Robust Inference Based on the Complementary Hamiltonian Monte Carlo

X. Xu, C. Liu, and H. Yang, *IEEE Golden Core member*

Abstract—The study aims to explore some reliability probability models using a Hamiltonian Monte Carlo sampler. There are certain concerns with the classic Hamiltonian Monte Carlo samplers. A notable aspect of the method –comess from its acceptance probability whose value is constant, "1", in theory. One practical problem is the possibility of divergence. A further issue emerges from the simulation trajectory, which preferentially traverses moves on the last principal component. Also, the method is extremely sensitive to run-time parameters. As an improvement, Riemann Manifold Hamiltonian Monte Carlo can travel along the first principal component. However, its overall accuracy on other components remains unclear. The No-U-Turn sampler can adjust the distance traveled, but it may also suffer from excessive simulation steps. To address these concerns, this study proposes: 1) to implement the conservation of energy by virtual collision among particles, 2) to adopt constant simulation steps and adjust the acceptance probability and step sizestep-size, and 3) to alternate two kind of trajectories to reconcile all principal components. Experiments show that the proposed algorithm is able to estimate the reliability and remaining useful life of the probability models. To bridge theory and practice, algorithm fragment is demonstrated with Mathematica language.

***Index Terms*—Bayes methods, Monte Carlo methods, Reliability engineering**

I. Introduction

THE reliability function R(t), known more usually as the survivor function S(t) in biomedical applications, expresses the probability that the life of a device or subject will exceed a given value, in which F(t) is the probability function that life time T will take a value less than or equal to t (1).

$$R(t) = S(t) = 1 - F(t) = \text{Prob}[T > t] \quad (1)$$

It is also possible to adopt Binomial distribution for timeless events. For software, the hazard rate representing the conditional probability intensity that an item of age t will fail in the next moment, could be constant. The Weibull distribution, extensively used in defining hardware reliability, can be studied using Hamiltonian Monte Carlo [1].

Two kinds of inferences can be drawn based on the sampled parameters of the reliability (survival) functions, that is, the probability of an item operating ordinarily for a certain amount of time as well as the useful life left on an item at a particular time of operation. The former is termed as reliability (probability) and the latter is measured with time span. In this study, for completeness, both estimations will be demonstrated.

There are plenty of uncommon probability models in reliability engineering that can adapt to Markov Chain Monte Carlo (MCMC). For example, it is known that, in the literature on software reliability, the exponential fault time class has the most articles written on it [2]. As the exponential distribution is a special case of the Weibull, therefore the software reliability can borrow many techniques from the reliability engineering. Using MCMC, it is possible to study some complex models that are out of the range of the traditional classical statistical methods, such as the change point model [3].

Considering causal relation, the reliability of a system can also be defined using Bayesian networks [4][5]. Traditionally, the research of Bayesian networks is mainly based on the methods originated from graph theory [6][7][8]. However, the Bayesian networks are special kind of Bayesian models and thus can be sampled with MCMC [9]. The benefit of MCMC will be clear for small data or uncertain environment.

Compared with others, survival data are often characterized by censored observations -- event times which are only known to be within a certain range [10]. Sampling reliability models with MCMC is possible, yet progressively censored data has been problematic for sampling by introducing non-normality.

Remaining Useful Life (RUL), the length of time a machine is likely to operate before it requires repair or replacement, is a top priority in predictive maintenance programs. One kind of the estimation that based on indirectly observed state processes is the Cox's covariate based hazard models [11][12]. In recent years, there has been extensive research on Cox's model using partial likelihood to model complex covariates [13][14]. The method of Cox Proportional Hazards model doesn't model the underlying hazard, which is both the model's great strength and one of its major drawbacks. It is called partial likelihood for

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

X. Xu is with the Nanjing University of Science and Technology, China (e-mail: xiaopeng.xu@aufe.edu.cn).

C. Liu is with the Nanjing University of Science and Technology, China (e-mail: chuancailiu@njust.edu.cn).

H. Yang is with the School of Informatics, University of Leicester, UK (e-mail: Hongji.Yang@Leicester.ac.uk).

带格式的: 左

带格式的: 字体: (默认) Times New Roman, (中文) Times New Roman

带格式的: 字体: (默认) Times New Roman, (中文) Times New Roman

ignoring censoring, thus is imprecise in theory.

Currently, there are several modelling and sampling tools, such as WinBUGS, JAGS, PyMC, PyMC3, and STAN [15-18]. Several books have been published to introduce Bayesian method by adopting these tools and their modeling languages [19-20]. Basically, these tools may be limited by their underlying sampling engines. It is worth to mention that a commercial software, called Mathematica, provides most of the required functions for MCMC, except the build-in sampling algorithm [21]. Some researchers may develop their own sampler for work or research [22]. One purpose of this research is to introduce a practicable implementation for the algorithm.

As a special kind of MCMC, Gibbs sampler has been widely used in Bayesian and reliability engineering for parameter estimation and hypothesis testing [23]. However, it often involves non-trivial technical details in the implementation and may suffer from convergence problems [24]. Therefore, this case study is based on a multidimensional metropolis sampler, which requires only the derivatives of probability (density) functions.

The most often used proposal distributions of MCMC are uniform or normal for random walk, which may cause inefficiency due to the curse of dimensionality for high-dimensional probability models. To tackle the problem, Hamiltonian Monte Carlo (HMC) is proposed by utilizing the gradient of potential energy (corresponding to probabilistic density) to generate simulation trajectory and conditionally accept new samples based upon the difference of the total energy [25][26][27].

One major criticism of the algorithm is that the acceptance probability is a constant 1 theoretically, as the simulation conserves total energy according to Hamiltonian dynamics. In addition to the problem of energy, the sampling effect is largely governed by simulation trajectories, which depends on ~~step size~~step-size, simulation steps, and choice of kinetic energies [28][29][30]. Therefore, the samples' quality is susceptible to several run-time parameters, especially ~~step size~~step-size and simulation steps [31]. The approach of the No-U-Turn Sampler (NUTS) is to adjust ~~step size~~step-size in advance and tune simulation steps for each sampling, until encountering a spatial U-turn [32]. Excessive simulation steps are a drawback of this approach, resulting in slower algorithm.

In general, the simulation trajectory may be somewhat inconsistent with the target probability density, therefore, the effectiveness is often hampered, especially for some extreme conditions. A method concerning the problem, called Riemann Manifold Hamiltonian Monte Carlo (RMHMC), can generate a different kind of trajectory by proposing a new kinetic energy function [33].

A lesser-known fact about RMHMC, however, is the improvement can only be guaranteed on the first principal component. This paper examines properties and consequences of simulation trajectories, focusing specifically on the first and last principal components. Furthermore, a series of bi-variate normal of increasingly high correlations have been employed to investigate the effectiveness of the proposed approach.

The proposed method comprises the following parts: (1) to exchange energy among particles of the system though virtual collision; (2) to substitute the controversial acceptance probability with strict Metropolis by maintaining a constant total energy; (3) to speeding up, fixes simulation steps, and adjusts ~~step size~~step-size based on the extremes of potential energy; (4) to adapt the acceptance probability in accordance with the absolute kinetic energy; (5) to simulate alternately with two kind of trajectories. In the experimental section, the proposed method is applied to study some reliability or survival models that based on the Weibull, exponential, or binomial distribution.

II. Hamiltonian Monte Carlo

The basic concept of the method is to sample on the simulation trajectory in agreement with the Metropolis-Hastings algorithm. The algorithm can be decomposed into subparts of Hamiltonian and Metropolis. Since both can notably influence the sampling outcome, therefore they will be examined respectively in the following sections to reveal some latent characteristics.

