


Global Likelihood Sampler for Multimodal Distributions

Si-Yu Yi, Ze Liu, Min-Qian Liu & Yong-Dao Zhou


To cite this article: Si-Yu Yi, Ze Liu, Min-Qian Liu & Yong-Dao Zhou (2023) Global Likelihood Sampler for Multimodal Distributions, Journal of Computational and Graphical Statistics, 32:3, 927-937, DOI: [10.1080/10618600.2023.2165499](https://doi.org/10.1080/10618600.2023.2165499)

To link to this article: <https://doi.org/10.1080/10618600.2023.2165499>

 View supplementary material 

 Published online: 14 Feb 2023.

 Submit your article to this journal 

 Article views: 162

 View related articles 

 View Crossmark data 

 Citing articles: 3 View citing articles 



Global Likelihood Sampler for Multimodal Distributions

Si-Yu Yi, Ze Liu, Min-Qian Liu , and Yong-Dao Zhou

School of Statistics and Data Science, Nankai University, Tianjin, China

ABSTRACT

Drawing samples from a target distribution is essential for statistical computations when the analytical solution is infeasible. Many existing sampling methods may be easy to fall into the local mode or strongly depend on the proposal distribution when the target distribution is complicated. In this article, the Global Likelihood Sampler (GLS) is proposed to tackle these problems and the GL bootstrap is used to assess the Monte Carlo error. GLS takes the advantage of the randomly shifted low-discrepancy point set to sufficiently explore the structure of the target distribution. It is efficient for multimodal and high-dimensional distributions and easy to implement. It is shown that the empirical cumulative distribution function of the samples uniformly converges to the target distribution under some conditions. The convergence for the approximate sampling distribution of the sample mean based on the GL bootstrap is also obtained. Moreover, numerical experiments and a real application are conducted to show the effectiveness, robustness, and speediness of GLS compared with some common methods. It illustrates that GLS can be a competitive alternative to existing sampling methods. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received June 2022
Accepted December 2022

KEYWORDS

Convergency; GL bootstrap;
Low-discrepancy point set;
Robustness

1. Introduction

Sampling from a target distribution is an important issue in statistics. For example, in Bayesian inference, the analytical solution of the posterior expectation is usually unavailable. Instead, the Monte Carlo (MC) and quasi-Monte Carlo (QMC) methods are two popular ways to approximate the integral. In the former case, random sampling is often used, while in the latter, low-discrepancy point sets can be used (Lemieux 2009).

To make the MC strategy work efficiently, many sampling methods have been studied, such as the Sampling/Importance Resampling (SIR) and the improved SIR (Skare, Bølviken, and Holden 2003). However, the success of SIR or improved SIR depends on a good proposal distribution that mimics the target distribution. Llorente et al. (2021) proposed a two-stage importance sampling method with an adaptive proposal, but it may also be difficult to mimic the target well in multimodal and high-dimensional cases. Markov chain Monte Carlo (MCMC, Robert and Casella 2013) is another well-established approach for sampling. When the target is strongly multimodal, some of the MCMC methods may be trapped in a local mode indefinitely, such as the Metropolis-Hastings (MH) algorithm. To overcome this problem, several strategies were proposed, such as the multiple-try Metropolis-type method (Calderhead 2014; Bern-ton et al. 2015; Martino 2018), local optimization (Tjelmeland and Hegstad 2001), and the tempering-based method (Liang and Wong 2001; Miasojedow, Moulines, and Vihola 2013). Liu, Liang, and Wong (2000) combined local optimization with the multiple-try method. The local optimization method may consume much computing resource and a particular optimization

procedure may not work well for all targets. For the tempering-based method, the design of the temperature ladder is a tricky issue and it tends to mix exponentially slowly in dimension if the modes have different covariance structures (Pompe, Holmes, and Łatuszyński 2020).

Compared with MC methods, QMC methods typically use the number-theoretic technique to generate the low-discrepancy points (e.g., Vandevoestyne and Cools 2010; Wang et al. 2015). It has been successfully used in many MC problems to speed up the error convergence rate. The randomized QMC can further improve the rate in some settings, such as randomly-shifted lattice rules. Vandevoestyne and Cools (2010) showed that quasi-random SIR has better convergence than SIR and Ning and Tao (2020) proposed the randomized quasi-random SIR (RQSIR) with higher accuracy than SIR and quasi-random SIR. In addition, QMC methods have also been used in MCMC. Owen and Tribble (2005) proposed a quasi-random version of MH (QMH) and proved the consistency for QMH and Schwedes and Calderhead (2018) considered a general parallel quasi-random MCMC framework. Nevertheless, these methods still have the problem that they are easy to fall into the local mode or fail in high-dimensional cases.

To draw samples from multimodal distributions more efficiently, we develop the Global Likelihood Sampler (GLS), which leverages the low-discrepancy points to obtain more representative samples. Given a target distribution, GLS uses a Good Lattice Point (GLP, Fang et al. 2018) set as the candidate samples, which provides a good representation for the support and does not need to choose a suitable proposal distribution. Thus, it

is problem-independent and more robust than MC and QMC methods which need the proposals. Another novelty is that GLS uses multiple random shifts on the GLP to enhance its space-filling property. Compared with SIR, MCMC, and some (randomized) QMC methods, GLS samples from multiple weighted randomly-shifted GLPs to generate multiple batches of samples, which increases the diversity of the samples and enables a more thorough exploration of the target distribution from a global perspective. Hence, it is essential for GLS to avoid falling into a local mode and work better than other randomized QMC methods. Both GLS and the multiple-try methods can produce multiple batches of samples. The main difference is that the samples in different batches are independent under GLS, while they are dependent under the multiple-try methods. Compared with the local optimization methods, there is no optimization procedure in GLS. There are fewer parameters to be set in GLS compared with the tempering-based approaches and the computational complexity of GLS can also be lower since it is without an iterative process. Moreover, we consider a multi-thread bootstrap method, named GL bootstrap, to estimate the MC error. Further, we thoroughly show the theoretical verifications of GLS and GL bootstrap and illustrate the effectiveness, robustness, and speediness of GLS for multimodal and multidimensional target distributions by comparisons in numerical studies and an application.

The outline of this article is as follows. Section 2 gives a detailed introduction of GLS. Some asymptotic results are derived in Section 3. Section 4 presents the GL bootstrap and derives the convergence for the approximate sampling distribution of the sample mean. Simulation results for various distributions are conducted to compare GLS with some commonly used methods in Section 5. A real application of Bayesian object detection is shown in Section 6 to further illustrate the validity of GLS. Section 7 gives some conclusions and discussions. All the proofs and additional results are given in the Appendix, supplementary materials.

2. Global Likelihood Sampler

In this section, we briefly introduce the GLP and the random shift technique in randomized QMC. Then, we show the GLS algorithm and discuss some details of GLS.

The discrepancy is a type of metric in QMC for measuring how close the QMC points to the true uniform distribution, such as the star discrepancy and wrap-around L_2 -discrepancy (WD), see Fang et al. (2018). An M -run point set is of low discrepancy if some discrepancy between the Empirical Cumulative Distribution Function (ECDF) of the points and the uniform distribution converges to 0 faster than the typical rate for independent random points, that is $O(M^{-1/2})$. The low-discrepancy points can scatter more uniformly than the uniform random points. One of the commonly used low-discrepancy point sets is the GLP. A GLP with M runs and d columns has star discrepancy $O(\kappa(M))$, where $\kappa(M) = 1/M$ if $d = 1$ and $(\log(M))^d/M$, otherwise. The general method for constructing a GLP is shown in the Appendix A.1. The random shift technique (L'Ecuyer 2016) is very common and useful in randomized QMC. It generates a single point uniformly on $[0, 1]^d$ and adds it to each point of a

given design on $[0, 1]^d$ mod 1, coordinate-wise. The Appendix A.1 proves that if we apply a random shift mod 1 to a GLP on $[0, 1]^d$, the low-discrepancy property is preserved. Hence, a (randomly-shifted) GLP in any region also should have a good space-filling property for that region.

By using the space-filling property, GLS randomly shifts the GLP multiple times to explore the distribution structure. In each shift, GLS defines the weights at the randomly-shifted GLP based on the scaled likelihoods. Compared with the case where GLP is replaced by the uniform random proposals, GLS can produce more representative weighted points for the target distribution. Then, GLS performs sampling from the multinomial distribution associated with the weighted shifted-GLP. Based on multiple random shifts, there are multiple multinomial samplings to generate multiple batches of samples, which increase the diversity and randomness of the sampling. Let $\Pi(\theta)$ be a d -dimensional cdf and $\pi(\theta) = C_0 \cdot p(\theta)$ be the pdf with support \mathcal{S} , where C_0 is the normalizing constant. Denote by m the batch size and n the number of batches. The detailed steps for GLS are described in Algorithm 1.

By Algorithm 1, after running GLS, n batches of m -size random samples will be generated. To attain the computational complexity of GLS, we denote by $O(\zeta)$ the complexity of calculating the likelihood p at a design point. For M points in a GLP and n times of shifts on the GLP, the total complexity for calculating the likelihoods is $O(\zeta Mn)$. The complexity of n times of random shifts on M d -dimensional points is $O(Mnd)$. Moreover, with nm samples, the total average-case complexity

Algorithm 1 The procedure of GLS

Input: The target d -dimensional pdf $\pi(\theta)$ with support \mathcal{S} , an M -point GLP Y_M^0 on $[0, 1]^d$, a predefined threshold ϵ_T , the sample size N .

