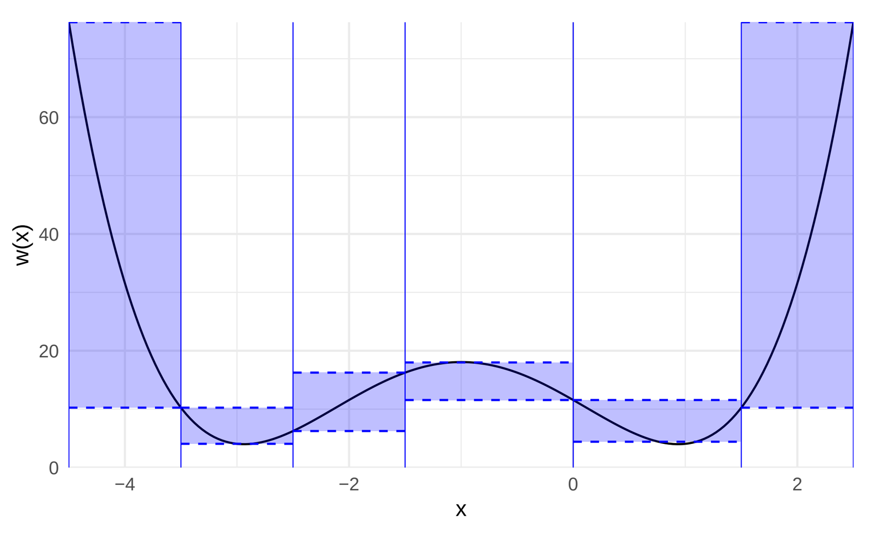

Figure 1

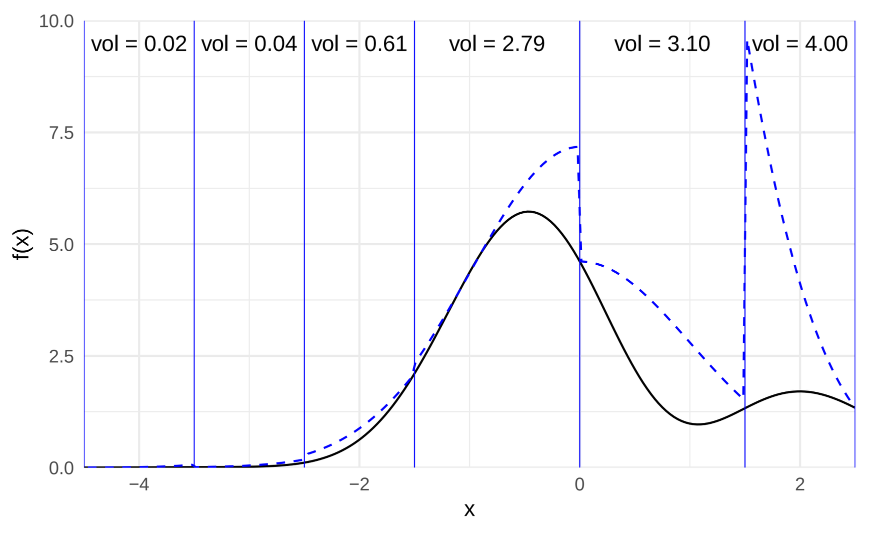

Figure 2

Rejection Sampling for Weighted Densities by Majorization

Andrew M. Raim*, **James A. Livsey**, **Kyle M. Irimata**
Center for Statistical Research and Methodology, U.S. Census Bureau
andrew.raim@census.gov   andrewraim.github.io

United States Census Bureau

Conway-Maxwell Poisson Distribution

- The Conway-Maxwell Poisson (CMP) distribution has become popular for modeling count data which may exhibit over- and/or underdispersion.
- The monograph by Sellers (2023) gives an overview of CMP and a number of recent developments. The R package COMPoissonReg (Raim and Sellers, 2022) implements basic CMP distribution functions and regression.
- A random variable X with distribution CMP(λ, ν) has probability mass function (pmf)

$$f(x) = \frac{\lambda^x}{(x!)^\nu Z(\lambda,\nu)}, \quad x = 0, 1, 2, \ldots, \quad Z(\lambda,\nu) = \sum_{x=0}^{\infty} \frac{\lambda^x}{(x!)^\nu}$$

  where $\lambda \geq 0$ and $\nu \geq 0$.
- The CMP(λ, ν) family includes some cases of interest.
  1. When $\nu = 1$, it corresponds to Poisson(λ). Here variance and mean are both λ.
  2. When $\nu < 1$, it becomes overdispersed so that the variance is larger than the mean. At the extreme $\nu = 0$, it corresponds to Geometric$(1 - \lambda)$.
  3. When $\nu > 1$, it becomes underdispersed so that the variance is smaller than the mean. As $\nu \to \infty$, it becomes Bernoulli$(\lambda/(1+\lambda))$.

Sampling

- Generating variates from CMP is non-trivial because the magnitude of $Z(\lambda, \nu)$ can vary wildly with λ and ν. The mass of the distribution can shift accordingly.
- For example, let $\lambda = 2$. If $\nu = 1$, then $Z(\lambda,\nu) = e^2$ and $\text{E}(X) = 2$. However, if $\nu = 0.05$, $Z(\lambda,\nu) \approx \exp(52{,}437.76)$ and $\text{E}(X) = 1{,}048{,}585$.
- The variate generating function `rcmp` in COMPoissonReg works by either: (1) truncating the series $Z(\lambda,\nu)$ to within a small tolerance, or (2) using an asymptotic approximation.
- Chanialidis et al. (2018) and Benson and Friel (2021) develop custom rejection sampling algorithms to generate exact CMP draws; they are used to implement the exchange algorithm for Bayesian analysis of CMP parameters.
- Raim (2023) uses the same decomposition that we now present to formulate an efficient rejection sampler, but the present method is easier to implement.

Underdispersion Case

- For the case $\nu \geq 1$, let g be the pmf of Geometric$(1/\{1+\lambda\})$ so that

$$f(x) \propto \frac{\lambda^x}{(x!)^\nu} = \underbrace{\left(\frac{\lambda}{1+\lambda}\right)^x \frac{1}{1+\lambda}}_{g(x)} \underbrace{(1+\lambda)^{x+1} \frac{\lambda^x}{(x!)^\nu}}_{w(x)}.$$

- For $\lambda = 10$ and $\nu = 1.2$, a sampler with $N = 21$ regions rejected 5 proposed draws to obtain 100,000 variates (rejection rate 0.005%).
- Figure 3 shows reduction in the log of total volume $\sum_{j=1}^N \text{vol}_j$. Figure 4 displays proportions of draws (bars) versus density values (points).