A. Hamiltonian Simulation

According to Hamiltonian dynamics (2), there is a close relation between the rate of change of phrase variables (momentum p and position q) and the gradient of energies (total energy H, potential U, and kinetic K). If K is also the function of q, its partial derivative with respect to q can be nonzero. The additional term of ∂K/∂q can be omitted for efficiency in exchange for additional systematic error.

$$\begin{cases} \frac{dp}{dt} = -\frac{\partial H}{\partial q} = -\frac{dU}{dq} - \frac{\partial K}{\partial q} \\ \frac{dq}{dt} = \frac{\partial H}{\partial p} = \frac{\cancel{dK}}{\cancel{dp}}\frac{\partial K}{\partial p} \end{cases} \quad (2)$$

From (3) we can see that the time derivative of total energy is constant 0 after being substitut~~ed~~ed~~ing~~ in Hamiltonian equations. This indicates a constant energy on the simulation trajectory. Thus, although HMC and RMHMC do not maintain constant total energy for different simulations, owing to ~~the~~ this ~~mentioned~~ reason, two terminals of the trajectory possess the same total energy.

$$\frac{dH(q(t),p(t))}{dt} = \frac{\partial H}{\partial q}\cdot\frac{dq}{dt} + \frac{\partial H}{\partial p}\cdot\frac{dp}{dt} = \frac{\partial H}{\partial q}\cdot\frac{\partial H}{\partial p} - \frac{\partial H}{\partial p}\cdot\frac{\partial H}{\partial q} = 0 \quad (3)$$

For a particular potential energy, the simulation trajectory is mainly determined by kinetic energy function. Because the sampling is carried out on the trajectory, therefore consequently, kinetic energy function has an effect on the correctness of result. Aiming to improve the effectiveness of sampling for high dimensional probability density, it is thus necessary to enquire into the interrelationship between kinetic energy and trajectory.

Two kinds of widely used kinetic energies are demonstrated in (4) and (5). The quadratic function for kinetic energy (4) adopted by HMC solely depends on the current momentum. In contrast, the major innovation of RMHMC comes from its kinetic energy (5), which is subject to both momentum and position.

带格式的： 字体: 倾斜, 复杂文种字体: 倾斜

带格式的： 字体: 倾斜, 复杂文种字体: 倾斜

带格式的： 字体: (默认) +中文正文 (等线), (中文) +中文正文 (等线)

$$K(p,q) = \frac{1}{2}p^T p \quad (4)$$

$$K(p,q) = \frac{1}{2}p^T (\frac{d^2U}{dq^2})^{-1} p \quad (5)$$

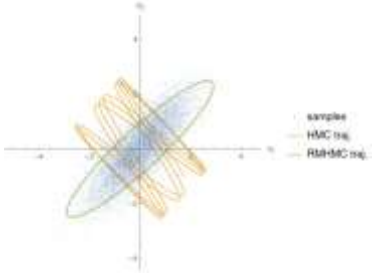

Fig. 1. Scatter plot and trajectories for a correlated bi-variate normal. The orange curve is the simulation trajectory of HMC, and the green curve is for the RMHMC, the blue dots indicate the scatter plot of the correlated bi-variate normal.

The difference can be better illustrated with an example. Fig. 1 presents the scatter plot and simulation trajectories for a correlated bi-variate normal density. The trajectory of HMC colored in orange is inclined to progress along the last principal component (i.e., the component with the lowest variance), yet the progression is much slower upon the other direction. The reason for the phenomenon is that the direction of the last principal component conforms to the gradient of potential energy, namely acceleration, and acceleration in turn affects trajectory. On the other side, the scheme of RMHMC is to instantly modify the direction of momentum by utilizing Hessian of potential energy during simulation, thereby forming another kind of trajectory. The trajectory of RMHMC colored in green can traverse upon the first principal component preferentially, therefore remedies the deficiency of HMC. However, the incidental drawback may be the insufficient ~~slower~~ progression upon the ~~direction of the~~ last principal component.

B. Metropolis Algorithm

As illustrated in Fig. 2, the trajectory of a Hamiltonian simulation starts from q_0 with an initial momentum p_0 and arrives q_1 with end momentum p_1.

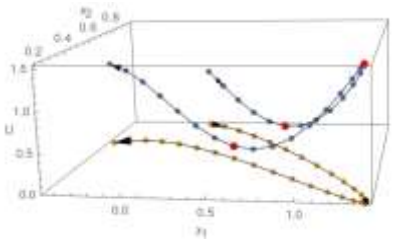

Fig. 3. The curve with orange dots is a spatial trajectory; the curve with blue dots is the joint spatial–potential trajectory; the red dots indicate extremes of potential energy; the black dot is the spatial U-turn; the arrows show the simulation direction.

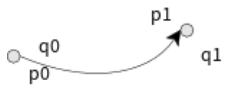

Fig. 2. The schematic diagram of a Hamiltonian simulation.

As a key component of MCMC, the Metropolis algorithm is often utilized to sample from proposal probability on a Markov chain. In the case of the balanced proposal probability, the first step is to walk randomly from q_0 to q_1, and the second step is to conditionally accept the new sample as stated by (6), in which π refers to an unnormalized posterior probability density [34].

$$\alpha = \min[1, \frac{\pi(q_1)}{\pi(q_0)}] = \min[1, e^{U(q_0)-U(q_1)}] \quad (6)$$

The prerequisite of the algorithm is that the probability of moving from q_0 to q_1 equals the reverse move. In case the condition is not met, it will be necessary to resort to the Metropolis–Hastings algorithm (7), in which $P(q_1 \to q_0)$ indicates the proposal distribution (or "jump" probability) moving from q_1 to q_0. If the mutual proposal probability between q_0 and q_1 is equivalent, then the Metropolis-Hastings will reduce to the Metropolis algorithm.

$$\alpha = \min[1, \frac{\pi(q_1)P(q_1 \to q_0)}{\pi(q_0)P(q_0 \to q_1)}] \quad (7)$$

As shown in (8), the acceptance probability of HMC and RMHMC has been proposed as a function of the difference of total energy between two terminals [35][36]. However, as Hamiltonian simulation conserves total energy, the acceptance probability is a constant *1* in theory. In other words, the terminal q_1 of the trajectory is treated as a new sample without regard to the Metropolis-Hastings. Treating the end terminal as a new sample may lead to logical inconsistency, as the probability to be sampled should be only subject to potential energy, whereas the trajectory of HMC (or RMHMC) also depends on the kinetic energy; therefore, trajectory cannot coincide with target probability. Considering that the acceptance probability is actually close to a constant *1*, some researchers seem to suggest disregarding Metropolis and acquiring samples from trajectory directly [37].

$$\alpha = e^{H(q_0,p_0)-H(q_1,p_1)} = e^{U(q_0)+K(p_0)-U(q_1)-K(p_1)} = 1 \quad (8)$$

C. ~~Step size~~Step-size and steps

The Hamiltonian simulation is generally very sensitive to the distance traveled, i.e., the multiplication of ~~step size~~step-size and steps. The time step has been identified as one of the major configurable parameters of the algorithm and is relevant to not only stability of simulation but also effectiveness of sampling. An overly small ~~-~~time ~~-~~step can substantially slow down the simulation, yet an unnecessarily large time step may result in instability.

NUTS provides a scheme to automatically adjust the number of steps (and ~~step size~~step-size) for HMC and RMHMC.

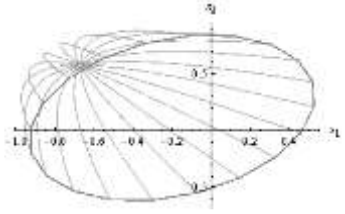

Fig. 4. Trajectories with the same total energy and initial position. The terminals are connected as an ellipse.

The basic idea is to treat spatial U-turn as the stop criterion. Fig. 3 shows a simulation trajectory ~~which~~ is linked by the orange dots in x_1 ~~-~~x_2 plane. NUTS tries to determine a proper step at the beginning. It then detects spatial U-turn for each subsequent simulation trajectory, exemplified by the black dot. As the simulation (of a single particle) is very sensitive to the length traveled, more accurate detection can be obtained with smaller step and more steps at the cost of computation.