- 1: **Truncation.** Choose $a_t < b_t, t = 1, \dots, d$ and let $\mathcal{D} = [\mathbf{a}, \mathbf{b}]$ with $\mathbf{a} = (a_1, \dots, a_d)^\top, \mathbf{b} = (b_1, \dots, b_d)^\top$, such that $\mathcal{D} \subseteq \mathcal{S}$ is a compact subregion containing most of the probabilities of interest, that is, $\int_{\mathcal{S} \setminus \mathcal{D}} \pi(\theta) d\theta < \epsilon_T$. Transform Y_M^0 onto \mathcal{D} and denote as $Y_M = \{\mathbf{y}_k = (y_{k,1}, \dots, y_{k,d}), k = 1, \dots, M\}$. Set n, m such that $N = nm$ and let $i = 1$.
- 2: **Random shift.** Generate a random shift vector $\mathbf{u}_i = (u_{i,1}, \dots, u_{i,d})^\top \in \mathcal{D}_S = [\mathbf{0}, \mathbf{b} - \mathbf{a}]$, where $u_{i,t} \sim U(0, b_t - a_t), t = 1, \dots, d$. Shift Y_M by $\mathbf{y}_k \oplus \mathbf{u}_i$, where the t th entry of $\mathbf{y}_k \oplus \mathbf{u}_i$ is

$$(\mathbf{y}_k \oplus \mathbf{u}_i)_t = \begin{cases} y_{k,t} + u_{i,t} & \text{if } a_t \leq y_{k,t} + u_{i,t} \leq b_t \\ y_{k,t} + u_{i,t} + a_t - b_t, & \text{otherwise.} \end{cases}$$

$Y_M \oplus \mathbf{u}_i = \{\mathbf{y}_k \oplus \mathbf{u}_i : k = 1, \dots, M\}$ is the randomly-shifted GLP in the i th batch.

- 3: **Likelihood.** Compute $q_{i,k} = p(\mathbf{y}_k \oplus \mathbf{u}_i) / \sum_{j=1}^M p(\mathbf{y}_j \oplus \mathbf{u}_i)$ and let $\tilde{F}_{M,i}$ be the multinomial distribution over $\{\mathbf{y}_k \oplus \mathbf{u}_i : k = 1, \dots, M\}$ with weight $q_{i,k}$ for $k = 1, \dots, M$.
- 4: **Batch sampling.** Sample from $\tilde{F}_{M,i}$ m times to obtain the i th batch of samples $B_i = \{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,m}\}$. If $i < n$, then set $i = i + 1$ and go to Step 2.

Output: Drawn sample set $\mathbb{X} = \{\mathbf{x}_{i,j} \in \mathcal{D} : i = 1, \dots, n, j = 1, \dots, m\}$.

of multinomial samplings is $O(Mnm/2)$. Thus, the overall complexity of GLS is $O((\zeta + d + m/2)Mn)$ in the average case.

For any $\mathbf{x}_{ij} \in \mathbb{X}$, $i = 1, \dots, n$, $j = 1, \dots, m$ and $\boldsymbol{\theta} \in \mathcal{D}$, we have

$$\begin{aligned} P(\mathbf{x}_{ij} \in (-\infty, \boldsymbol{\theta}]) &= E(\tilde{F}_{M,i}(\boldsymbol{\theta}|\mathbf{u}_i)) \\ &= \frac{1}{v(\mathcal{D}_S)} \int_{\mathcal{D}_S} \sum_{k=1}^M q_{i,k} I_{[\mathbf{a}, \boldsymbol{\theta}]}(\mathbf{y}_k \oplus \mathbf{u}_i) d\mathbf{u}_i, \end{aligned}$$

where $v(\mathcal{D}_S)$ is the volume of \mathcal{D}_S and $I_{[\mathbf{a}, \boldsymbol{\theta}]}$ is the indicator function on $[\mathbf{a}, \boldsymbol{\theta}]$. It shows that any element in \mathbb{X} follows the same distribution, denoted by $F_M(\boldsymbol{\theta})$. Moreover, since the shift vectors \mathbf{u}_i , $i = 1, \dots, n$ in different batches are independent and all the points in the GLP are shifted by the same vector in one batch, we can obtain the following proposition.

Proposition 1. Any GLS sample follows the distribution $F_M(\boldsymbol{\theta})$ and the drawn samples in different batches are independent whereas the samples in the same batch are dependent.

By Proposition 1, for fixed N , the setting $n = N$ and $m = 1$ brings iid samples. When $m > 1$, the generated samples in the same batch are dependent. However, for complicated distribution, the complexity of GLS is mainly controlled by the term $O(\zeta Mn)$. Thus, larger m and smaller n can bring lower computational complexity because of fewer evaluations on the likelihoods. In practical implementation, if N is moderate, we can set $n = N$ and $m = 1$ to obtain good performance; otherwise, a larger m is required to tradeoff the performance against the complexity, if iid samples are not the case. Section 5 shows that a larger m would sharply reduce the running time and this scheme does not degrade the performance of GLS much compared with the case $m = 1$. As for M , its setting is associated with the parameters that characterize the target distribution. Let ρ be the parameter to approximately characterize the area proportion of the support of Π compared with \mathcal{D} . Since \mathcal{D} is pre-determined to contain most of the support, the value of ρ is usually smaller than 1. For some thin or narrow distributions, ρ can be small, for example, $\rho < 10\%$, especially when the dimension is high. When ρ is smaller, M should be larger to catch the structure of the target distribution. According to the empirical results in Section 5, if ρ can be roughly pre-determined, $M = 5/\rho$ is a relatively suitable choice to obtain good performance for GLS.

In addition, because of the truncation, the generated sample set \mathbb{X} is associated with the conditional distribution $\Pi(\boldsymbol{\theta}|\mathcal{D})$ essentially, while, it can be derived that for any $\boldsymbol{\theta} \in \mathcal{S}$,

$$|\Pi(\boldsymbol{\theta}) - \Pi(\boldsymbol{\theta}|\mathcal{D})| = \left| \int_{(\mathcal{S} \setminus \mathcal{D}) \cap (-\infty, \boldsymbol{\theta}]} \pi(\mathbf{x}) d\mathbf{x} + \frac{\int_{\mathcal{D}} \pi(\mathbf{x}) d\mathbf{x} - 1}{\int_{\mathcal{D}} \pi(\mathbf{x}) d\mathbf{x}} \int_{\mathcal{D} \cap (-\infty, \boldsymbol{\theta}]} \pi(\mathbf{x}) d\mathbf{x} \right|,$$

in which $(\int_{\mathcal{D}} \pi(\mathbf{x}) d\mathbf{x} - 1) / (\int_{\mathcal{D}} \pi(\mathbf{x}) d\mathbf{x}) < 0$, $\int_{(\mathcal{S} \setminus \mathcal{D}) \cap (-\infty, \boldsymbol{\theta}]} \pi(\mathbf{x}) d\mathbf{x}$ and $\int_{\mathcal{D} \cap (-\infty, \boldsymbol{\theta}]} \pi(\mathbf{x}) d\mathbf{x}$ are nondecreasing with respect to each component of $\boldsymbol{\theta}$. Moreover, the compact region \mathcal{D} is limited to satisfy $\int_{\mathcal{S} \setminus \mathcal{D}} \pi(\mathbf{x}) d\mathbf{x} < \epsilon_T$ in Step 1 of GLS. Thus, we have

$$\sup_{\boldsymbol{\theta} \in \mathcal{S}} |\Pi(\boldsymbol{\theta}) - \Pi(\boldsymbol{\theta}|\mathcal{D})| < \epsilon_T, \quad (1)$$

which implies that the difference can be controlled by the threshold ϵ_T . Based on the fact, we further develop the convergence results for GLS in the next section.

3. Theoretical Results

Let $F(\boldsymbol{\theta}) = \Pi(\boldsymbol{\theta}|\mathcal{D})$ and $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathcal{D})$, where $\pi(\boldsymbol{\theta}|\mathcal{D}) = \pi(\boldsymbol{\theta}) / \int_{\mathcal{D}} \pi(\mathbf{x}) d\mathbf{x}$, if $\boldsymbol{\theta} \in \mathcal{D}$; $\pi(\boldsymbol{\theta}|\mathcal{D}) = 0$, otherwise. Denote $\epsilon_1 = \sup_{\boldsymbol{\theta} \in \mathcal{D}, \mathbf{u} \in \mathcal{D}_S} V_{\text{HK}}(f \cdot I_{[\mathbf{a}, \boldsymbol{\theta}]}) D^*(Y_M \oplus \mathbf{u}) v(\mathcal{D})$, where $V_{\text{HK}}(f \cdot I_{[\mathbf{a}, \boldsymbol{\theta}]})$ is the total variation of $f \cdot I_{[\mathbf{a}, \boldsymbol{\theta}]}$ on \mathcal{D} in the sense of Hardy and Krause (Hlawka 1961), $D^*(Y_M \oplus \mathbf{u})$ is the star discrepancy of $Y_M \oplus \mathbf{u}$, and $v(\mathcal{D})$ is the volume of \mathcal{D} . To establish the asymptotic results for GLS, we need the following assumptions.

Assumption 1. $f(\mathbf{x}) I_{[\mathbf{a}, \boldsymbol{\theta}]}(\mathbf{x})$ is of uniformly bounded variation in the sense of Hardy and Krause for any $\boldsymbol{\theta} \in \mathcal{D}$.