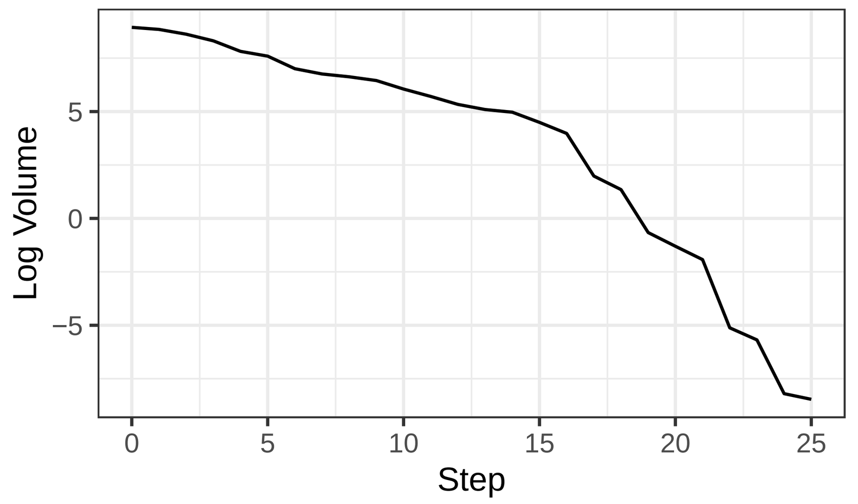

Figure 3

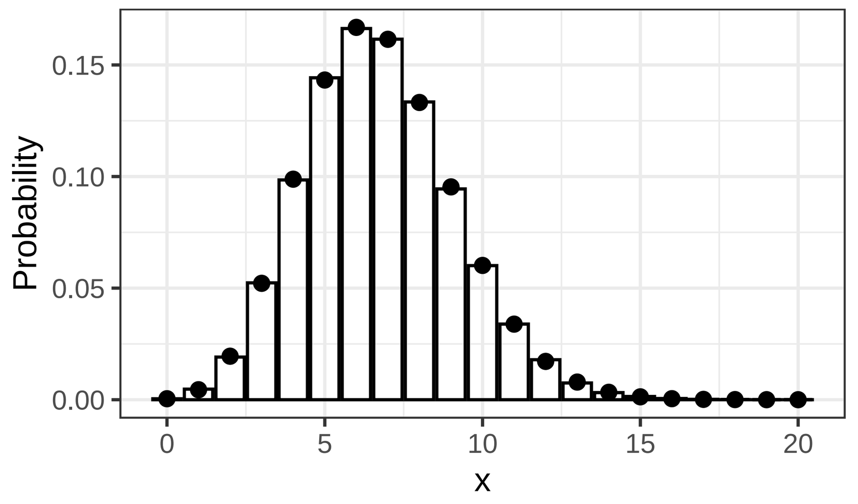

Figure 4

Overdispersion Case

- For $\nu < 1$, Geometric$(1/\{1+\lambda\})$ may be an inefficient base because its mass can be practically disjoint from CMP(λ, ν).
- Here let $\mu = \lambda^{1/\nu}$ and

$$f(x) \propto \frac{\mu^{\nu x}}{(x!)^\nu} = \underbrace{\left(\frac{\mu}{1+\mu}\right)^x \frac{1}{1+\mu}}_{g(x)} \underbrace{(1+\mu)^{x+1} \frac{\mu^{x(\nu-1)}}{(x!)^\nu}}_{w(x)}.$$

- For $\lambda = 1.5$ and $\nu = 0.05$, a sampler with $N = 101$ regions rejected 2,922 proposed draws to obtain 100,000 variates (rejection rate 2.84%).
- Figure 6: empirical density of draws (solid black) with pmf (red dashed).

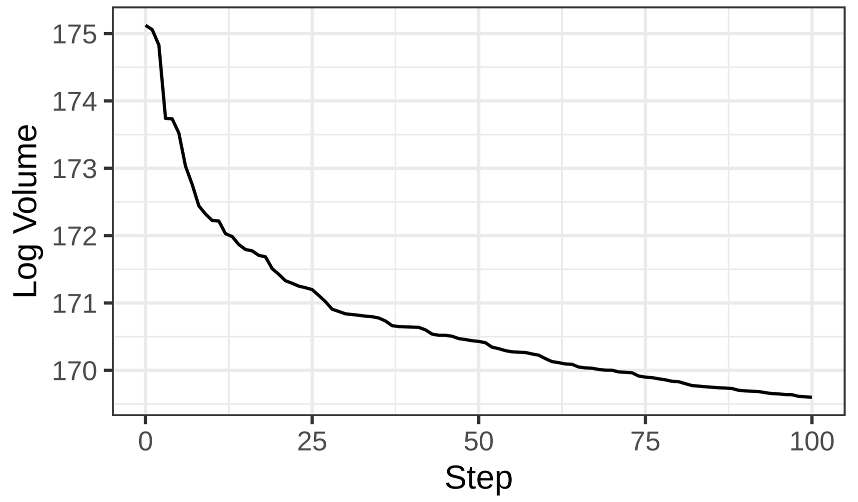

Figure 5

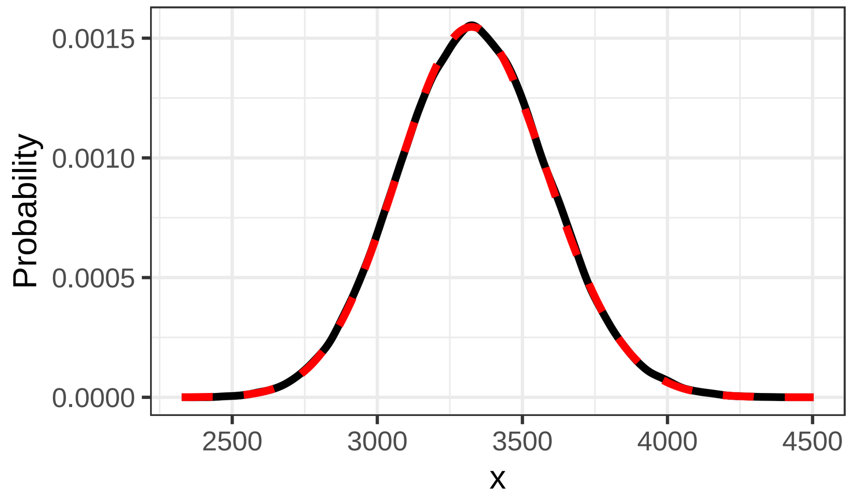

Figure 6

Acceptance Rates (%)

A brief study of acceptance % was carried out with $R = 100{,}000$,

- $\lambda \in \{0.25, 0.5, 0.75, 1, 1.25, 2, 5, 10\}$,
- $\nu \in \{0.01, 0.05, 0.5, 1, 1.5, 5, 10\}$,
- where $\lambda^{1/\nu} \leq 50{,}000$.