Considering energy may produce a better criterion. The trajectory with blue dots illustrates the corresponding spatial ~~-~~potential trajectory. The red dots indicate the locations of either minimal or maximal energy. Compared with this, the spatial U-turn depends on a single measure. ~~Obviously~~In this example, the trajectory has only one U-turn, but has three extremal energies. Thus, the criterion ~~criteria~~ of spatial U-turn may be weaker than ~~-~~that of the extreme of energy.

III. Complementary Hamiltonian Monte Carlo

This research seeks to address the problems of classic method with the scheme of conservation of energy, virtual collision, and alternating trajectories.

A. The Conservation of Total Energy

In this research, we adopt the acceptance probability of the Metropolis, which is related with the difference of potential energies (9). The precondition of Metropolis, the equi-probability of mutual jumps of the two position, can be implemented with energy conservation.

$$\alpha = \min[1, e^{U(q_0)-U(q_1)}] \quad (9)$$

Despite the fact that Hamiltonian simulation conserves total energy, the multiple simulations carried out by the HMC algorithm do not maintain a constant total energy. Therefore, two consequent simulations share a same terminal, which corresponds to different total energies, introducing an additional state transition which needs an explicit Metropolis-Hastings sampling. To avoid the trouble brought upon by the complex situation, this study makes use of the same initial total

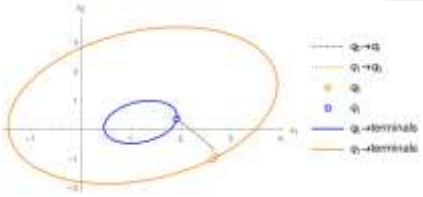

Fig.5. HMC trajectories between two positions, and terminal surfaces. The blue ellipse is the terminal started from q0, and the orange is the terminal started from q1. The dashed lines are one of such simulation trajectories.

energy for all simulations.

On account of the fact that potential energy only depends on the position, adopting a constant total energy indicates that the initial kinetic energy is also fixed. Thus, in the random initialization of the momentum, only the direction can be ~~changed~~changed, whereas the magnitude of momentum is exclusively determined by the initial kinetic energy, which can be calculated using total energy and initial potential energy.

Fig. 4 displays multiple trajectories with the same total energy and initial position. The simulation is carried out using the kinetic energy defined in (4) and the potential energy defined in (10) (ρ=0.9). The initial momentum of these trajectories possesses the equally spaced angles of direction and the same kinetic energy, and a closed hyper-surface (closed curve for the two-dimensional case) is formed by the terminal locations. The figure reveals that if the orientation of the initial momentum is evenly distributed, then the terminal positions will be spread unevenly on a hyper-surface. In broad terms, the Hamiltonian simulation with constant total energy may be regarded as a random walk upon the angular dimension of the initial momentum, which is analogous to the random walk of MCMC.

$$U = \frac{1}{2}(x_1, x_2)\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad (10)$$

Adopting the same energy functions as before, Fig. 5 presents two positions with the same total energy and their mutual simulation trajectories. Starting from q_0 and taking arbitrary momentum with designated kinetic energy, within the same simulation steps, the terminals make up a spatial surface colored blue, and any position (e.g., q_1) on the surface possesses the same total energy as q_0. Since q_1 possesses the same total energy, its kinetic energy can be calculated straightforwardly. Starting from q_1, and drawing a random momentum with the calculated kinetic energy, the new terminals also form a surface and q_0 is located on it. Besides above properties, another noticeable feature is that the reciprocal simulation trajectories between q_0 and q_1 are coincidence. This indicates that with the same simulation steps, if a simulation trajectory starts from q_0 with the initial momentum p_0 and arrives in q_1 with terminal

momentum p_1, then there is a reverse simulation starts from q_1 with initial momentum $-p_1$ and arrives in q_0 with terminal momentum $-p_0$. Similar discussion can be found in a textbook [38].

To apply Metropolis algorithm, it is necessary to clarify the equi-probable jump between q_0 and q_1, as illustrated in a two-dimensional space. If the simulation starts from q_0 with momentum p_0 and terminates at q_1, then p_0 can be created as follows:

Algorithm 1: the procedure for generating p_0 and q_1

1) choose an arbitrary angle for p_0 ($0<\theta<2\pi$);
2) regulate the magnitude of p_0 according to the expected kinetic energy;
3) carry out Hamiltonian simulation starting from q_0 with the momentum p_0;
4) after a specified period of time, the simulation terminates at q_1 position.

As p_0 is drawn from a uniform distribution on the interval $[0,2\pi]$, thinking of the one-to-one correspondence between p_0 and q_1, thus the jump probability from q_0 to q_1 follows the uniform distribution. Likewise, the jump probability from q_1 to q_0 is also uniformly distributed on the same interval. As a result, the mutual jump probabilities between q_0 and q_1 are identical, thus, Metropolis algorithm is applicable for the subsequent sampling.

The foregoing discussion does not restrict the form of kinetic energy; therefore, it is appropriate for any kind of kinetic energy. The results of simulation using RMHMC are displayed in Fig. 6. The difference between this and previous result is originated from the kinetic energy. Owing to the fact that the trajectory conforms to the first principal component, therefore a highly correlated probability can be applied with RMHMC samplers.

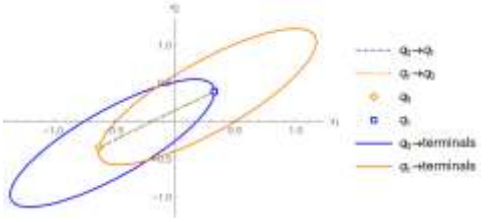

Fig. 6. RMHMC trajectories between two positions, and terminal surfaces. The blue ellipse is the terminal started from q_0, and the orange is the terminal started from q_1. The dashed lines are one of such simulation trajectories.

B. Multi-particle system

One reason why classic HMC or RMHMC disregard energy conservation may be the necessity of balancing the stability and exploration. As their system contains only a single particle, fixing its total energy will seriously limit exploration. To reconcile the controversy, our proposal is to introduce multiple particles and exchange their energy though "collision". In this way, the total energy of the system is conserved whereas the energy of each particle can vary extensively, thus ensuring both adequate exploration and overall stability.

Accurate collision detection requires tiny simulation ~~step, which~~step, which will considerably lower computational efficiency. In effect, the only necessity is to redistribute energy among particles. Therefore, "virtual collision" is carried out to ensure: 1) conservation of system energy during collision, 2) the direction of momentum is equi-probable after colliding. Intuitively, the particles of system are indistinguishable, and the two terminals of trajectories have equal jump probability. Consequently, Metropolis algorithm can be applied.

C. Alternating Trajectories

To sample accurately in all dimensions, both trajectories are considered in the algorithm. The proposed approach is to choose the Hamiltonian or Riemannian simulation with equal probability, to carry out the simulation, and to sample with the Metropolis algorithm. The rationale is that the simulation is based on the previous sample, which can be generated from either HMC or RMHMC. Adopting two kind of complementary trajectories, the samples can approach target probability density gradually to maintain the benefit of both trajectories and overcome their deficit.

D. Algorithm Description

Algorithm 2 presents the proposed algorithm that conserves the total energy. From the line 8 to line 17, the momentum of each particle is initialized, and the total kinetic and potential energies ~~for~~ of the system are also accumulated here. The designated overall kinetic energy is calculated in line 19 or line 21 depending on ~~the~~ the type of the trajectory. The magnitude of each particle's momentum is regulated in line 25 to ensure the conservation of system energy. Starting from line 28, Hamiltonian simulation using Euler integration and Metropolis sampling are carried out for each particle. For better effect, HMC and RMHMC are controlled with respective step δ and system energy H. In line 39, the potential energy is saved during simulation for subsequently tuning step~~-~~size. The number of steps, the constant S, can be as small as 3, since the step~~-~~size is adjusted instead. Fixing the number of steps is to reduce computation load. The main novelty for this is the conservation of total energy of the system and the alternating algorithm. The adjustment of the energy and step~~-~~size is associated with implementation details, and will be discussed separately.