Assumption 2. M is sufficiently large such that $\epsilon_1 \in (0, 1/2)$.

Assumption 3. m is fixed.

Assumption 4. $\kappa(M) \sqrt{m} = O(1)$,

Assumption 5. There exists a constant $\beta' > 0$, such that $\kappa(M) \sqrt{m} / (\log(m))^{\beta'} = O(1)$.

Assumption 1 is required to guarantee the bounded error in Koksma-Hlawka inequality, which serves to achieve uniform asymptotic results. Assumptions 2–5 impose some assumptions on M and m to achieve the convergence in probability or with probability 1. When Assumption 1 holds, $V_{\text{HK}}(f \cdot I_{[\mathbf{a}, \boldsymbol{\theta}]})$ is bounded. $D^*(Y_M \oplus \mathbf{u}) = O(\kappa(M))$ for any $\mathbf{u} \in \mathcal{D}_S$ and $\kappa(M)$ is decreasing with respect to M . The value of $v(\mathcal{D})$ is bounded. Hence, Assumption 2 can be satisfied when M is enlarged to some degree. When N is given, $m(\leq N)$ can be set based on the user's requirement for the performance and the complexity of GLS. Assumptions 4–5 can also be satisfied when M is larger than m to a certain extent.

Based on these assumptions, we derive the convergence of GLS. First, we can obtain how the cdf of the samples $F_M(\boldsymbol{\theta})$ converges to the conditional target cdf $F(\boldsymbol{\theta})$ on \mathcal{D} .

Lemma 1. (1) If Assumption 1 holds, then we have

$$\sup_{\boldsymbol{\theta} \in \mathcal{D}} |F_M(\boldsymbol{\theta}) - F(\boldsymbol{\theta})| = O(\kappa(M));$$

(2) if Assumptions 1, 2 holds, then for any $\mathbf{z} \sim F(\boldsymbol{\theta})$ and closed hypercube $\mathcal{B} \subseteq \mathcal{D}$, where \mathcal{B} is small enough such that $P(\text{Card}((Y_M \oplus \mathbf{u}_i) \cap \mathcal{B}) \leq 1) = 1$, we further have

$$|P(\mathbf{x}_{ij} \in \mathcal{B}) - P(\mathbf{z} \in \mathcal{B})| \leq 2\epsilon_1 P(\mathbf{z} \in \mathcal{B}),$$

where $\text{Card}(\cdot)$ is the cardinality of a set.

Lemma 1(1) shows that $F_M(\boldsymbol{\theta})$ converges to $F(\boldsymbol{\theta})$ on \mathcal{D} uniformly as $M \rightarrow \infty$. The convergence rate $\kappa(M)$ is faster than the rate $O(M^{-1/2})$ for the uniform random proposals. From Lemma 1(2), the relative error between the distributions (or densities) of \mathbf{x}_{ij} and \mathbf{z} can be as small as possible when M is large enough. It also implies that, (i) for any integrable function, the integrals under the two distributions can be arbitrarily close,

and (ii) for any nonnegative integrable function, the relative error between the integrals can be arbitrarily small when M is sufficiently large.

In addition, as previously discussed in Section 2, if $m > 1$, the samples in each batch are dependent. For any $1 \leq k \leq m$, let $F_{(k)}(\theta)$ and $F_{M,(k)}(\theta)$ be the joint cdfs of k independent random vectors that follow $F(\theta)$ and the GLS samples $\mathbf{x}_{1,1}, \dots, \mathbf{x}_{1,k}$, respectively. In the following, we discuss how the correlation and dependence perform among the samples.

Lemma 2. If Assumption 1 holds, then we have

- (1) for any $1 \leq i \leq n$ and $1 \leq k < l \leq m$,
 $\sup_{\theta \in \mathcal{D}} |\text{cov}(I_{[a,\theta]}(\mathbf{x}_{i,k}), I_{[a,\theta]}(\mathbf{x}_{i,l}))| = O(\kappa^2(M))$;
- (2) for any $1 \leq k \leq m$, $\sup_{\theta \in \mathcal{D}} |F_{(k)}(\theta) - F_{M,(k)}(\theta)| = O(\kappa(M))$.

By Lemma 2, both the correlation and dependence among the GLS samples diminish with the increase of M . In practical application, even though enlarging m brings some correlation among the samples in the same batch, it can be weak if M is large.

Let $F_{M,n,m}(\theta)$ be the ECDF of the GLS samples. Based on Lemma 1(2), we show how $F_{M,n,m}(\theta)$ converges to $F_M(\theta)$ uniformly in the following.

Lemma 3. (1) If Assumptions 1, 2, 3 or Assumptions 1, 2, 4 hold, then

$$\sup_{\theta \in \mathcal{D}} |F_{M,n,m}(\theta) - F_M(\theta)| = O_p\left(\frac{1}{\sqrt{nm}}\right);$$

(2) for any $\beta > 0$, if Assumptions 1, 2, 3 or Assumptions 1, 2, 5 with $\beta' < \beta$ hold, then

$$\sup_{\theta \in \mathcal{D}} |F_{M,n,m}(\theta) - F_M(\theta)| = o\left(\frac{1}{\gamma_{n,m}}\right),$$

where $\gamma_{n,m} = \sqrt{nm}/(\log(nm))^{1/2+\beta}$.

From Lemma 3, we obtain the uniform convergence rates of $F_{M,n,m}(\theta)$ to $F_M(\theta)$, in probability and with probability 1, respectively. The latter is a stronger result, while the rate is lower. Based on Lemma 1(1) and Lemma 3, we can further obtain the uniform convergence between $F_{M,n,m}(\theta)$ and the conditional target distribution $F(\theta)$ on \mathcal{D} as follows.

Theorem 1. (1) If Assumptions 1, 2, 3 or Assumptions 1, 2, 4 hold, then

$$\sup_{\theta \in \mathcal{D}} |F_{M,n,m}(\theta) - F(\theta)| = O_p\left(\max\left\{\kappa(M), \frac{1}{\sqrt{nm}}\right\}\right);$$

(2) for some $\beta > 0$, if Assumptions 1, 2, 3 or Assumptions 1, 2, 5 with $\beta' < \beta$ hold, then

$$\sup_{\theta \in \mathcal{D}} |F_{M,n,m}(\theta) - F(\theta)| = O\left(\max\left\{\kappa(M), \frac{1}{\gamma_{n,m}}\right\}\right).$$

Next, we discuss how to improve the convergence rate of GLS. Vandevostyne and Cools (2010) proved that under some conditions, using low-discrepancy points $D_m = \{(2k-1)/2m : k = 1, \dots, m\}$ to do multinomial resampling in quasi-random SIR can accelerate the convergence compared with SIR. In GLS,

let $\varrho^{(i)} : [0, 1] \rightarrow Y_M \oplus \mathbf{u}_i$ be the mapping to represent the multinomial sampling from the shifted GLP with weights $q_{i,j}, j = 1, \dots, M$, in the i th batch, that is, $B_i = \{\varrho^{(i)}(u_m^{(i,j)}) : j = 1, \dots, m\}$, where $u_m^{(i,1)}, \dots, u_m^{(i,m)} \stackrel{\text{iid}}{\sim} U(0, 1)$, $1 \leq i \leq n$, $\varrho_\theta^{(i)} := I_{[a,\theta]} \circ \varrho^{(i)}$, $i = 1, \dots, n$, and \circ represents the composition of two functions. To obtain the improved result, we need another assumption.

Assumption 6. For any $1 \leq i \leq n$, $\varrho_\theta^{(i)}(u)$ is of uniformly bounded variation in the sense of Hardy and Krause for any $\theta \in \mathcal{D}$, run size of GLP M , and shift vector \mathbf{u}_i .

Inspiringly, we can demonstrate that if the randomized QMC points are used in the multinomial samplings of GLS, the convergence rate can be higher.

Theorem 2. If Assumptions 1, 2, 6 hold, and for $i = 1, \dots, n$, the randomized QMC point set $D_m \oplus u_m^{(i)}, u_m^{(i)} \sim U(0, 1)$, are used in Step 4 of GLS for multinomial sampling, that is, $B_i = \{\varrho^{(i)}(d_m^{(i)}) : d_m^{(i)} \in D_m \oplus u_m^{(i)}\}$, then

$$\sup_{\theta \in \mathcal{D}} |F_{M,n,m}(\theta) - F(\theta)| = O_p\left(\max\left\{\kappa(M), \frac{1}{m\sqrt{n}}\right\}\right).$$

By Theorem 2, if the conditions can be satisfied, the convergence rate can be improved from $1/\sqrt{nm}$ to $1/(m\sqrt{n})$ for the convergence of GLS in probability.

Finally, we convert the convergence result of GLS on \mathcal{D} to that on the support \mathcal{S} . From the previous discussion and (1), it is known that for any $\epsilon_T > 0$, we can truncate a tail and obtain a hypercube \mathcal{D} such that $\sup_{\theta \in \mathcal{S}} |\Pi(\theta) - F(\theta)| < \epsilon_T$. Thus, based on Theorem 1, we can also show that $F_{M,n,m}(\theta)$ uniformly converges to $\Pi(\theta)$ on \mathcal{S} .

Theorem 3. For any $\epsilon_T > 0$, we can cut out a closed and bounded hypercube \mathcal{D} and perform the GLS algorithm on \mathcal{D} . If Assumptions 1, 2, 3 or Assumptions 1, 2, 5 with $\beta' < \beta$ holds for some $\beta > 0$, then

$$\sup_{\theta \in \mathcal{S}} |F_{M,n,m}(\theta) - \Pi(\theta)| < \epsilon_T,$$

when M and nm are sufficiently large.