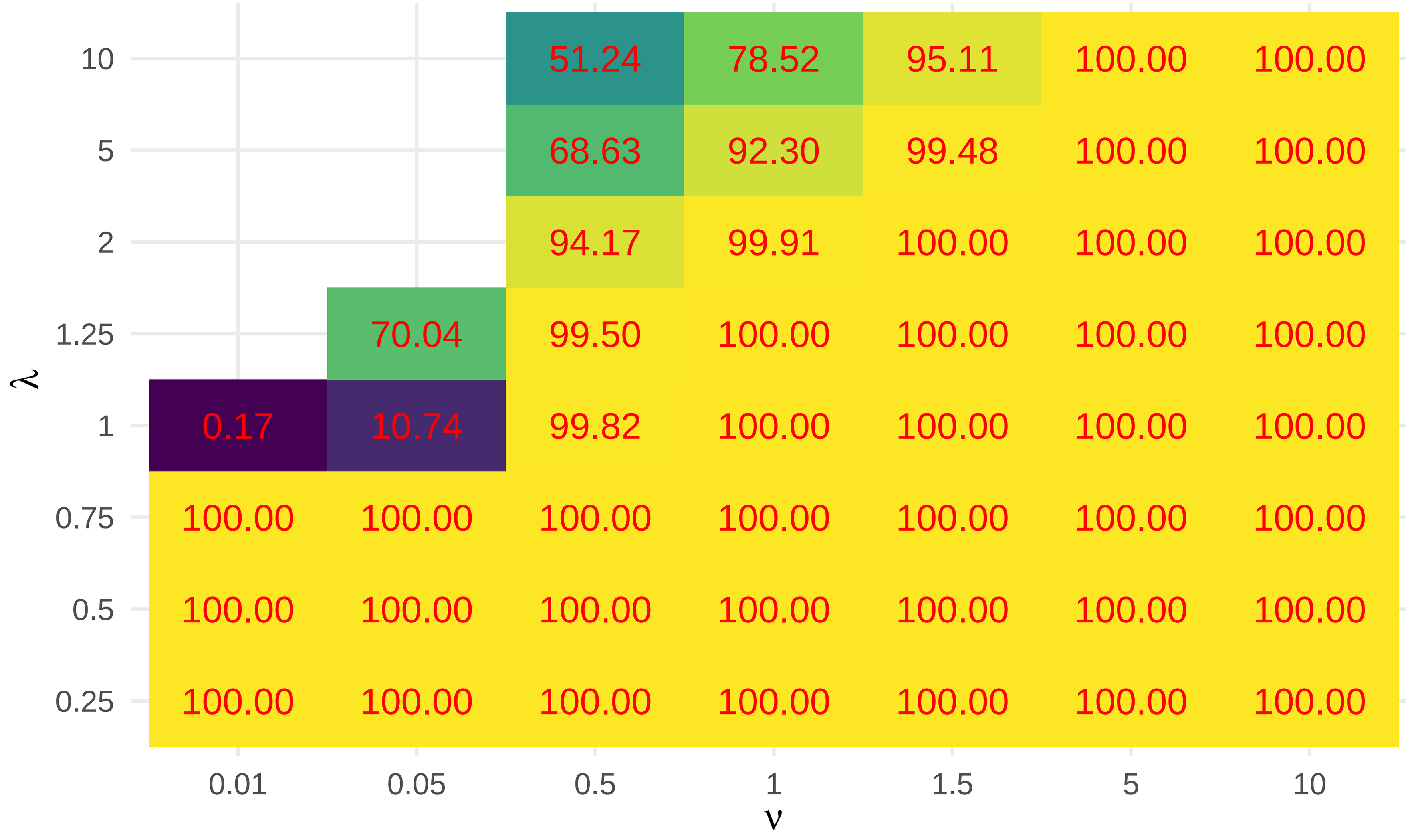

Figure 7: $N = 10$.

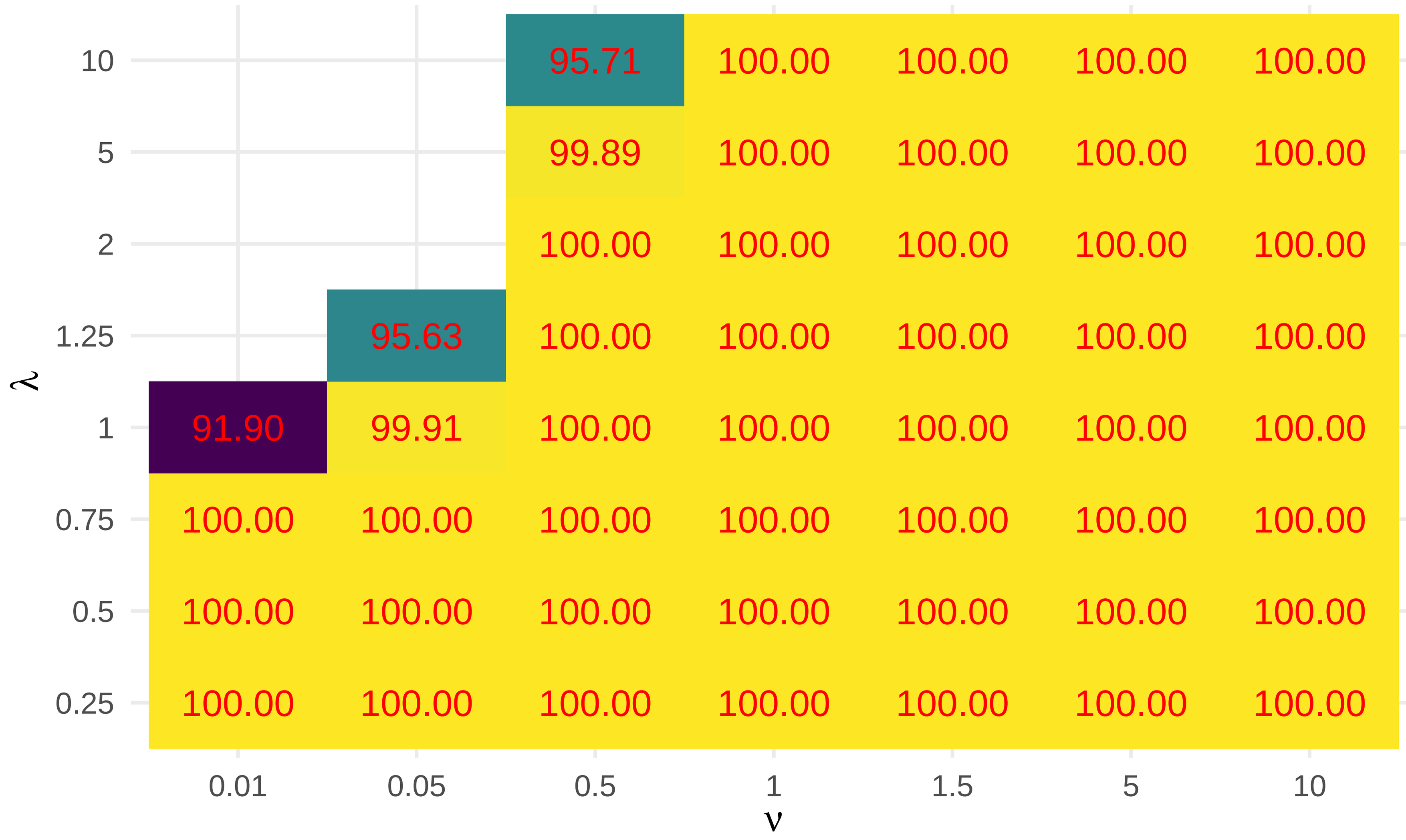

Figure 8: $N = 50$.