Algorithm 2: the proposed method

1	-- initializing –
2	initialize H_{HMC}, H_{RMHMC}, δ_{HMC}, and δ_{RMHMC}
3	method=HMC
4	for i = 1 to N
5	U_{total}=0
6	K_{total}=0
7	-- M particles--
8	for j=1 to M
9	-- equal probability for momentum--
10	p_j=NormalRandom(0,1)
11	if method is HMC
12	$K_{total} = K_{total} + \frac{1}{2} p_j^T p_j$
13	else

```
14          K_total = K_total + (1/2) p_j^T (d^2U/dq_j^2)^-1 p_j
15        end
16        U_total=U_total+U(q_j)
17      end
18      If method is HMC
19        K_d=H_HMC-U_total
20      else
21        K_d=H_RMHMC-U_total
22      end
23      -- total energy conservation for the system –
24      for j=1 to M
25        p_j = p_j sqrt(|K_d / K_total|)
26      end
27      -- M particles --
28      for j=1 to M
29        q_0=q_j
30        -- Hamiltonian simulation --
31        for s = 1 to S
32          if method is HMC
33            p_j = p_j − δ_HMC dU/dq_j
34            q_j = q_j + δ_HMC p_j
35          else
36            p_j = p_j − δ_RMHMC dU/dq_j
37            q_j = q_j + δ_RMHMC (d^2U/dq_j^2)^-1 p_j
38          end
39          save U(q_j)
40        end
41        -- Metropolis --
42        α = e^(U(q_0)−U(q_j))
43        if α_j < UniformRandom(0,1)
44          q_j=q_0
45        end
46        save q_j
47        save α_j
48      end
49      -- tuning --
50      if method is HMC
51        tune H_HMC and δ_HMC
52        method=RMHMC
53      else
54        tune H_RMHMC and δ_RMHMC
55        method=HMC
56      end
57    end
```

E. Algorithm Parameters

The total energy as well as the ~~step size~~step-size is required to be determined for the proposed methods. The step-~~–~~size is regulated based on the energy extremes for fixed simulation strides, whereas the energy can be adjusted ~~by setting the magnitude of the initial momentum~~according to the average acceptance probability of all particles. Generally speaking, the simulation and acceptance probability are very sensitive to the length traveled, so suitable parameters are crucial. It is found that coarse tuning the ~~step size~~step-size and fine tuning the total system energy can achieve the designated average acceptance probability.

The time-step is adjusted according to the occurrence of maximal and minimal ~~engines~~energies. In contrast to NUTS, multiple particles are considered, so the tuning is more stable. Whenever the ~~step size~~step-size is unsuitably small, the extreme potential energy will take place ~~in~~ at the start or end of the trajectory. On the contrary, if the ~~step size~~step-size is excessively large, the minimal potential energies will happen at the beginning terminals and maximal energies at the end terminals. It is small probability event that the phenomenon occurs for all particles simultaneously, in which case, the ~~step size~~step-size needs to be adjusted.

This study explores the relationship between the average acceptance probability and (absolute) kinetic energy for the system. It is apparent from intuition that excessively large kinetic energy will drive the simulation to the region of low probability and therefore issues a low acceptance probability. Thus, a small average acceptance probability implies excessively large kinetic~~an overly large average acceptance probability indicates an insufficient total kinetic~~. Contrariwise, ~~a small value implies excessively large kinetic~~ an overly large acceptance probability indicates an insufficient total kinetic. The absolute value accounts for the negative kinetic energy caused by the negative definite of the hessian matrix.

Current implementation is to adapt with the proportion of 0.1, causing exponential trends, namely 1.1^{n} or 1.1^{-n}, which span huge range for hundreds of n. Applications and experiments reveal that proper tuning can be achieved in hundreds of iterations in most cases. Owing to the scheme of multiple particles, the adjustment based on their statistics ensure success with high probability.

F. Partial implementation

The following code fragment outlines the Mathematica implementation. Two helper functions for derivative are defined in line 2 and line 3. Line 7 to line 9 is the potential function and its first and second order derivatives. Defining a potential function is the only requirement for the sampling, for derivatives can be computed automatically by the help of Mathematica. The simulation is carried out from line 12 to line 15. The acceptance probability of Metropolis, listed in line 17, is based on the difference of potential energies. It is deviated from the classic approach of HMC and RMHMC, which is based on the difference of total energies.

Code: basic implementation

```
1   (* helper functions for derivatives *)
2   HessianH[f_,x_List?VectorQ]:= D[f,{x,2}];
3   GradientG[f_,x_List?VectorQ]:= D[f,{x,1}];
4   (* potential energy and its derivatives *)
5   rho = 1 - 1/10^9;
6   SIGMA = {{1, rho}, {rho, 1}};
7   U[x_, y_] = 1/2 {x, y}.LinearSolve[SIGMA, {x, y}];
```

带格式的: 字体: 倾斜, 复杂文种字体: 倾斜

带格式表格

带格式的: 缩进: 首行缩进: 0 厘米

带格式的: 缩进: 首行缩进: 0 厘米

带格式的: 缩进: 首行缩进: 0 厘米

带格式的: 缩进: 首行缩进: 0 厘米

带格式的: 缩进: 首行缩进: 0 厘米

带格式的: 缩进: 首行缩进: 0 厘米

带格式的: 缩进: 首行缩进: 0 厘米

```
dU[x_, y_] = GradientG[U[x, y], {x, y}];
ddU[x_, y_] = HessianH[U[x, y], {x, y}];
(* Hamiltonian simulation *)
q0=q;
Do[p = p - dt Apply[dU, q];
    q = q + dt If[method == HMC, p,
                LinearSolve[Apply[ddU, q], p]],
     STEPS];
(* Metropolis *)
a=Exp[Clip[Apply[U, q0]-Apply[U,q],{-20, 0}]];
If[a<RandomVariate[UniformDistribution[]],q=q0];
```

IV. Experiments and Analysis

In the experimental section, it is shown that the improved HMC has the capability to estimate parameters of probability models. Firstly, limitation of classic methods is illustrated with the bi-variate normal distribution and a ring-shaped distribution. Secondly, the proposed method is applied to study reliability using Weibull, Hierarchical Weibull Model, and Hierarchical Binomial Model. Finally, the remaining useful life is predicted for the proportional hazards (Cox) model using the simulated data.

A. Bi-variate Normal Distribution

Fig. 7 displays the scatter plots for the samples of a bi-variate normal density using the proposed method. As the transformed samples resemble the standard bi-variate normal, the sampling results resemble the target probability density.

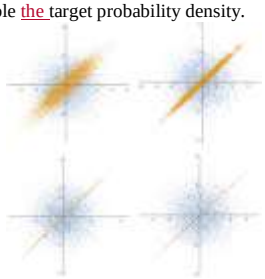

Fig. 7. Samples (orange) and transformed samples (blue) of the correlated bi-variate normal distribution using (10) with ρ= 0.9, 0.99, 0.9999, and 0.99999999.

In order to assess the accuracy of the sampling, the error is measured as the relative entropy (KL-divergence) between the fitted sampling distribution q(x) and the target distribution p(x):

$$\int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx \tag{11}$$

For the bi-variate distribution adopted here, each dimension needs a KL-divergence respectively. The covariance matrix is formed as $\begin{bmatrix}1 & 0\\ 0 & \sigma^2\end{bmatrix}$, in other words, the standard deviation of the first (and last) principal component is 1 (and σ). By varying σ, it is possible to study the unitary sampling error of each dimensions, ~~which is characterized by the KL-divergence between sampling and target distributions~~.

Fig. 8 draws a distinction among methods for the bi-variate normal on the first principal component, with greater value of KL-divergence representing larger error of the samples. HMC0 and RMHMC0 denote the classical algorithms, whereas HMC and RMHMC are the modified algorithms that conserving total energy. CHMC (Complementary Hamiltonian Monte Carlo) denotes the proposed method that conserving energy and alternating trajectories.

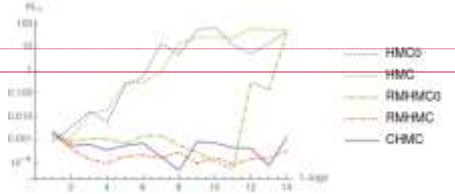

Fig. 8. The error on the first principal component.