Theorem 3 shows that the ECDF of the GLS samples can approximate the target cdf $\Pi(\theta)$ arbitrarily, which verifies the validity of the GLS algorithm.

4. GL Bootstrap

Generally, the bootstrap can approximate the sampling distribution of a “statistic” based on resampling to construct the hypothesis testing. We introduce the multi-thread GL bootstrap in Algorithm 2, which is mainly used to estimate the MC error for the n -size GLS samples. Compared with bootstrapping from the single-thread sampler with $m = 1$ in GLS, GL bootstrap increases m in GLS and in each T_b , $b = 1, \dots, B$, it only draws one sample from each batch of the GLS samples, which decreases the possibility of repeated samples and increases the diversity of the bootstrap samples. Hence, GL bootstrap allows a

Algorithm 2 The procedure of GL bootstrap

- 1: Let $\xi = \xi(\Pi)$ be the interested vector parameter of $\Pi(\theta)$ and $\hat{\xi} = \xi(F_{M,n,1})$.
- 2: Take one \mathbf{x}_{i,w_i} ($1 \leq w_i \leq m$) randomly from each batch $B_i = \{\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,m}\}$, $i = 1, \dots, n$, and pool them into a pseudo-sample set $T_b = \{\mathbf{x}_{i,w_i} : i = 1, \dots, n\}$. Let $\hat{G}(T_b)$ be the ECDF of T_b . Define $\hat{\xi}_b = \xi(\hat{G}(T_b))$ as the estimate of ξ based on T_b .
- 3: Repeat Step 2 B times to generate B sets T_b and B estimates $\hat{\xi}_b$, $b = 1, \dots, B$, whose ECDF is an approximation for the sampling distribution of $\hat{\xi}$.
- 4: The estimates of ξ and the MC error are, respectively,

$$\bar{\xi} = \frac{1}{B} \sum_{b=1}^B \hat{\xi}_b \quad \text{and} \quad \text{MCE} = \frac{1}{B} \sum_{b=1}^B (\hat{\xi}_b - \bar{\xi})(\hat{\xi}_b - \bar{\xi})^\top.$$

better chance to efficiently reduce the bias for the estimation of some statistic than bootstrapping from the single-thread sampler multiple times. Wang et al. (2015) also illustrated this point of view empirically. With the multi-thread strategy, the sampling distribution of a statistic can also be approximated. Given the GLS sample set $\mathbb{X} = \{\mathbf{x}_{i,j} \in \mathcal{D} : i = 1, \dots, n, j = 1, \dots, m\}$, we explore the sampling distribution of the sample mean $\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{i,1}$ under GL bootstrap. Let $\{\mathbf{x}_i^* : i = 1, \dots, n\}$ be a GL bootstrap pseudo-sample and $\bar{\mathbf{X}}^* = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^*$ be a GL bootstrap estimate of the sample mean, where the pseudo-sample is generated following Step 2 of Algorithm 2. Denote by μ the mean and by Σ the covariance of the target cdf $\Pi(\theta)$. Then, we obtain the asymptotic result for the approximate sampling distribution of the sample mean in the following theorem.

Theorem 4. Let $G_n(\tau) = P(\sqrt{n}(\bar{\mathbf{X}} - \mu) \leq \tau)$ and $G_n^*(\tau) = P(\sqrt{n}(\bar{\mathbf{X}}^* - \bar{\mathbf{X}}) \leq \tau | \mathbb{X})$. If the conditions in Theorem 3 hold, as $M \rightarrow \infty$, $m \rightarrow \infty$ and $n \rightarrow \infty$, we have

$$\sup_{\tau \in \mathbb{R}^d} |G_n(\tau) - G_n^*(\tau)| \rightarrow 0.$$

Conditional on \mathbb{X} , let $\hat{G}_B^*(\tau) = \frac{1}{B} \sum_{b=1}^B I_{(-\infty, \tau]}(\sqrt{n}(\bar{\mathbf{X}}_b^* - \bar{\mathbf{X}}))$. By Glivenko-Cantelli theorem, the simulation error between $\hat{G}_B^*(\tau)$ and $G_n^*(\tau)$ can be made arbitrarily small uniformly for τ by increasing the number of replications B . Hence, Theorem 4 indicates that GL bootstrap works for the sample mean. The validities of GL bootstrap under other statistics can be demonstrated similarly with suitable modifications.

5. Numerical Simulations

In this section, we compare GLS with some popular methods, that is, SIR, MH, EMC, RQSIR, and QMH, through numerical experiments under various settings. In Section 5.1, by sampling from multivariate mixed normal distributions with different numbers of dimensions and modes, we demonstrate the effectiveness and superiority of GLS. In Section 5.2, we compare these methods under some other distributions, whose proposal distributions are not easy to determine. Comparisons are also

conducted under shifted distributions, where the locations or shapes of the modes are shifted deterministically or randomly. Based on these results, the robustness of GLS is illustrated. In the simulations, the sizes of the GLPs in GLS, RQSIR, and the number of the proposals in SIR are set the same, denoted by M .

5.1. The Effectiveness of GLS

In this section, the target distribution is set to be a d -dimensional multimodal distribution, which consists of a mixture of n_m normal distributions,

$$\Pi(\theta) = \frac{1}{n_m} \sum_{i=1}^{n_m} \Phi(\theta | \mu_i, \Sigma_i), \quad (2)$$

where $\Phi(\theta | \mu_0, \Sigma_0)$ is the cdf of the multivariate normal distribution with mean μ_0 and covariance matrix Σ_0 , n_m is the number of modes, $\Sigma_1 = \dots = \Sigma_{n_m} = (\rho/n_m)^{2/d}/64 \cdot \mathbf{I}_d$, and ρ is an approximated area proportion for the support of Π relative to $\mathcal{D} = [0, 1]^d$. The mean vectors μ_1, \dots, μ_{n_m} are decided by a nearly uniform design with size n_m such that most of the area with nonzero density falls into $[0, 1]^d$.

For GLS, we denote it by GLS_1 if $m = 1$, GLS_m otherwise. Two cases are discussed, that is, (i) GLS_1 with $n = N$, $m = 1$; (ii) GLS_m with $n = m = \sqrt{N}$. For the multimodal target distribution, it may be difficult to determine an appropriate proposal distribution. Hence, the proposal distributions for SIR, MH, and EMC are set as the random uniform distribution, which may be a robust choice. Other settings for the considered methods are shown in the Appendix B.1. Given the target cdf Π with pdf π and the ECDF Π_N of the drawn samples $\{\mathbf{x}_i = (x_{i,1}, \dots, x_{i,d}) : i = 1, \dots, N\} \subseteq [0, 1]^d$, instead of $\sup_{\theta \in [0, 1]^d} |\Pi(\theta) - \Pi_N(\theta)|$, we adopt a generalization of WD in QMC to compare the samples under different methods for saving time. The generalized WD (GWD) is similar to WD, except that the uniform distribution is replaced by Π . The smaller the GWD, the closer the ECDF of the samples is to the target. Specifically, GWD has the form

$$\begin{aligned} \text{GWD} &= \left[\int_{[0,1]^{2d}} \prod_{\ell=1}^d k(t_\ell, z_\ell) \pi(\mathbf{t}) \pi(\mathbf{z}) \, d\mathbf{t} \, d\mathbf{z} \right. \\ &\quad - \frac{2}{N} \sum_{i=1}^N \int_{[0,1]^d} \prod_{\ell=1}^d k(t_\ell, x_{i,\ell}) \pi(\mathbf{t}) \, d\mathbf{t} \\ &\quad \left. + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \prod_{\ell=1}^d k(x_{i,\ell}, x_{j,\ell}) \right]^{1/2} \\ &\approx \left[\frac{1}{N_{\text{qmc1}}} \sum_{q_1=1}^{N_{\text{qmc1}}} \prod_{\ell=1}^d k(t_{q_1,\ell}, z_{q_1,\ell}) \pi(\mathbf{t}_{q_1}) \pi(\mathbf{z}_{q_1}) \right. \\ &\quad - \frac{2}{NN_{\text{qmc2}}} \sum_{i=1}^N \sum_{q_2=1}^{N_{\text{qmc2}}} \prod_{\ell=1}^d k(e_{q_2,\ell}, x_{i,\ell}) \pi(\mathbf{e}_{q_2}) \\ &\quad \left. + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \prod_{\ell=1}^d k(x_{i,\ell}, x_{j,\ell}) \right]^{1/2}, \end{aligned}$$