Rejection Sampling for Weighted Densities by Majorization

Andrew M. Raim*, **James A. Livsey**, **Kyle M. Irimata**
Center for Statistical Research and Methodology, U.S. Census Bureau
andrew.raim@census.gov   andrewraim.github.io

United States Census Bureau

von Mises Fisher Distribution

- A random variable $\boldsymbol{V}$ with von Mises Fisher distribution $\text{VMF}_d(\boldsymbol{\mu}, \kappa)$ is on the d-dimensional sphere $\mathbb{S}^d = \{\boldsymbol{v} \in \mathbb{R}^d : \boldsymbol{v}^\top \boldsymbol{v} = 1\}$ and has density

$$f_{\text{VMF}}(\boldsymbol{v}) = \frac{\kappa^{d/2-1}}{(2\pi)^{-d/2} I_{d/2-1}(\kappa)} \exp(\kappa \cdot \boldsymbol{\mu}^\top \boldsymbol{v}) \cdot \mathrm{I}(\boldsymbol{v} \in \mathbb{S}^d),$$

where parameters $\kappa > 0$ and $\boldsymbol{\mu} \in \mathbb{S}^d$ determine concentration and modal direction, respectively, and $I_\nu(x) = \sum_{m=0}^{\infty} \{m! \cdot \Gamma(m+\nu+1)\}^{-1} (\frac{x}{2})^{2m+\nu}$ is modified Bessel function of the first kind.

- A draw from $\text{VMF}_d(\boldsymbol{\mu}, \kappa)$ with $\boldsymbol{\mu} = (1, 0, \ldots, 0)$ can be obtained as

$$\boldsymbol{V}_0 = \left(\sqrt{1 - X^2} \cdot \boldsymbol{U}, X \right), \quad \text{(Ulrich, 1984)},$$

where $\boldsymbol{U} \sim \text{Uniform}(\mathbb{S}^{d-1})$ and X has density

$$f(x) = \frac{(\kappa/2)^{d/2-1} (1 - x^2)^{(d-3)/2} \exp(\kappa x)}{\sqrt{\pi} \cdot I_{d/2-1}(\kappa) \cdot \Gamma((d-1)/2)} \cdot \mathrm{I}(-1 < x < 1).$$

- Transform to $\boldsymbol{V} \sim \text{VMF}_d(\boldsymbol{\mu}, \kappa)$, for any desired $\boldsymbol{\mu}$, using $\boldsymbol{V} = \boldsymbol{Q} \boldsymbol{V}_0$ with $\boldsymbol{Q}$ an orthonormal matrix whose first column is $\boldsymbol{\mu}$.
- A draw of $\boldsymbol{U}$ can be obtained as $\boldsymbol{Z} / \sqrt{\boldsymbol{Z}^\top \boldsymbol{Z}}$ with $\boldsymbol{Z} \sim \text{N}(\boldsymbol{0}, \boldsymbol{I}_{d-1})$ (Muller, 1959); therefore, drawing $\boldsymbol{V}$ reduces to univariate generation of X.

Our Rejection Sampler

- To apply our proposed rejection sampler, consider the decomposition

$$f(x) \propto \underbrace{2(1 - x^2)^{(d-3)/2} \exp(\kappa x)}_{w(x)} \cdot \underbrace{\frac{1}{2} \mathrm{I}(-1 < x < 1)}_{g(x)}$$

so that g is the density of $\text{Uniform}(-1, 1)$.

Example

- Consider the setting $d = 3$, $\kappa = 10$, and $\boldsymbol{\mu} = (1, 0, \ldots, 0)$.
- To sample X, Figure 9 shows reduction in log-volume after adapting to $N = 101$ regions.
- To obtain $R = 50,000$ draws of X, 1,393 proposed draws were rejected (rejection rate: 2.71%).
- Figure 10 compares the empirical CDF of the X draws (solid black) to the sCDF of X (dashed red) computed by numerical integration of $f(x)$.
- The R draws of X were used to construct R draws of $\boldsymbol{V}$. Figure 12 displays these draws projected to two dimensions from the perspective of $\boldsymbol{\mu}$. Yellow bins indicate higher counts and blue bins indicate lower counts.
- Figure 11 shows the $\text{VMF}_3(\boldsymbol{\mu}, \kappa)$ density in three dimensions.