HMC0 suffers from imprecision on the first principal component, furthermore, the error increases substantially with the decreasing σ. HMC shows a similar trend. On the contrary, both RMHMC0 and RMHMC generate low error for a wide range of σ. If σ becomes extraordinarily small, the error of RMHMC0 will soar highly, implying the inherent instability of the classic approach. Both RMHMC and CHMC maintain a low error, indicating the improved stability for the proposed approach on the first principal component.

Similarly, the estimation error on the last principal component is revealed in Fig. 9, in which σ decreases to the right. It clearly points out that, contrary to previous figure, both RMHMC0 and RMHMC generate large error and both HMC0 and HMC achieve small error. The reason for this phenomenon is that the trajectory of HMC conforms to the last principal component and of RMHMC complies with the first principal component. As the methods that conserving energy acquire lower error for small σ, they are evidently more stable than its classic counterparts. HMC0 is capable to maintain small error for relatively large σ. However, as σ is reduced to a certain degree, the error of HMC0 escalates quickly. In contrast, within the same range, HMC preserves small error; its error will only increase with an even smaller σ. On the other hand, CHMC mainly preserves low error rate. Thus the proposed algorithm is more stable on the last principal component.

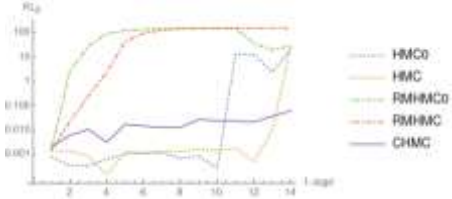

Fig. 9. The error on the last (i.e. 2nd) principal component.

Comparing Fig. 8 with Fig. 9, it is clear that both HMC and RMHMC will be imprecise on certain principal components and more accurate on the other principal component. Under

带格式的： 缩进: 首行缩进: 0 厘米
带格式的： 缩进: 首行缩进: 0 厘米
带格式的： 缩进: 首行缩进: 0 厘米
带格式的： 缩进: 首行缩进: 0 厘米
带格式的： 缩进: 首行缩进: 0 厘米
带格式的： 字体: (中文) +中文正文 (等线)
带格式的： 缩进: 首行缩进: 0 厘米
带格式的： 缩进: 首行缩进: 6.51 字符
带格式的： 缩进: 首行缩进: 0 厘米
带格式的： 两端对齐
带格式的： 字体: (默认) +标题 CS (Times New Roman), 复杂文种字体: +标题 CS (Times New Roman)
带格式的： 缩进: 首行缩进: 0 厘米
带格式的： 字体: (中文) +中文正文 (等线), (中文) 中文(中国)
带格式的： 缩进: 首行缩进: 0 厘米
带格式的： 字体: (默认) +标题 CS (Times New Roman), 复杂文种字体: +标题 CS (Times New Roman)

same experimental conditions, similar or more stable result is achievable for the proposed methods that conserve energy. Overall, the method alternating two trajectories and conserving total energy can obtain low error simultaneously on both principal components.

A noteworthy benefit of the energy-conserved approach is the promise of adjusting acceptance probability though total (or kinetic) energy. There is hardly any direct way for the traditional samplers to control acceptance probability, which can only be influenced by other indirect means (such as ~~step size~~step-size). As mentioned early, the theoretical acceptance probability of the classic methods is constant 1. Accordingly, if the ~~step size~~step-size is suitably tuned, the actual acceptance probability might approach constant 1. Usually, choosing a suitable ~~step size~~step-size is a difficult task for HMC0 as its running is often correlated with ~~step size~~step-size. Fig. 10 illustrates that when the ~~step size~~step-size is set to 10^{-4}, the average acceptance probability for HMC0 approaches zero, indicating excessively large ~~step size~~step-size.

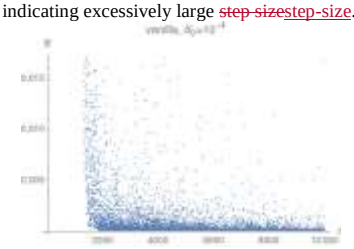

Fig. 10. The average acceptance probability for the HMC0 setting initial ~~step size~~step-size to 10^{-4}.

Under the same condition, if the ~~step size~~step-size is set to smaller value, such as 10^{-9}, the acceptance probability will be marginally enhanced as in Fig. 11, although still unstable and overly small. One main deficit of HMC0 is the sensitivity to ~~step size~~step-size: different value may lead to exceedingly low or high acceptance probability.

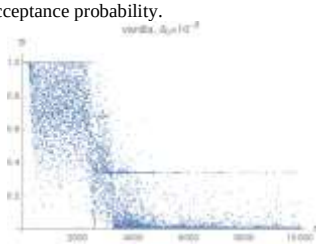

Fig. 11. The average acceptance probability for the HMC0 setting initial ~~step size~~step-size to 10^{-9}.

As shown in Fig. 12, for the proposed CHMC, the average acceptance probability is more evenly distributed. The experiment uses three particles. Whenever the acceptance probability of one particle approaches 1 and others approach 0, the average acceptance probability will be about 0.33. The cases of 0.66 and 1 are similar. The chains are simulated with a burn-in period of 5000 iterations, causing variation around 5000th

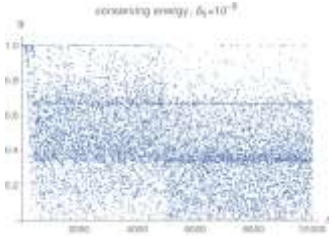

Fig. 12. The average acceptance probability for the proposed method setting initial ~~step size~~step-size to 10^{-9}.

iteration. The reason of better result, owing to Metropolis algorithm, is that the acceptance probability is not constant 1 theoretically. Due to the tiny initial ~~step size~~step-size, ~~the~~the average acceptance probability starts near 1. After hundreds of iterations, its value is almost scattered evenly. This research proposes that adjusting energy according to average acceptance probability of multiple Markov chains (i.e., particles) can align the acceptance probability with some target value or range. A tiny acceptance probability implies an excessively large kinetic energy, which should be reduced, and vice versa. In order to adapt energy steadily, also it may be necessary to allow the average acceptance probability to vary in certain range, such the range of [0.3, 0.7] or [0.1, 0.9]. In other words, it will influence sampling effect to restrict ~~the~~ 99% of the acceptance probability to [0.49,0.51].

B. Ring-shaped distribution

As shown in (12), the radius of a ring follows a normal with mean of 10 and standard deviation of σ. To fully examine the characteristics and effect of the trajectories, different distribution will be generated according to σ to study the sampling effect of simulation trajectories. Although it is not applicable to principal components anymore, it can be described using polar coordinates. The new variables, radius and angle, is transformed from x and y. The radius coordinate follows a normal distribution, while the angular coordinate is a uniform.

$$e^{-U(x,y)} = e^{\frac{-(\sqrt{x^2+y^2}-10)^2}{2\sigma^2}} \tag{12}$$

Fig. 13 demonstrates the trajectories of a ring-shaped distribution ~~and the corresponding trajectories~~. The trajectory of HMC mainly traverses along the radial coordinate, moreover, as shown in the following experiment, its sampling on this coordinate is more accurate. In contrast, the trajectory of RMHMC mainly progresses around the ~~outer~~ circle, therefore, the sampling on the angular coordinate is more accurate. Additionally, —the outward trajectory caused by ignoring

∂K/∂q may imply a systematic error.~~It is apparent from the trajectories that HMC and RMHMC are roughly complementary to each other as each trajectory preferably moves along one of the coordinates.~~

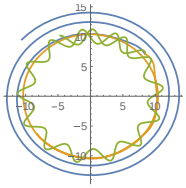

Fig. 13. Trajectories for a ring-shaped distribution,~~:~~ blue: RMHMC without ∂K/∂q, orange: corrected RMHMC, green: HMC.