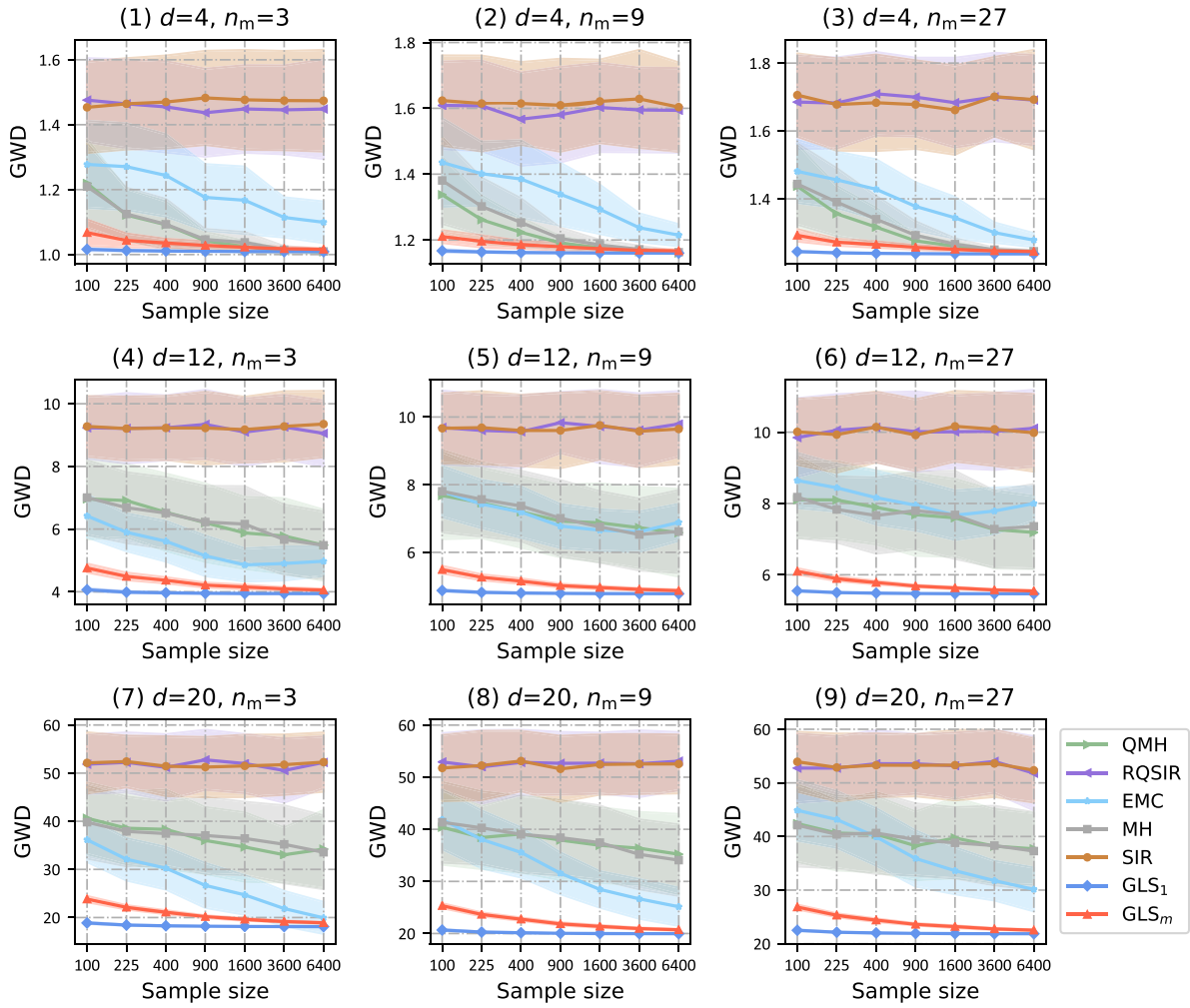


Figure 1. The comparison of GWD with $\rho = 2\%$. Under each method, the line and the corresponding bandwidth of the shaded band denote the average and double standard deviation of GWDs for 100 repetitions, respectively.

where \mathcal{K} is the WD kernel $\mathcal{K}(\mathbf{t}, \mathbf{z}) = \prod_{\ell=1}^d k(t_\ell, z_\ell)$ with $k(t, z) = 3/2 - |t - z| + |t - z|^2$, $\{\mathbf{t}_1, \dots, \mathbf{t}_{\text{qmc1}}\}$, $\{\mathbf{z}_1, \dots, \mathbf{z}_{\text{qmc1}}\}$ and $\{\mathbf{e}_1, \dots, \mathbf{e}_{\text{qmc2}}\}$ are quasi-random numbers on $[0, 1]^d$.

For Π in (2), we consider $d = 4, 16, 64$, $n_m = 3, 9, 27$, and $\rho = 0.5\%, 1\%, 2\%$. For each case, the sample size $N = 100, 225, 400, 900, 1600, 3600, 6400$. In general, when ρ is smaller, M should be larger for better exploration. We set $M = 5/\rho$. Samplings and computations are carried out on a server Intel(R) Xeon(R) Gold 5118 CPU @ 2.30 GHz. Under each setting, we replicate each method (i.e., QMH, RQSIR, EMC, MH, SIR, GLS₁ and GLS_m) 100 times. The average CPU time for running each method once is recorded. Based on the 100 samples, we compute the means and the standard deviations of GWDs, L_∞ -norms of the bias of sample mean (BSM), and L_∞ -norms of the bias of sample covariance matrix (BSC). Here we show the results for GWD and running time with $\rho = 2\%$ for saving space. Other results can be found in the Appendix B.

In Figure 1, we show the comparison of GWDs under different methods. SIR and RQSIR have relatively bad performance with large GWDs whatever d and n_m are. When d is small, for example, $d = 4$, QMH performs better than MH with smaller averages for GWDs and both of them are better than EMC. If d is larger and n_m is small, for example, $d = 12, 20$ and

$n_m = 3$, QMH and MH show similar performance and EMC can perform better than them. However, the advantage of EMC diminishes with the increase of n_m . As for GLS_m and GLS₁, they are always more stable and show better performance than other methods, especially when d is higher and n_m is larger. GLS₁ performs better than GLS_m because of the independence of the samples and more exploration (i.e., larger n) for the distribution structure. The comparisons for cases $\rho = 0.5\%, 1\%$ show similar conclusions in the Appendix B.1. In general, when ρ is smaller, and d and n_m are larger, GLS₁ and GLS_m can maintain relatively good performance, while other methods show worse performance. Moreover, we display the L_∞ -norms of BSM and BSC in the Appendix B.2 and B.3, respectively. The result for BSM gives similar conclusions to those based on GWD, and in terms of BSC, the superiority of GLS is also demonstrated. Hence, these results illustrate that GLS has an advantage in exploring the structure of the multimodal distribution.

In addition, we also compare the average CPU time in Figure 2 and Appendix B.4. When M is small, that is, ρ is large (e.g., $\rho \geq 2\%$), the computational complexity of GLS₁ is lower than that of EMC; when M increases (e.g., $\rho \leq 1\%$ and N is large), the complexity of GLS₁ could be higher than that of EMC. Nevertheless, it is worth noting that the running time of GLS_m is

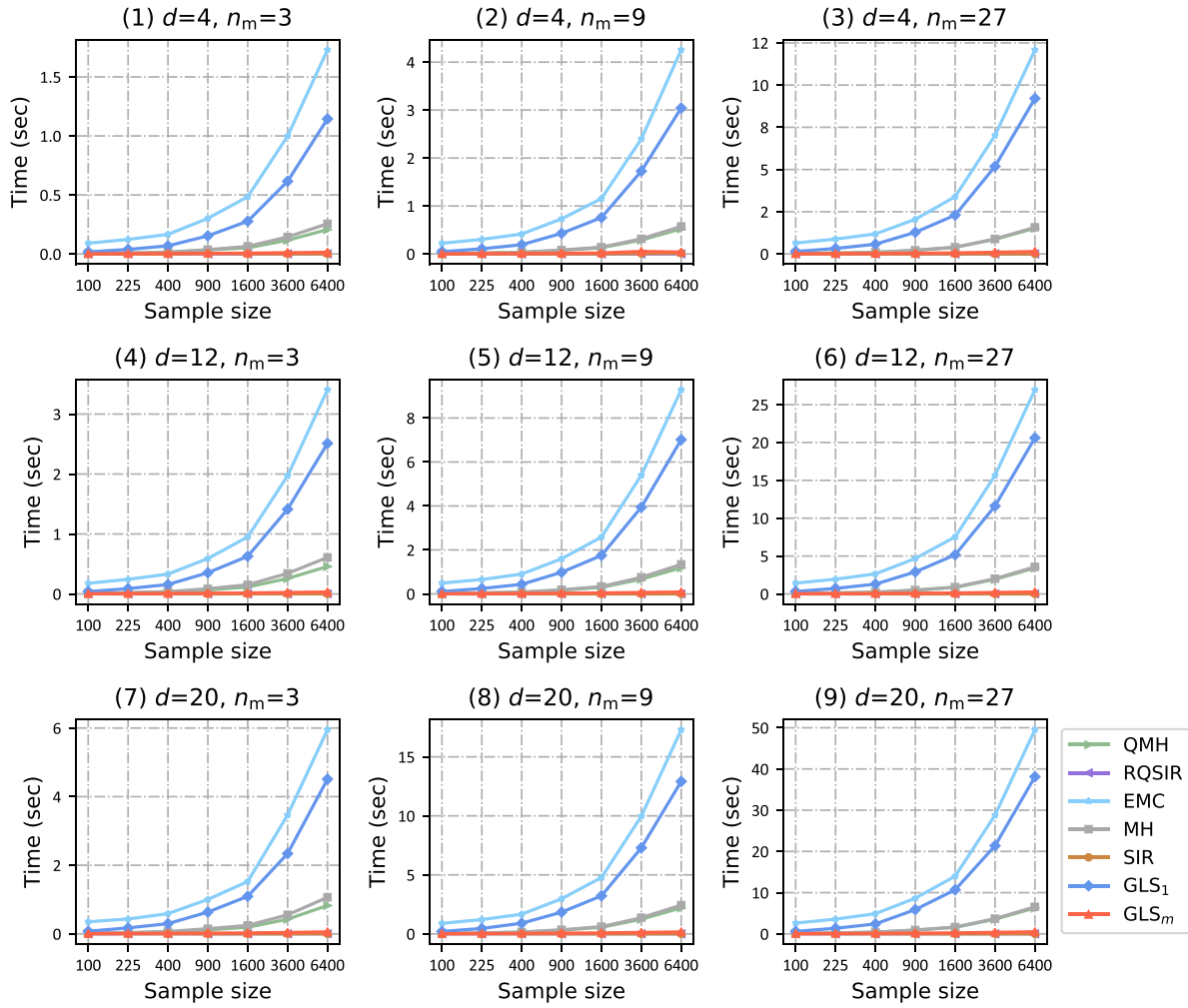


Figure 2. The comparison of the average CPU time with $\rho = 2\%$.

always short, which is close to those of SIR and RQSIR. It implies that the computational complexity can be decreased sharply for GLS by increasing the batch size m , when the total sample size is fixed. Hence, considering both computational complexity and performance, GLS_1 is a good choice with better performance and acceptable running time when the sample size is small, while GLS_m maybe better with less running time and comparable effect when the sample size is large.