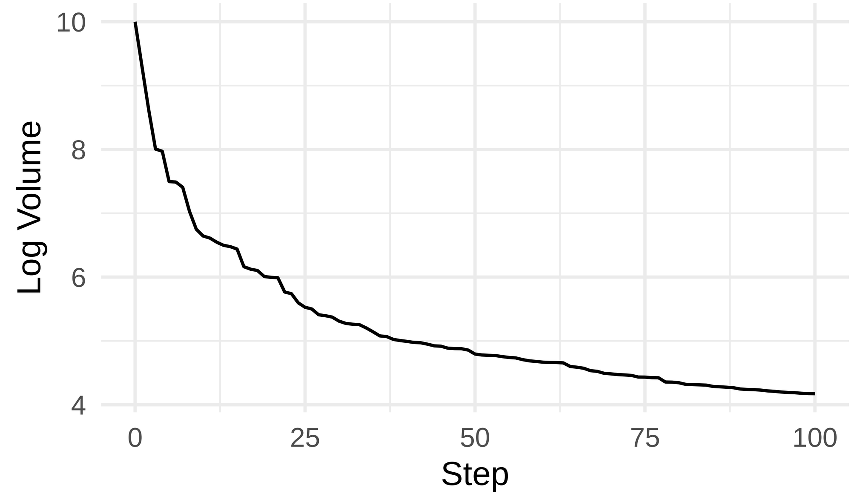

Figure 9

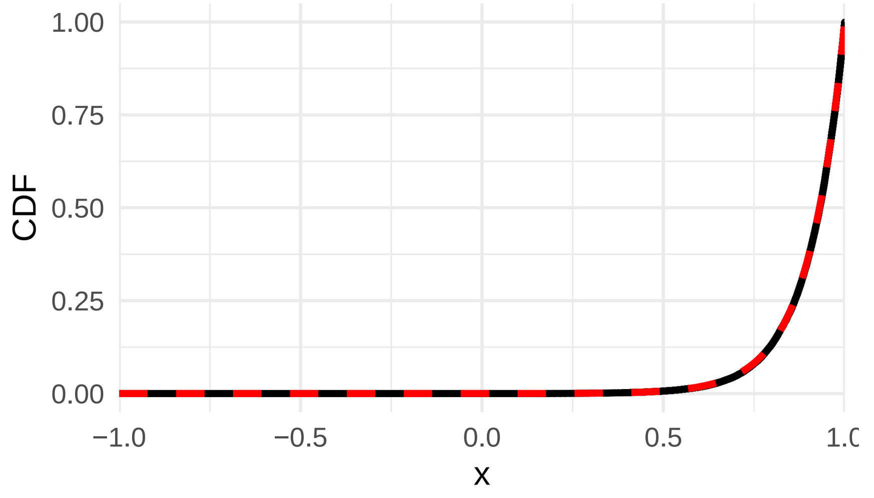

Figure 10

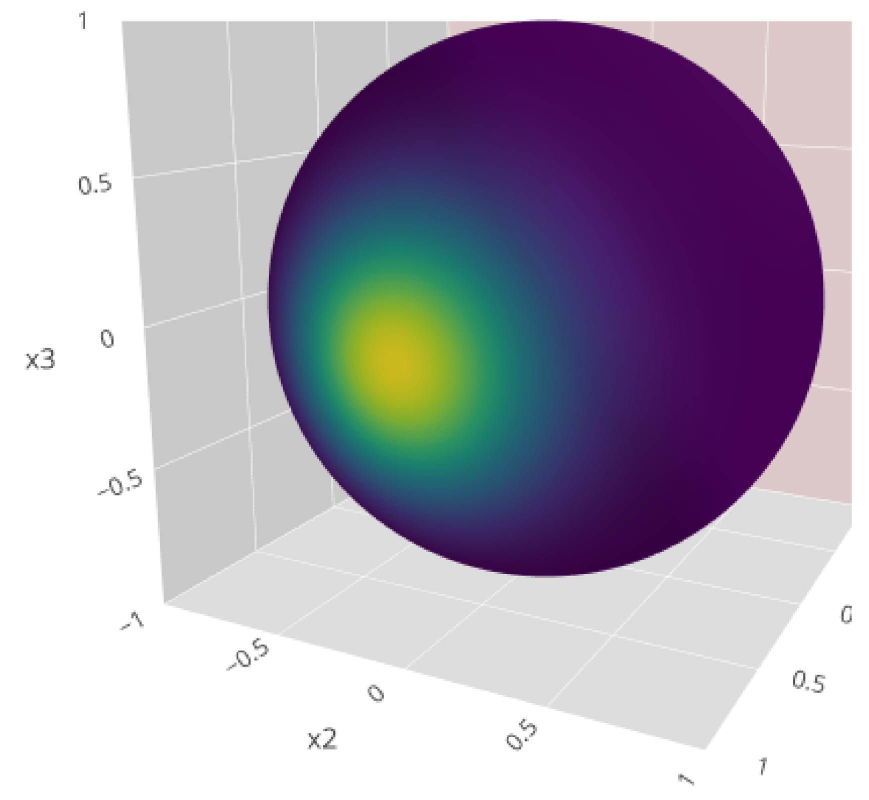

Figure 11

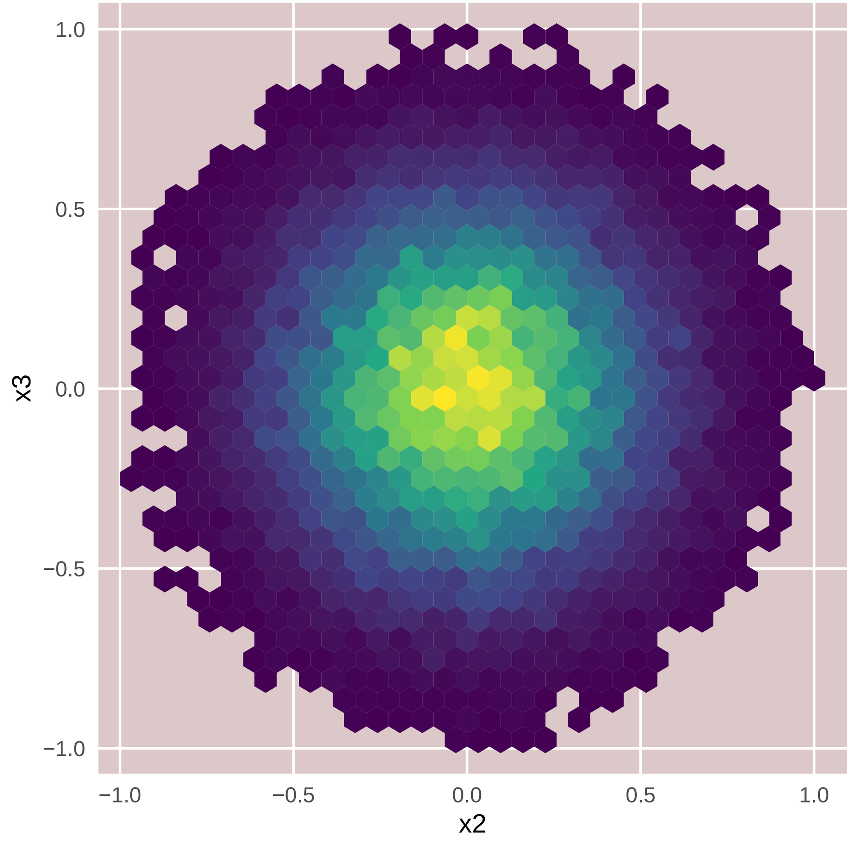

Figure 12

Ulrich & Wood's (UW) Sampler

- Ulrich (1984) and Wood (1994) develop a custom rejection sampler for X. It is still used in a number of software packages, including movMF (Hornik and Grün, 2014) and Rfast (Tsagris and Papadakis, 2018).
- Proposal is random variable $X_0 = [1 - (1 + b)Z]/[1 - (1 - b)Z]$ where $Z \sim \text{Beta}((d-1)/2, (d-1)/2)$ and b is fixed; density is

$$f_0(x \mid b) = \frac{2 \cdot b^{(d-1)/2} (1 - x^2)^{(d-3)/2}}{B(\frac{d-1}{2}, \frac{d-1}{2}) \cdot [(1 + b) - (1 - b)x]^{d-1}}, \quad x \in (-1, 1). \quad (1)$$

- To obtain the smallest M such that $f(x)/\{M f_0(x \mid b)\} \leq 1$ for all $x \in (-1, 1)$:

$$x_* = \frac{1 - b_*}{1 + b_*}, \quad b_* = \frac{-2\kappa + \sqrt{4\kappa^2 + (d-1)^2}}{d - 1}.$$

- The algorithm proceeds with $c = \kappa x_* + (d - 1) \log(1 - x_*^2)$:
  1. Draw x from proposal (1) and u from $\text{Uniform}(0, 1)$.
  2. Accept x as a draw from the target if $\log u < \kappa x + (d-1) \log(1 - x \cdot x_*) - c$; otherwise reject x and return to step 1.