Fig. 14 displays the error on the angular coordinate. When σ is decreased, the error of both HMC0 and HMC rise quickly. By contrast, both RMHMC0 and RMHMC attains smaller error on this coordinate. HMC and RMHMC perform slightly better than their classic counterparts for small σ.

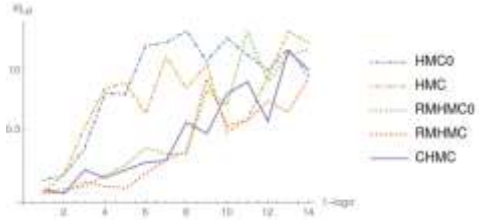

Fig. 14. Error on the angular dimension.

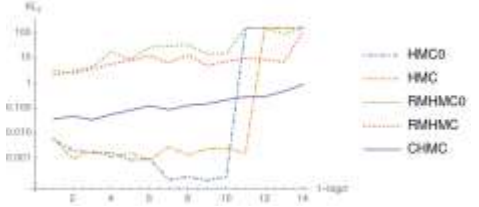

Fig. 15. Error on the radial dimension.

Turning now to the error on radial dimension (Fig. 15), it is apparent that HMC0 and HMC maintain significantly smaller error than RMHMC0 and RMHMC in a wide range of σ. However, as σ decreases furthermore, the error of the classic methods will inflate. By contrast, the proposed methods that conserve energy remain moderately stable. One notable aspect of this graph is that, for a certain range of σ, HMC0 may obtain smaller error than HMC on the radial dimension; however, the benefit depends on initial condition, and may be offsetted by angular coordinate. In addition, as far as we know, the benefit may not always be possible, especially for higher dimension models.

It may be unnoticeable that both RMHMC and RMHMC0 maintain a high error on the radical coordinate owing to the systematic error due to lacking ∂K/ ∂q of (2).~~of Euler integration.~~ The samples of RMHMC incline outward, instead of balancing around the ring. It indicates that systematic error may be introduced by original scheme of RMHMC for non-normal models.

In summary, the results of bi-variate normal and ring-shaped distributions suggest that the proposed approach is equivalent or more stable than its classic counterpart. In cases that the classic samplers fail to converge, the proposed approach will be a feasible alternative.

C. Weibull distribution

It is widely held that the two-parameter Weibull distribution is a particularly difficult case, because it requires a two-dimensional joint prior distribution of the Weibull parameters [39]. Nonetheless, these parameters can be sampled out in a straightforward manner using the proposed method. Suppose the prior of parameters of Weibull are uniformly distributed in the interval [0,1000]. Having sampled from the posterior using completed or censored data, we can compare the result with frequentist method and Gibbs sampler (JAGS [40]).

1) Frequentist Method

The location of the parameters of Weibull can be estimated using maximum likelihood method. A convenient approach to estimate the variance of the parameter is to use the Hessian matrix (13) to compute the covariance (14).

$$\frac{\mathrm{d}^2\mathrm{U}}{\mathrm{d}\theta^2} = \begin{bmatrix} \frac{d^2U(\theta)}{d\theta_1^2} & \cdots & \frac{d^2U(\theta)}{d\theta_1 \partial\theta_p} \\ \vdots & \ddots & \vdots \\ \frac{d^2U(\theta)}{d\theta_p \partial\theta_1} & \cdots & \frac{d^2U(\theta)}{d\theta_p^2} \end{bmatrix} \tag{13}$$

$$\mathrm{V} = \left(\frac{\mathrm{d}^2\mathrm{U}}{\mathrm{d}\theta^2}|_{\hat{\theta}}\right)^{-1} = \begin{bmatrix} Var(\hat{\theta}_1) & \cdots & Cov(\hat{\theta}_1, \hat{\theta}_p) \\ \vdots & \ddots & \vdots \\ Cov(\hat{\theta}_1, \hat{\theta}_p) & \cdots & Var(\hat{\theta}_p) \end{bmatrix} \tag{14}$$

Thus the confidence interval of the parameters can be derived in the following steps: firstly, to find the maximum likelihood estimate, and then to compute the covariance matrix, finally to compute the confidence interval with normal or student's *t* distribution using the diagonal elements of the covariance matrix [41].

2) Censored Data

One major distinction between reliability models and ordinary probability models is the presence of censoring. Right censoring occurs when an item has not failed by the last inspection [42]. If the lifetime of a product is modeled by probability density function of f(t), then the probability of not observing a failure in [0, t_L] is 1-F(t_L), where the F(t) is the cumulative distribution function of f(x). Suppose M indicates the number of failed products and N for the number of workable products by the last inspection. Considering both cases, the total likelihood follows:

$$f(t|\theta) = \prod_{i=1}^{M} f(t_i|\theta) \times \prod_{j=1}^{N} (1 - F(t_j|\theta)) \quad (15)$$

To reveal the influence of heavily censoring, 100 items are sampled according to the above mentioned true value. The values greater than 100 are right-censored, and 3 uncensored items are found. Fig. 16 shows the scatter plot and histograms for the parameters of the proposed method. The scatter plot demonstrates a strong non-Normal characteristic and spreads in both dimensions contrasted with the completed data case. As the parameters are actually bi-variate correlated, sampling directly in the two-dimensional space is possibly more accurate than sampling each dimension individually, which exemplifies the potential advantage of HMC over Gibbs sampler.

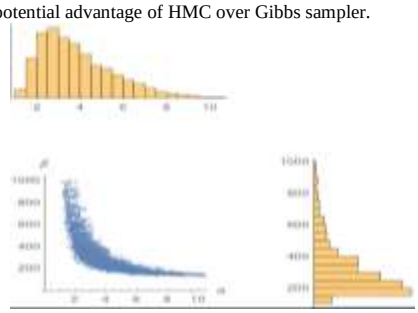

Fig. 16. Sampled Parameters of the Weibull for heavily censored data.

It is conceivable that non-Normal or highly correlated multivariate probability density may cause problems for Gibbs sampler. To illustrate an example, the type II right censoring is carried out for the experimental data by censoring 5,36 ,68 ,84 ,92, 96 or 98 items in a total 100 products. In this way, we can reveal the relation between the width of credible (and confidence) interval and the level of censoring. Fig. 17 shows, raising the level of censoring, both width of the credible interval of the proposed method (blue) and the confidence interval (gray) of the frequentist method increase. Attributing to Bayesian methodology, the proposed method has narrower interval therefore are more accurate. Yet, although the Gibbs sampling with JAGS (red) is also based on the same Bayesian model, the result is imprecise due to the fact that its credible interval does not contain the true value of parameter (dashed line). Only for low level of censoring, the credible interval of JAGS can be coincidence with the true value.

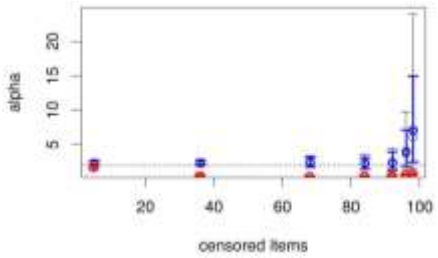

Fig. 17. 95% interval under various censoring conditions for the *a* parameter, grey: frequentist confidence interval, blue: credible interval of the proposed method, red: credible interval using Gibbs sampler (JAGS). The horizontal dash line shows *a*'s true value. The dots represent median.

Fig. 18 displays the similar relation between the level of censoring and the width of interval for scale parameter β. For slightly censored data, all three intervals are near the parameter's true value, despite the interval of Gibbs sampler does not contain the true value. With more censored data, both interval of the proposed and the frequentist method widen, but still contain the parameter's true value. On the other hand, the interval of Gibbs sampler is deviated from the true value, suggesting uncorrected sampling result. As shown in the rightmost of the figure, only for very high level of censoring, both interval of the proposed and the frequentist method do not contain the true value. The example indicates that potential problem may exist for applying Gibbs sampler to Weibull models with heavily censored data.

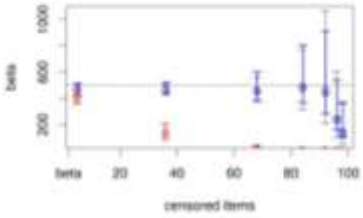

Fig. 18. 95% confidence and credible interval for the β parameter.