Finally, based on the samples from GLS_m and GLS_1 , we study the convergence of GL bootstrap in terms of the standard deviations of the L_∞ -norm of BSM and GWD. Specifically, for ξ being the L_∞ -norm of BSM or GWD, denote by $MCE(\xi)$ and $SD(\xi)$ the estimated standard deviation by GL bootstrap and the standard deviation by repeated samplings, respectively. For each N in each case, we perform GL bootstrap $B = 100$ times based on one $10N$ -size sample from GLS_m with $n = N$ and $m = 10$. We calculate $MCE(\xi)$ based on the 100 N -size GL bootstrap pseudo-samples and compare them with the corresponding $SD(\xi)$ obtained from 100 repeated samplings by GLS_1 with $n = N$ and $m = 1$. Then, we compute $Diff(\xi) = \sqrt{N}(SD(\xi) - MCE(\xi))$, which are shown in Figure 3. For the L_∞ -norm of BSM, in many cases, there is almost no obvious increasing or decreasing tendency; for GWD, there is also no obvious trend with the increase of N . Hence, we can nearly claim that both $Diff$ s of L_∞ -norm of BSM and GWD can be bounded

when $N(= nm)$ is large. That is, both convergence rates of the estimated standard deviations of L_∞ -norm of BSM and GWD by GL bootstrap are approximately $O(1/\sqrt{n})$ with fixed M and m .

5.2. The Robustness of GLS

In this section, we discuss some other distributions to study the robustness of GLS. Following Ning and Tao (2020), we first apply SIR, RQSIR, MH, and GLS to a bivariate posterior distribution and a unimodal Kotz-type distribution to approximate the (posterior) expectation. We also consider two cases for GLS, that is, (i) GLS_1 with $m = 1$ and $n = N$, (ii) GLS_m with $m = 10$ and $n = N/m$. Other settings for the considered methods are in the Appendix B.5. For these methods, similar to Ning and Tao (2020), we set (i) $M = 1000$, $N = 100$, (ii) $M = 3000$, $N = 200$. We replicate each method 1000 times and compare the MSEs of the mean estimations under different methods, that is,

$$MSE = \frac{1}{1000} \sum_{k=1}^{1000} \left[\hat{E}_k(X) - \overline{\hat{E}_k(X)} \right]^2,$$

where $\hat{E}_k(X) = \sum_{i=1}^N \mathbf{x}_i^{(k)} / N$, $\overline{\hat{E}_k(X)} = \sum_{k=1}^{1000} \hat{E}_k(X) / 1000$ and $\{\mathbf{x}_i^{(k)} : i = 1, \dots, N\}$ is the sample set for the k th replication. When the true expectation is known, $\overline{\hat{E}_k(X)}$ is replaced by the

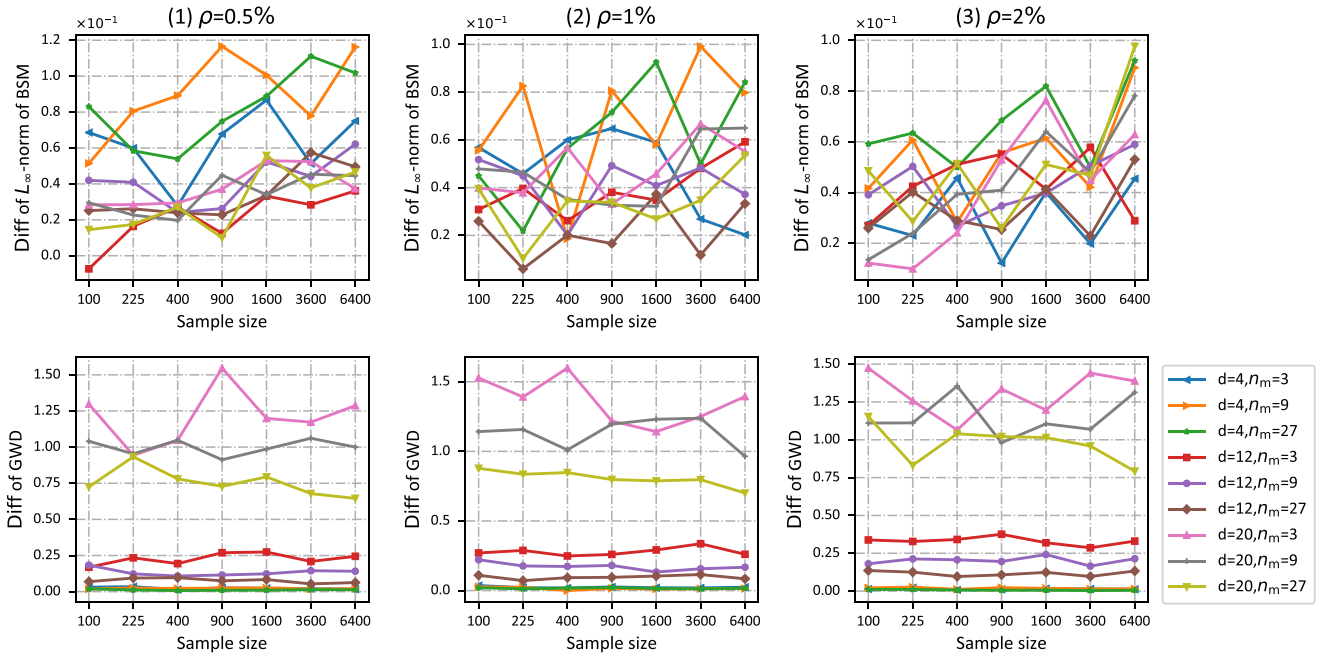


Figure 3. The comparisons of Diff for L_∞ -norm of the bias of sample mean (BSM) and GWD with $\rho = 0.5\%, 1\%, 2\%$.

Table 1. The MSEs of the mean estimations for the Kotz-type distribution with $M = 1000, N = 100$.

Method	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	sum
SIR	0.5974	0.7902	1.9724	5.7121	21.6671	113.7552	144.4946
RQSIR	0.5833	0.6357	1.7315	5.1814	15.3513	70.2728	93.7560
MH	0.3336	0.4719	1.2200	3.6727	13.4874	65.0283	84.2140
GLS ₁	0.1851	0.2327	0.2650	0.2635	0.2438	0.2528	1.4430
GLS _m	0.1879	0.2360	0.2655	0.2633	0.2447	0.2526	1.4500

NOTE: sum is the summation of the MSE for each component.

true value. The sampling details for the bivariate posterior distribution are given in the Appendix B.5. In this case, SIR, RQSIR, and MH have suitable proposals. GLS shows comparable performance to approximate the posterior expectation compared to the methods with good proposal distributions.

For the Kotz-type distribution, the corresponding pdf has the form

$$\pi(\theta) = C_a |\Sigma_a|^{1/2} \left[(\theta - \mu_a)^\top \Sigma_a^{-1} (\theta - \mu_a) \right]^{J-1} \exp \left\{ -r \left[(\theta - \mu_a)^\top \Sigma_a^{-1} (\theta - \mu_a) \right]^s \right\},$$

where $r > 0, s > 0, J > (2 - d)/2$, C_a is the normalizing constant, μ_a is the mean vector, and Σ_a is the dispersion matrix. We set $r = 1, s = 2, J = 2$ and consider a six-dimensional Kotz-type distribution with μ_a being the 6×1 vector of zeros and Σ_a being the six-dimensional Pascal matrix, which is shown in the Appendix B.5.

Following Ning and Tao (2020), the proposal distributions in SIR, RQSIR, and MH are set as the multivariate normal distribution $\Phi(\theta | \mu_a, \Sigma_p)$ with $\Sigma_p = \text{diag}(1, 2, 6, 20, 70, 252)$, the diagonal elements of Σ_a . In this setting, the proposal distribution goes far away from the target, since the target distribution is quite skewed. When the coordinate index is larger, the degree of the skewness is greater. Table 1 shows the MSEs for the mean estimation under different methods with $M = 1000, N =$

100. Except for GLS₁ and GLS_m, other methods bring large MSEs, especially when the coordinate index is large. Another case presents a similar phenomenon in the Appendix B.5. It implies that for SIR, RQSIR, and MH, when the chosen proposal distribution is not suitable or it is difficult to find a suitable proposal distribution, the performance would be pretty bad, while GLS can well avoid such situations and shows certain robustness to the target distribution.