Rejection Rates (%)

- The table below presents a small study comparing rejection rates (as percentages) from the UW sampler with ours. Displayed is percent of rejected proposals to obtain 50,000 draws.
- The UW sampler rejects more frequently as κ is increased and d is small. It is very fast in practice but developing it involved a clever transformation.
- Rejection rates for our sampler reduce slowly with N in some cases. E.g., when $d = 2$, $w(x)$ is a bowl-shaped function with steep sides. A different choice of majorization than constants $\overline{w}_j$ may be more effective here.

	UW Sampler					Our Sampler ($N = 100$)				
d	0.1	0.5	1	5	$\kappa = 10$	0.1	0.5	1	5	10
2	0.28	4.99	13.33	30.31	32.39	6.21	8.09	8.19	7.10	6.81
3	0.10	1.92	6.11	24.82	28.80	0.16	0.65	1.30	2.52	2.66
4	0.04	0.99	3.45	21.03	26.02	1.04	1.11	1.44	2.47	2.46
5	0.03	0.56	2.26	17.72	23.86	1.52	1.56	1.73	2.42	2.72
10	0.01	0.14	0.56	8.40	16.40	2.52	2.32	2.32	2.64	2.74
20	0.00	0.05	0.13	2.84	8.10	2.87	2.53	2.69	2.61	2.81
50	0.00	0.01	0.02	0.45	1.84	2.87	3.06	2.71	2.96	2.96

Rejection Sampling for Weighted Densities by Majorization

Andrew M. Raim*, **James A. Livsey**, **Kyle M. Irimata**
Center for Statistical Research and Methodology, U.S. Census Bureau
andrew.raim@census.gov andrewraim.github.io

Gaussian Process Regression

- Suppose $\mu(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a function whose form may be unknown, and $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, are data where y_i is a noisy observation of $\mu(\boldsymbol{x}_i)$.
- Consider the GP model

$$\begin{aligned} y_i &= \mu(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad i = 1, \ldots, n, \\ \mu &\sim \mathrm{GP}(0, k(\cdot, \cdot)), \quad \sigma^2 \sim \mathrm{Gamma}(a_\sigma, b_\sigma), \end{aligned}$$

shape a_σ, rate b_σ, and covariance kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\}$.

- Likelihood portion of the model in vector form is

$$\boldsymbol{y} = \mu(\boldsymbol{X}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \quad \mu(\boldsymbol{X}) \sim \mathrm{N}\big(\boldsymbol{0}, k(\boldsymbol{X}, \boldsymbol{X})\big).$$

Rejection Sampler

- Using the proposed rejection sampler, we can draw exactly from the posterior distribution $[\sigma^2 \mid \boldsymbol{y}]$ without MCMC.
- Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the spectral decomposition of $k(\boldsymbol{X}, \boldsymbol{X})$ with $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$.
- Using $\sigma^2 \boldsymbol{I} + k(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{U}[\sigma^2 \boldsymbol{I} + \boldsymbol{\Lambda}]\boldsymbol{U}^\top$, we can transform the marginal likelihood of $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} + k(\boldsymbol{X}, \boldsymbol{X}))$ to $\boldsymbol{z} = \boldsymbol{U}^\top \boldsymbol{y}$ where $z_i \stackrel{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2 + \lambda_i)$.
- Let weight function be the unnormalized posterior with respect to $\boldsymbol{z}$:

$$\begin{aligned} \log w(\sigma^2) = &-\frac{1}{2}\sum_{i=1}^n \log(\sigma^2 + \lambda_i) - \frac{1}{2}\sum_{i=1}^n \frac{z_i^2}{\sigma^2 + \lambda_i} \\ &+ (a_\sigma - 1)\log \sigma^2 - b_\sigma \sigma^2. \end{aligned} \tag{2}$$

- We take base distribution g as the density of $\mathrm{Uniform}(0, 1000)$.
- Form (2) avoids repeating large matrix operations in the sampler, though these may be needed to initially obtain $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$.
- This sampler can be used with other priors on σ^2 and other covariance kernels with fixed hyperparameters.

Prediction

- We can sample from the posterior predictive distribution for (potentially new) inputs $\boldsymbol{X}_0 = (\boldsymbol{x}_{01} \cdots \boldsymbol{x}_{0n_0})^\top$.
  1. Draw $\sigma^{2(r)}$, $r = 1, \ldots, R$, from posterior (2).
  2. Draw $\mu(\boldsymbol{X}_0)$ from $[\mu(\boldsymbol{X}_0) \mid \boldsymbol{y}, \sigma^2]$ for each $\sigma^2 = \sigma^{2(r)}$.
- The distribution $[\mu(\boldsymbol{X}_0) \mid \sigma^2, \boldsymbol{y}]$ can be obtained from

$$[\boldsymbol{y}, \mu(\boldsymbol{X}_0) \mid \sigma^2] \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma^2 \boldsymbol{I} + k(\boldsymbol{X}, \boldsymbol{X}) & k(\boldsymbol{X}, \boldsymbol{X}_0) \\ k(\boldsymbol{X}_0, \boldsymbol{X}) & k(\boldsymbol{X}_0, \boldsymbol{X}_0) \end{bmatrix} \right)$$

as $\mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where

$$\begin{aligned} \boldsymbol{\mu}_0 &= k(\boldsymbol{X}_0, \boldsymbol{X})[k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}]^{-1}\boldsymbol{y}, \\ \boldsymbol{\Sigma}_0 &= k(\boldsymbol{X}_0, \boldsymbol{X}_0) - k(\boldsymbol{X}_0, \boldsymbol{X})[k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}]^{-1} k(\boldsymbol{X}, \boldsymbol{X}_0). \end{aligned}$$

Example

- To learn about the sinc function $\mu(x) = \sin(\pi x)/\{\pi x\}$, we observe $n = 25$ outcomes $y_i = \mu(x_i) + \epsilon_i$ with $\epsilon_i \stackrel{\text{iid}}{\sim} \mathrm{N}(0, 0.1^2)$, x_i on an evenly spaced grid in $[-6, 6]$. Let $a_\sigma = 2$ and $b_\sigma = 1/2$.
- Figure 15 shows the observed data and $\mu(x)$.
- Figure 13 shows decrease in log-volume of the rejection sampler over 100 adapt steps. With $N = 101$ regions, 1,502 proposed draws were rejected to obtain $R = 50,000$ (rejection rate 2.92%)
- Figure 14 compares the empirical distribution of draws from rejection sampler (solid blue) to R draws computed via Stan (Carpenter et al., 2017) with NUTS (dashed black).
- Figure 16 displays posterior predictive mean of $\mu(x)$ (blue curve), for x on a fine grid on $[-6, 6]$, and associated 95% pointwise interval from 0.025 and 0.975 quantiles (blue shaded area).