D. Hierarchical Weibull Model

It has been shown that hierarchical Weibull model is applicable to estimate the reliability of a product family [43]. For this and next experiments, the provided model, data, scripts, and true value of parameters are used [43]. From each of the ten types covered, 100 products are randomly picked out and tested regularly within a time period. According to (16), the time to fail of the i^{th} item (ttf_i) follows Weibull distribution. Furthermore, the reliability can be calculated using the estimated parameters, as shown in (17). Therefore, the precision of the reliability is relied on the parameter estimation.

$$ttf_i \sim \text{Weibull}(\alpha_i, \beta_i) \quad (16)$$

$$R_i = e^{-(84/\beta_i)^{\alpha_i}} \quad (17)$$

The Weibull's parameter a_i and β_i of the i[th] type are assumed to be Gamma distribution with parameter *a*, *b*, *c* and *d*.

$$\beta_i \sim \text{Gamma}(a, b) \quad (18)$$

$$\eta_i \sim \text{Gamma}(c, d) \quad (19)$$

As a_i and β_i of each type interfere with each other through *a*, *b*, *c* and *d*, the prior and initial value of parameters *a* to *d* will influence the running of the algorithm. In the following, the same convention in the book is adopted as uniform distribution:

$$a, b, c, d \sim \text{Uniform}(0.001, 100) \quad (20)$$

The experiment is implemented with the right censoring by permitting successful test within the period of time. In this case, the time to fail is an unknown value that is greater than the upper bound. From the original data table provided by the book, the number of failed products for each type during the experimental period is counted. For brevity, the original data table is not shown. From the above discussion, it is known that the more uncensored products, the higher

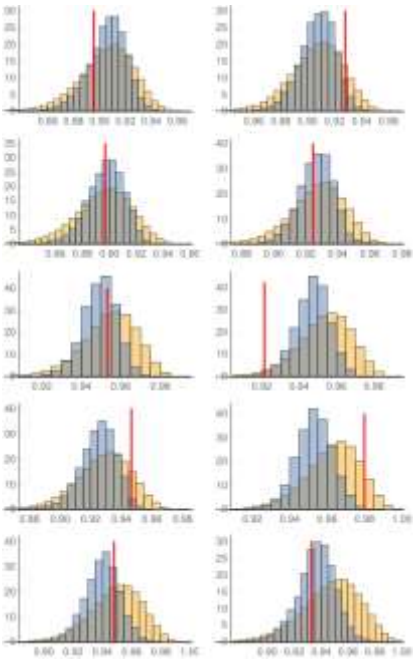

Fig. 19. The reliability histograms of 10 types (left to right and top to bottom), blue: the proposed method, orange: Gibbs sampler (JAGS), red line: true value of the reliability.

precision for parameter estimation as well as the reliability.

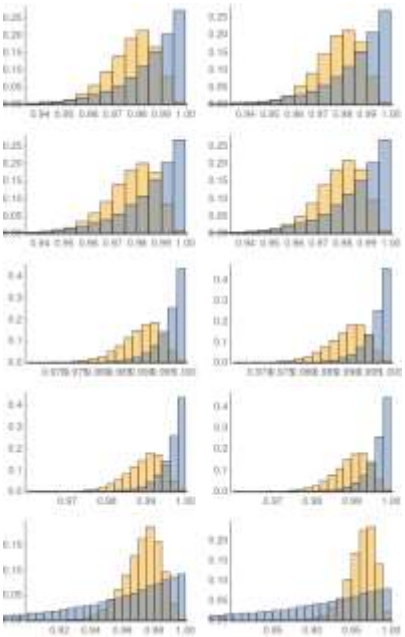

Fig. 20. Histogram of reliability of the proposed method (yellow) and the JAGS(blue) for θ_1 to θ_{10}.

As shown in Table I, the uncensored products for the 8th, 9th, and 10th types may be too few to apply Gibbs sampler.

TABLE I
Number of uncensored product

i	1	2	3	4	5	6	7	8	9	10
Mi	45	45	40	30	13	9	12	2	2	0

As shown in Table II, for quantitatively evaluating the methods, the true values of parameters and reliability of each type are provided by the book in Table II [43]. One notable feature is that the shape of the 8th type product is much higher than other values, resulting in the highest reliability prominently.

TABLE II
TRUE VALUE OF THE PARAMETERS

Product Name	Shape	Scale	Reliability
Gen 1	1.05271	681.011	0.895419
Gen 2	1.30417	610.3	0.92747
Gen 7	1.38997	674.867	0.946269
Gen 8	**2.155**	510.555	**0.979744**
Gen 9	1.47402	610.915	0.947733
Gen 10	1.2842	667.784	0.932594

As shown in Fig. 19, the blue histogram illustrates the computed reliability with (17) using the proposed method and the orange histogram is the result using JAGS, whereas the red

lines indicate the true value of these reliability. Firstly, both methods generate similar histograms for product type 1 to type 4, yet for other types, the proposed produces lower reliability than that of JAGS. Listed in Table I, the 1st to 4th type have significant more uncensored items than others, thus according to previous discussion, their parameter estimation should also be more precise. As shown in Table II, the true reliability of these types is substantially lower than others, therefore possibly lowers down the overall reliability. Secondly, the histograms using JAGS cover the parameter's true value, yet the 95% credible interval of the proposed method for 8th type does not contain the true value. After examining the true values in Table II, it is supposed that the shape and reliability of the 8th type is unusually higher than others, which may imply an uncommon type in the family. The level of censoring may be another potential contributing factor, since the type has only two uncensored data, high uncertainty is presence in parameter estimation.

E. Hierarchical Binomial Model

It has been shown that the reliability of a family of product can be studied using the hierarchical Binomial model [43]. The family encompasses 9 generations of product, and the 10th generation has not been tested: its reliability is predicted based on the other generations. As shown in Table IV provided by the book, for the i^{th} generation, N_i items are randomly picked and tested, and found Y_i success.

TABLE IV
TESTED AND PASSED NUMBER FOR EACH TYPE

i	1	2	3	4	5	6	7	8	9
Y_i	59	59	59	59	299	299	299	299	12
N_i	59	59	59	59	299	299	299	299	12

Assuming independent fault among products within the same generation, the number of successes follow a binomial distribution:

$$y_i \sim Binomial(n_i, \theta_i) \tag{21}$$

As shown in (22), if the prior of reliability θ_i of the of the i^{th} generation is a uniform distribution or Beta (1,1), then according to the conjugate between prior and likelihood, the posterior reliability is also a beta distribution therefor can be applied with the Gibbs sampler. By contrast, based on Metropolis, the proposed method is not subjected to the above constraint and affordable to any prior continuous density. Nevertheless, same prior is adopted for both the propose method and JAGS to compare the sampling results subsequently.

$$\theta_i \sim Beta(\alpha + 1, \beta + 1) \tag{22}$$

The convention in the published research is adopted here by assuming both α and β follow Gamma distribution (23) with hyper-parameter 1 and 1 and initializing the Markov chains for both parameters with 1 [43]. The prior and initial value for these two parameters is configurable in the experiment, which may affect sampling result. Therefore, to compare with the JAGS's result in the published research, the experiment is carried out using the same configuration.

$$\alpha, \beta \sim \text{Gamma}(1,1) \tag{23}$$

Traditionally, the Gibbs sampler implemented in BUGS tools such as JAGS is the major computational facility for hierarchical models. Nonetheless, this research applies CHMC to the same model to verify the feasibility of the proposed method, to comprehend the pros and cons of both methods, and to fully investigate the underlying distribution by incorporate multiple source of information.

What stands out in the Fig. 20 is the difference in shape between the proposed (yellow) and the Gibbs sampling with JAGS (blue). The histograms of the former are concentrated uni-modals, however the histograms of the latter increase monotonously, which may be arguably inconsistent with the distribution Beta(α+1, β+1) for positive α and β. Considering that both convergences are diagnosed with similar sample size, the disparity may be more likely caused by their fundamental difference: the proposed method explores the multidimensional space simultaneously, whereas the Gibbs sampler accesses each coordinate individually. It is reasonable that an adequately implemented multidimensional sampler is advantageous for better sampling precision.