In addition, we also perform some shifts on the target distributions in (2) with $d = 12, n_m = 9$, and $\rho = 1\%, 3\%$. The details are shown in the Appendix B.6. From the result, it can be concluded that GLS is robust to the shape shifting of the modes and has good performance. When the distribution has relatively dispersed modes, GLS can perform well and stably; even though the modes are centralized, GLS_m can also bring good performance with limited computing resources.

6. Application—Bayesian Object Detection

In this section, we consider a real application, detecting the discrete objects hidden in some background noise, to further illustrate the validity of GLS. A Bayesian approach for this problem in an astrophysical context was first presented by Hobson and McLachlan (2003). Similar task is also discussed in other literature on the multimodal distributions, such as Pompe, Holmes, and Łatuszyński (2020) that focused on the inference about the locations of sensors.

Following Hobson and McLachlan (2003), we wish to detect and characterize some Gaussian-shaped discrete objects in a single image, each of which is described by

$$\tau(\mathbf{x}; \mathbf{c}) = A \exp \left\{ -1/2 \cdot [\mathbf{x} - \mathbf{X}, \mathbf{y} - \mathbf{Y}] \mathbf{R}^{-1} [\mathbf{x} - \mathbf{X}, \mathbf{y} - \mathbf{Y}]^\top \right\}, \quad (3)$$

where $\mathbf{c} = \{X, Y, A, \mathbf{R}\}$, $\mathbf{x} = (x, y)$, (X, Y) is the location of the object, A is its amplitude, and $\mathbf{R} \in \mathbb{R}^{2 \times 2}$ is some measure

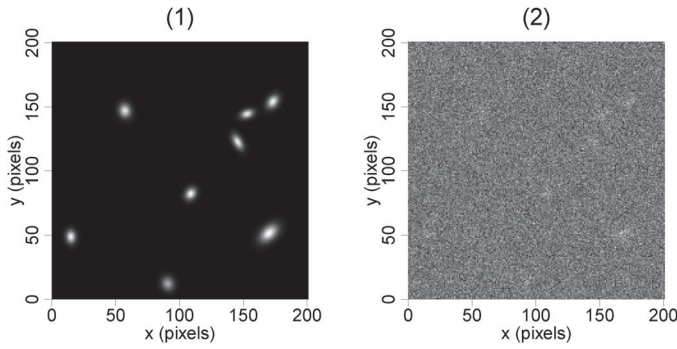


Figure 4. (1): The 201×201 -pixel test image; (2): the corresponding data map with independent Gaussian pixel noise added to (1).

of its spatial extent. Denote the vector \mathbf{D} as the pixel values in the image. If there exist N_{obj} objects defined by (3) and the contribution of each object to the data is additive, we have $\mathbf{D} = \mathbf{n} + \mathbf{s}(\tilde{\mathbf{c}})$, where $\tilde{\mathbf{c}} = \{c_1, \dots, c_{N_{\text{obj}}}\}$, $\mathbf{s}(\tilde{\mathbf{c}}) = \sum_{i=1}^{N_{\text{obj}}} \mathbf{s}(c_i)$ denotes the contribution to the data from the i th discrete objects, and \mathbf{n} denotes the noise. For simplicity, N_{obj} is fixed and known. We wish to use the data \mathbf{D} to detect the objects. Figure 4(1) shows the 201×201 -pixel test image containing $N_{\text{obj}} = 8$ objects. The details for data generation are shown in the Appendix B.7. In Figure 4(2), we plot the data map by adding independent Gaussian pixel noise with a variance of 1.5. With this level of noise, only a few objects are visible.

A desirable approach is to detect all the objects in the image simultaneously by sampling from the posterior distribution of the full set of parameters $\tilde{\mathbf{c}}$. Specifically, since the noise \mathbf{n} follows $\Phi(\mathbf{n}|\boldsymbol{\mu}_b, \Sigma_b)$ with $\boldsymbol{\mu}_b = \mathbf{0}_{N_{\text{pix}}}$ and $\Sigma_b = 1.5 \cdot \mathbf{I}_{N_{\text{pix}}}$, the likelihood function is

$$L(\tilde{\mathbf{c}}) = \frac{\exp\{-1/2 \cdot [\mathbf{D} - \mathbf{s}(\tilde{\mathbf{c}})]^\top \Sigma_b^{-1} [\mathbf{D} - \mathbf{s}(\tilde{\mathbf{c}})]\}}{(2\pi)^{N_{\text{pix}}/2} |\Sigma_b|^{1/2}}, \quad (4)$$

where $N_{\text{pix}} = 201^2$ is the length of \mathbf{D} . Let $\pi_p(\tilde{\mathbf{c}})$ denote the joint prior distribution of $\tilde{\mathbf{c}}$. Then the (unnormalized) posterior distribution is $P(\tilde{\mathbf{c}}|\mathbf{D}) = L(\tilde{\mathbf{c}})\pi_p(\tilde{\mathbf{c}})$. If we focus on all the location parameters, the parameter space is $\{X_1, Y_1, \dots, X_8, Y_8\}$, which corresponds to a 16-dimensional sampling problem. In addition, by Hobson and McLachlan (2003), an alternative approach is simply to set $N_{\text{obj}} = 1$. We directly sample from the corresponding posterior and try to detect the objects. Even if all the parameters are taken into account, the parameter space is still reduced, that is, $\{X, Y, A, \mathbf{R}\}$, which is a six-dimensional problem. If we only focus on the location parameters and fix A, \mathbf{R} , the space can be further reduced to $\{X, Y\}$, which is only two-dimensional. Fixing $N_{\text{obj}} = 1$ does not restrict us to detecting just a single object. Indeed, the posterior distribution would possess numerous local maxima, each corresponding to the location of one of the objects. Hobson and McLachlan (2003) showed that this method is reliable when the objects of interest are separated spatially.

Here we consider the 2-, 6-, and 16-dimensional cases discussed above and let the sample size $N = 2500$. We use GLS₁ ($n = N, m = 1$), GLS_m ($n = m = \sqrt{N}$), SIR, MH, and EMC to sample from the posterior distribution and compare the performance of these methods for detecting the

objects. The prior distributions and the settings for the considered sampling methods are shown in the Appendix B.7. The sampling results are shown in Figure 5. To judge the detection performance, we also plot the contours for the objects in Figure 5, which form 8 elliptical areas. Clearly, for the three cases, SIR and MH cannot capture most of the objects and they are easy to fall into a local mode. For EMC, the performance for the 16-dimensional case is the best with samples falling into seven elliptical areas, whereas for the other two cases, more than half of the objects are not detected. SIR, MH, and EMC all produce many repeated samples. Under GLS₁ and GLS_m, it can be seen from Figure 5 that most of the objects are captured for the three cases. Moreover, the drawn samples have more diversity compared with other methods. From Figure 5(11)–(15), compared with other methods in the 16-dimensional case, there are more samples out of the elliptical areas under GLS₁ and GLS_m after projecting the eight pairs of the location samples into the (X, Y) -subspace. In fact, the number of samples in each elliptical area is more than that of any other area of the same size. Based on the true location of each object, we also calculate the bias of the sample mean and standard deviation of the samples in each elliptical area of Figure 5 and the numbers of samples in each area under GLS₁ and GLS_m. The details of these quantitative results are shown in the Appendix B.7. Since SIR, MH, and EMC draw more repeated samples that lead to smaller standard deviations, we mainly focus on the comparison of the biases. Compared with other methods, many of the biases under GLS₁ and GLS_m are smaller. Thus, GLS can detect more objects with higher accuracy. Moreover, in most cases, the variance of the samples in each elliptical area under GLS₁ is larger than that under GLS_m, while the bias of the sample mean under GLS₁ is smaller in many cases. It implies that GLS₁ generates more diverse samples and brings more accurate detection. In terms of the number of samples in each elliptical area, GLS₁ is more inclined to produce more samples than GLS_m at the modes with relatively small densities. It illustrates that GLS₁ takes a more thorough exploration of the target distribution. In summary, for object detection, GLS₁ and GLS_m perform more effectively than other considered methods for multimodal distribution. GLS₁ takes more thorough exploration and brings more accurate detection compared with GLS_m, while GLS_m requires less computational cost.

7. Conclusions and Discussions

For GLS, we theoretically prove that the ECDF of the samples converges to the target distribution. When the randomized QMC points are used in the multinomial samplings of GLS, the convergence rate can be improved. We also demonstrate the convergence of the approximate sampling distribution for the sample mean under GL bootstrap.