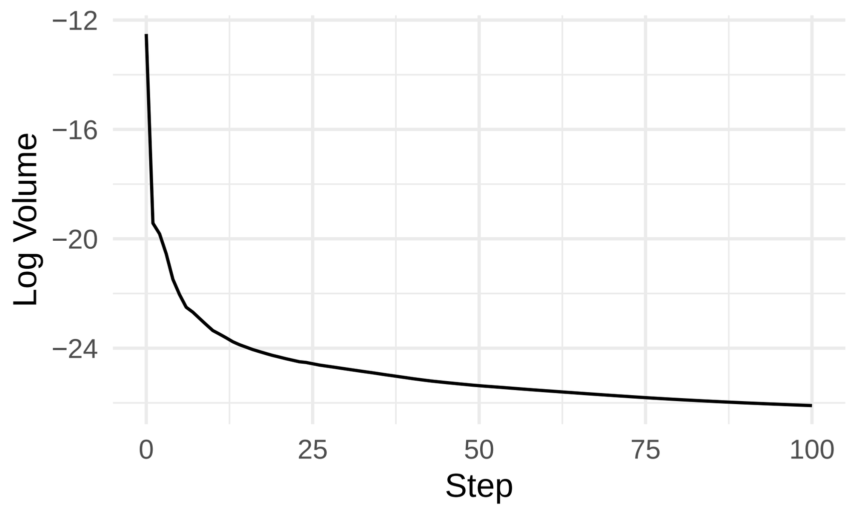

Figure 13

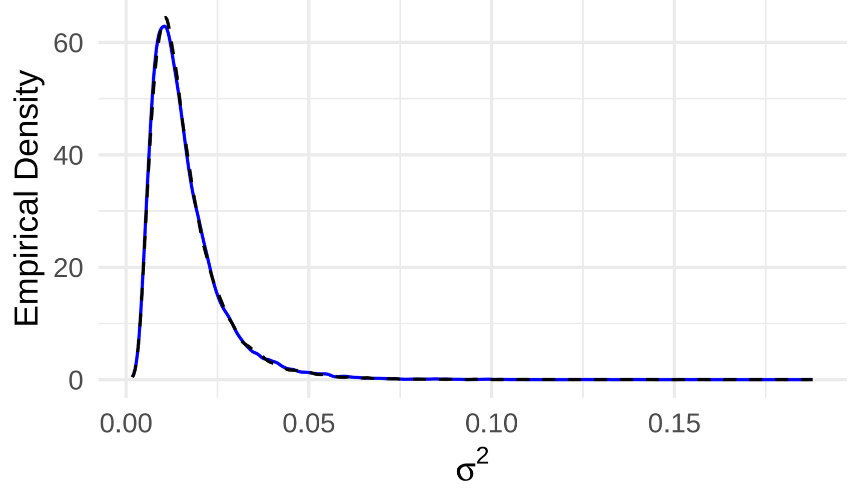

Figure 14

Example (cont'd)

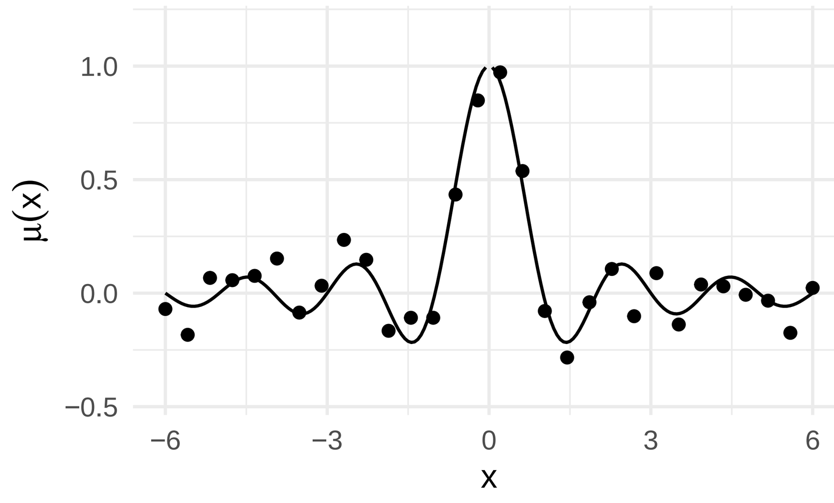

Figure 15

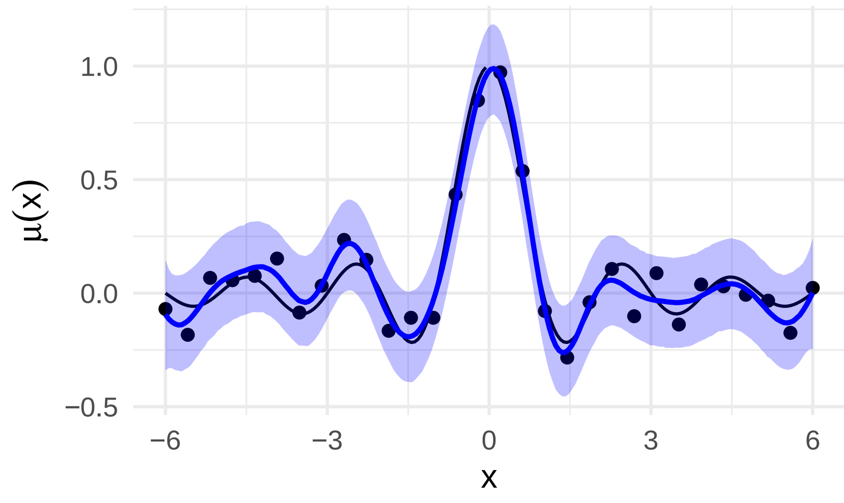

Figure 16

Spatial Regression

- The spatial linear regression model presented in Chapter 6 of Banerjee et al. (2015) is an application of the GP.
- Suppose $\boldsymbol{x}_i$ are locations on a spatial domain with fixed covariate $\boldsymbol{s}(\boldsymbol{x}_i) \in \mathbb{R}^m$ and observation y_i, $i = 1, \ldots, n$, and

$$\begin{aligned} y_i &= \boldsymbol{s}(\boldsymbol{x}_i)^\top \boldsymbol{\beta} + \zeta(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2), \\ \zeta &\sim \mathrm{GP}(0, k(\cdot, \cdot)), \quad \boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}), \quad \sigma^2 \sim \mathrm{Gamma}(a_\sigma, b_\sigma). \end{aligned}$$

- With kernel $k(\cdot, \cdot)$ completely specified, we can draw from the exact posterior using the proposed rejection sampler.
- Marginally, $[\boldsymbol{y} \mid \sigma^2] \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} + \sigma_\beta^2 \boldsymbol{S}\boldsymbol{S}^\top + k(\boldsymbol{X}, \boldsymbol{X}))$, where $\boldsymbol{S}$ is a matrix with $\boldsymbol{s}(\boldsymbol{x}_i)$ as the ith row.
- Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the spectral decomposition of $\sigma_\beta^2 \boldsymbol{S}\boldsymbol{S}^\top + k(\boldsymbol{X}, \boldsymbol{X})$, and consider the posterior with respect to data $\boldsymbol{z} = \boldsymbol{U}^\top \boldsymbol{y}$ where $z_i \stackrel{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2 + \lambda_i)$ as before.
- Draws of $\boldsymbol{\beta}$ can be recovered from $[\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y}] \propto [\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2] \cdot [\boldsymbol{\beta}]$ using conjugacy of $\mathrm{N}(\boldsymbol{y} \mid \boldsymbol{S}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I} + k(\boldsymbol{X}, \boldsymbol{X}))$ and $\mathrm{N}(\boldsymbol{\beta} \mid \boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I})$.
- Draws of $\zeta(\boldsymbol{X}_0)$ from posterior predictive distribution $[\zeta(\boldsymbol{X}_0) \mid \boldsymbol{y}]$ may be obtained using $[\zeta(\boldsymbol{X}_0) \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{y}] \equiv \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$,

$$\begin{aligned} \boldsymbol{\mu}_0 &= k(\boldsymbol{X}_0, \boldsymbol{X})[k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}]^{-1}(\boldsymbol{y} - \boldsymbol{S}\boldsymbol{\beta}), \\ \boldsymbol{\Sigma}_0 &= k(\boldsymbol{X}_0, \boldsymbol{X}_0) - k(\boldsymbol{X}_0, \boldsymbol{X})[k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}]^{-1} k(\boldsymbol{X}, \boldsymbol{X}_0). \end{aligned}$$

- The R package spBayes (Finley et al., 2007) considers a fully conjugate variation of this model with σ^2 fixed. Also, full Bayesian treatments of more general variants with MCMC via Metropolis-Hastings.

Rejection Sampling for Weighted Densities by Majorization

Andrew M. Raim*, **James A. Livsey**, **Kyle M. Irimata**
Center for Statistical Research and Methodology, U.S. Census Bureau
andrew.raim@census.gov andrewraim.github.io

Additional Notes

- The approach in this work—a finite mixture proposal based on disjoint regions—is a vertical strip method, which is discussed in Devroye (1986, Chapters II and VIII) and Martino et al. (2018, Chapter 3). Martino et al. refer to this as "Ahrens method".
- A weighted density form presents an opportunity for improved efficiency with vertical strips. Evaluating and drawing from the reweighted proposal can be more involved, but is tractable in the presented examples.
- Raim (2023) uses some of these ideas to implement a hybrid of the direct sampling method of Walker et al. (2011) and rejection sampling. The present method is more straightforward and easier to implement.
- Adaptive rejection sampling methods build an envelope using rejected draws Martino et al. (2018, Chapter 7).
  1. The ARS algorithm produces independent draws but requires the target to be log-concave.
  2. Adaptive Rejection Metropolis Sampling (ARMS) removes the log-concave restriction; however, it produces a chain of non-independent draws and proposal is not guaranteed to converge to the target as rejections increase.
  3. The Independent Doubly Adaptive Rejection Metropolis Sampling (IA2RMS) algorithm addresses the ARMS convergence issue which also reduces dependence.

References: Rejection Sampling

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer, 1986.

M. Evans and T. Swartz. Random variable generation using concavity properties of transformed densities. *Journal of Computational and Graphical Statistics*, 7(4):514–528, 1998.

Chris J. Maddison, Daniel Tarlow, and Tom Minka. A∗ sampling. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Luca Martino, David Luengo, and Joaquín Míguez. *Independent Random Sampling Methods*. Springer, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Andrew M. Raim. Direct sampling with a step function. *Statistics and Computing*, 33(1), 2023.

John von Neumann. Various techniques in connection with random digits. In A.S. Householder, G.E. Forsythe, and H.H. Germond, editors, *Monte Carlo Methods*, National Bureau of Standards Applied Mathematics Series, pages 36–38. U.S. Government Printing Office, Washington, DC, 1951.

Stephen G. Walker, Purushottam W. Laud, Daniel Zantedeschi, and Paul Damien. Direct sampling. *Journal of Computational and Graphical Statistics*, 20(3):692–713, 2011.

References: Conway-Maxwell Poisson

Alan Benson and Nial Friel. Bayesian Inference, Model Selection and Likelihood Estimation using Fast Rejection Sampling: The Conway-Maxwell-Poisson Distribution. *Bayesian Analysis*, 16(3):905–931, 2021.

Charalampos Chanialidis, Ludger Evers, Tereza Neocleous, and Agostino Nobile. Efficient Bayesian inference for COM-Poisson regression models. *Statistics and Computing*, 23:595–608, 2018.

Andrew M. Raim and Kimberly F. Sellers. COMPoissonReg: Usage, the normalizing constant, and other computational details. Research Report Series: Computing #2022-01, Center for Statistical Research and Methodology, U.S. Census Bureau, 2022. URL https://www.census.gov/library/working-papers/2022/adrm/RRC2022-01.html.

Kimberly F. Sellers. *The Conway-Maxwell-Poisson Distribution*. Cambridge University Press, 2023.

References: von Mises Fisher

Kurt Hornik and Bettina Grün. movMF: An R package for fitting mixtures of von Mises-Fisher distributions. *Journal of Statistical Software*, 58(10): 1–31, 2014.

Mervin E. Muller. A note on a method for generating points uniformly on N-dimensional spheres. *Communications of the ACM*, 2(4):19–20, 1959.

Michail Tsagris and Manos Papadakis. Forward regression in R: From the extreme slow to the extreme fast. *Journal of Data Science*, 16(4):771–780, 2018.

Gary Ulrich. Computer generation of distributions on the m-sphere. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 33(2):158–163, 1984.

Andrew T. A. Wood. Simulation of the von Mises Fisher distribution. *Communications in Statistics—Simulation and Computation*, 23(1):157–164, 1994.

References: Gaussian Process

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC Press, 2nd edition, 2015.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

Andrew O. Finley, Sudipto Banerjee, and Bradley P. Carlin. spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1–24, 2007.