A variety of numerical experiments are conducted under different distributions to compare GLS with other sampling methods. The results show that GLS has a good performance and is stable for multimodal and multidimensional distributions. In terms of computational complexity, we can increase m and decrease n to reduce the running time for a fixed N . By the

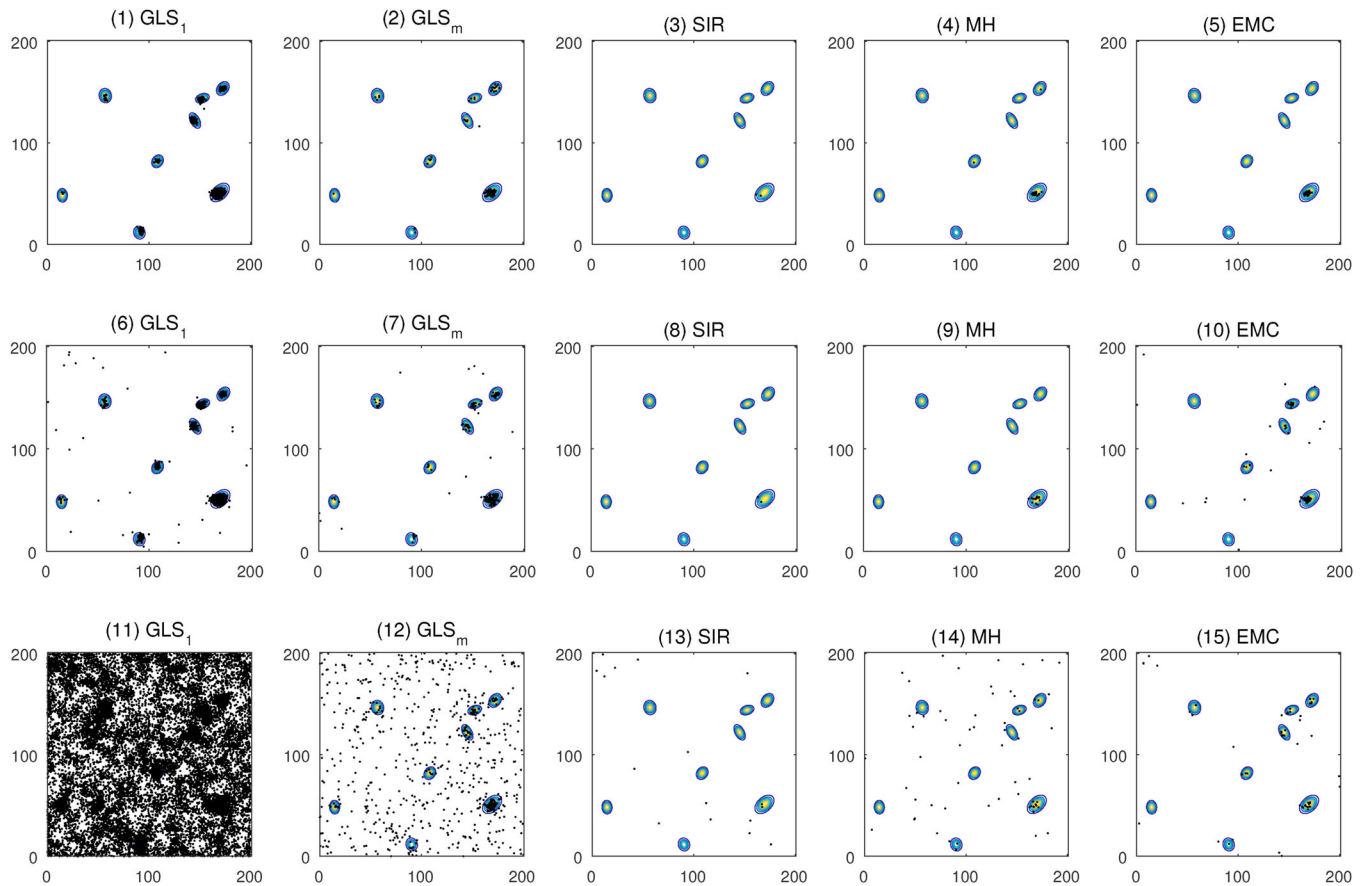


Figure 5. The sample plots from the posterior for the considered three cases. The black dots are the samples and the ellipses are the contours corresponding to Figure 4(1). The upper, middle, and lower panels depict the (projected) samples in the (X, Y) -subspace for the 2-, 6-, and 16-dimensional cases, respectively.

comparisons of sampling performance and running time, if $M < 500$ and $N < 400$, GLS_1 is recommended with good performance and acceptable computational cost; otherwise, GLS_m is a better choice with less cost and relatively good result. The set of m in GLS_m shall be a compromise between performance and complexity. By the simulations, $m = \sqrt{N}$ is a good choice. Moreover, when the proposal distribution is well set for the methods that need proposals, GLS gives comparable performance; when the proposal distribution cannot be given suitably, GLS outperforms those methods. The distribution shifts also do not affect the superior performance of GLS. It illustrates that the GLS is robust against the distribution structure. The application in Bayesian object detection further verifies the validity of GLS with multiple detected objects. In addition, GLS also performs well under other unimodal distributions, which are not presented for saving space. Hence, without the intention to replace MCMC or other advanced sampling methods, GLS can be a promising and competitive alternative with effectiveness, robustness, and speediness.

Additionally, GLS uses random shifts on GLP to enhance the ability for exploring the target distribution structure. To improve the performance of GLS, the deterministic shifts for the GLP can be studied for a better space-filling property. Moreover, the current version of GLS is probably not suitable when the target is extremely heavy-tail or the concern is extreme events. It is of particular importance to improve GLS for accommodating heavy-tail distributions. It is beyond the scope of the article but worthy of further investigation.

Supplementary Materials

Code: The supplemental file includes all the programs to reproduce the results in the article. (GLS_CODE_ALL.zip)

Appendix: The supplemental file includes the Appendix which gives all the proofs and additional results. (GLS-appendix.pdf)

Acknowledgments

The authors thank the editor, the associate editor, and two reviewers for their helpful comments. The authors would like to thank Yuchung Wang for his help.

Funding

This work was supported by the National Natural Science Foundation of China (11871288 and 12131001), the National Ten Thousand Talents Program, the Fundamental Research Funds for the Central Universities, LPMC, and KLMDASR.

ORCID

Min-Qian Liu  <http://orcid.org/0000-0003-1954-6241>

References

- Bernton, E., Yang, S., Chen, Y., Shephard, N., and Liu, J. S. (2015), "Locally Weighted Markov Chain Monte Carlo," arXiv preprint arXiv:1506.08852. [927]

- Calderhead, B. (2014), “A General Construction for Parallelizing Metropolis-Hastings Algorithms,” *Proceedings of the National Academy of Sciences*, 111, 17408–17413. [927]
- Fang, K. T., Liu, M. Q., Qin, H., and Zhou, Y. D. (2018), *Theory and Application of Uniform Experimental Designs*, Singapore: Springer. [927,928]
- Hlawka, E. (1961), “Funktionen von beschränkter Variation in der Theorie der Gleichverteilung,” *Annali Di Matematica Pura Ed Applicata*, 54, 325–333. [929]
- Hobson, M. P., and McLachlan, C. (2003), “A Bayesian Approach to Discrete Object Detection in Astronomical Data Sets,” *Monthly Notices of the Royal Astronomical Society*, 338, 765–784. [934,935]
- L’Ecuyer, P. (2016), “Randomized Quasi-Monte Carlo: An Introduction for Practitioners,” in *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, Cham: Springer, pp. 29–52. [928]
- Lemieux, C. (2009), *Monte Carlo and Quasi-Monte Carlo Sampling*, New York: Springer. [927]
- Liang, F., and Wong, W. H. (2001), “Real-Parameter Evolutionary Monte Carlo with applications to Bayesian Mixture Models,” *Journal of the American Statistical Association*, 96, 653–666. [927]
- Liu, J. S., Liang, F., and Wong, W. H. (2000), “The Multiple-try Method and Local Optimization in Metropolis Sampling,” *Journal of the American Statistical Association*, 95, 121–134. [927]
- Llorente, F., Martino, L., Delgado-Gómez, D., and Camps-Valls, G. (2021), “Deep Importance Sampling based on Regression for Model Inversion and Emulation,” *Digital Signal Processing*, 116, 103104. [927]
- Martino, L. (2018), “A Review of Multiple Try MCMC Algorithms for Signal Processing,” *Digital Signal Processing*, 75, 134–152. [927]
- Miasojedow, B., Moulines, E., and Vihola, M. (2013), “An Adaptive Parallel Tempering Algorithm,” *Journal of Computational and Graphical Statistics*, 22, 649–664. [927]
- Ning, J. H., and Tao, H. Q. (2020), “Randomized Quasi-Random Sampling/Importance Resampling,” *Communications in Statistics-Simulation and Computation*, 49, 3367–3379. [927,933,934]
- Owen, A. B., and Tribble, S. D. (2005), “A quasi-Monte Carlo Metropolis Algorithm,” *Proceedings of the National Academy of Sciences*, 102, 8844–8849. [927]
- Pompe, E., Holmes, C., and Łatuszyński, K. (2020), “A Framework for Adaptive MCMC Targeting Multimodal Distributions,” *The Annals of Statistics*, 48, 2930–2952. [927,934]
- Robert, C. P., and Casella, G. (2013), *Monte Carlo Statistical Methods*, New York: Springer. [927]
- Schwedes, T., and Calderhead, B. (2018), “Quasi Markov Chain Monte Carlo Methods,” arXiv preprint arXiv:1807.00070. [927]
- Skare, Ø., Bølviken, E., and Holden, L. (2003), “Improved Sampling-Importance Resampling and Reduced Bias Importance Sampling,” *Scandinavian Journal of Statistics*, 30, 719–737. [927]
- Tjelmeland, H., and Hegstad, B. K. (2001), “Mode Jumping Proposals in MCMC,” *Scandinavian Journal of Statistics*, 28, 205–223. [927]
- Vandevostyne, B., and Cools, R. (2010), “On the Convergence of Quasi-Random Sampling/Importance Resampling,” *Mathematics and Computers in Simulation*, 81, 490–505. [927,930]
- Wang, Y. C., Ning, J. H., Zhou, Y. D., and Fang, K. T. (2015), “A New Sampler: Randomized Likelihood Sampling,” in *Souvenir Booklet of the 24th International Workshop on Matrices and Statistics*, pp. 255–261. [927,931]