TABLE V
RELIABILITY: THE PROPOSED

	Mean	Sd.	2.5%	50%	97.5%
θ_1	0.977736	0.00999381	0.954865	0.979083	0.993165
θ_5	0.989736	0.00479928	0.978151	0.990464	0.996855
θ_9	0.972567	0.0117747	**0.945639**	0.974276	0.990805
θ_{10}	0.96302	0.0178858	**0.920321**	0.965921	0.989597

As shown in Table V and Table VI, for the 1st to 8th products, the mean of the reliability using the proposed method is lower than that of JAGS. The reliability of θ_{10} for the future type shows the most dissimilarity. Its 95% credible interval is [92.0%,99.0%] for our method, and [60.9%,99,7%] for Gibbs sampler (JAGS). The wider credible interval of the JAGS may be attributed to its higher standard error. Regarding all products pass the test, the lower bound 60.9% of the Gibbs sampling may be too conservative.

TABLE VI
RELIABILITY: THE GIBBS SAMPLING (JAGS)

	Mean	Sd.	2.5%	50%	97.5%
θ_1	0.984933	0.0146366	0.946069	0.989242	0.999547
θ_5	0.996555	0.00342154	0.987525	0.997562	0.999904
θ_9	0.952628	0.0454001	**0.830467**	0.966313	0.998639
θ_{10}	0.889943	0.104913	**0.609385**	0.92128	0.996834

F. The proportional hazards (Cox) model

The model decomposes the i^{th} individual specific hazard function $h_i(t)$ into two terms:

$$h_i(t) = h_0(t, \lambda) e^{x_i^T \beta} \tag{24}$$

With the help of the proposed algorithm, it is not necessary

to depend on the partial likelihood at present, as the sampling on the full likelihood with censoring is trivial. Comparing to partial likelihood, it is a computation-intensive approach, yet the benefit will be clear for complex modeling and analysis, such as more precise RUL prediction.

To model censoring, firstly, we can link the cumulative distribution function with the survival function and cumulative hazard function by:

$$\mathrm{F}(\mathrm{t_i}, \lambda) = 1 - \mathrm{S}(\mathrm{t_i}, \lambda) = 1 - e^{-H(t_i, \lambda)} \quad (25)$$

Assuming the i^{th} observation is the triple (t_i, d_i, x_i), its likelihood of the i^{th} observation given the covariates can be defined in (26), which is identical to a handbook effectively [443].

$$\mathrm{F}'(t_i, \lambda)^{d_i}(1 - \mathrm{F}(t_i, \lambda))^{1-d_i} \quad (26)$$

The t_i is the event or censoring time, d_i is censoring indicator, and x_i is the covariate (vector). If an event occurs at t_i, then d_i equals 1. On the other hand, if nothing happens eventually, then d_i will be 0. Multiplying likelihood of all individual observations results in a total likelihood with the parameter λ and β.

As researches based on real data may have limitation [454], a novel data generating method is adopted here by sampling data from the "likelihood" using CHMC. As shown in Fig. 21, adopting the convention of a textbook [7], the data (x and t) play the role of the distribution's parameter to be sampled, and the values of λ and β are the observed "data". To highlight multivariate characteristics, both x and β are defined as two dimensional vectors. Without loss of generality, h_0 is based on exponential distribution with parameter λ.

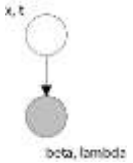

Fig. 21. Graphical model for data generating.

One major advantage as well as controversy of the Bayesian methodology is the inclusion of prior knowledge. When generating synthetic data, we can introduce a prior probability over data x (and t, which is positive), then we obtain the posterior distribution given β and λ. In practice, improper priors, such as a uniform distribution over an infinite domain, may be appropriate candidate. For this data generating process, however, improper prior may raise convergence problem.

Fig. 22 shows that the samples of posterior over x are distributed on a thin hyper-plane (or surface) when such improper uniform prior is adopted. Under the condition of β_1=1, β_2=5, and λ=1, positive t and unrestricted x are sampled using 5 Markov chains. The scatter plot is shown in two different angles of view in Fig. 23. The left figure shows the uneven distribution, which suggests inadequate convergence. On the right, we can visualize a thin hyper-plane, which may cause problems for simulation and sampling algorithm. Although convergence is generally indispensable for sampling algorithm, the requirement of strict convergence may be relaxed for the sampling of data. In any case, we are interested in parameter estimation instead of data simulation.

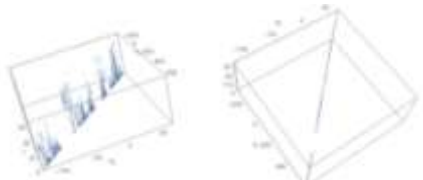

Fig. 22. Scatter plots of the ideal data with two angles of view.

The major difference from the Cox's is the coincidence of the parameter β of covariates and parameter λ of the base line probability. It is demonstrated with the simplest multivariate model, as models based on higher dimensional covariates or more complex probabilities are similar. As mentioned before, unrestricted x can make both the data and parameter spreading on hyper surfaces, which often impede convergence.

As before, we simulate 5000 samples for each of 5 chains, and randomly choose 20 items from the total of 25000 samples. The right censoring threshold of t_i is set to 1, which equals to the mean value of the exponential distribution due to the true value of λ, resulting in triples (t_i, d_i, xc_i) as data. A graphical model representing the process by which data are generated is shown in Fig. 23, in which the xc_i denotes the outcome of the censoring process and N is the number of items (i.e., 20).

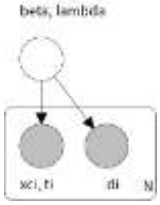

Fig. 23 Graphical model of parameter estimation.

Sampling the parameters conditioned on the above mentioned data, the scatter plot in two different angle of view is shown in Fig. 24, which hints convergence. The scatter plot shown in the left-hand plot in Fig. 24 presents a strong non-normality, thus the inference with normal assumption may

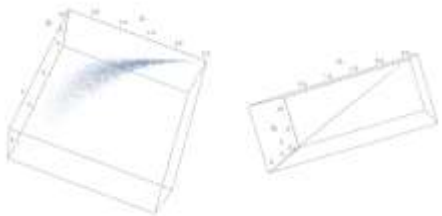

Fig. 24. Scatter plots of the parameters for the ideal data.

thereby introduce systematic error. The right-hand plot shows that the parameter samples lie on a thin plane (surface). The probability distribution on multidimensional surface can cause great trouble for MCMC (and HMC) algorithms, since it is very easy to move into the low probability region. Although RMHMC may be free from this problem, ~~it can~~ only the sampling on the first principal components can be guaranteed ~~for the sampling on the first principal components~~. Nevertheless, due to several enhancements adopted, the proposed method is usually suitable to this kind of situation.

Fig. 25 reveals the sample histogram, location and interval of the parameter β_1. The blue dashed and solid lines indicate the mean and credible interval using the proposed method, and similarly, the black dashed and solid lines correspond to the Cox's PHM algorithm. The red dashed line locates the true value. Compared with the MLE location (black dashed line) obtained using PHM algorithm, the samples' mean of our method (blue dashed line) is nearer to the true value (red dashed line). Furthermore, the credible interval of the proposed method is also slightly narrower than the confidence interval of PHM. The plausible lower precision of PHM may be attributed to the part of censored data, which have been treated ordinarily.

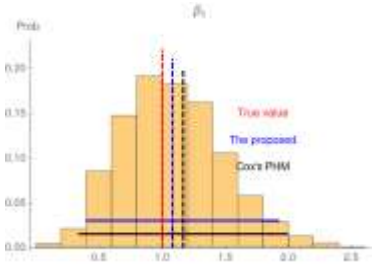

Fig. 25. Inference of β_1 for the ideal data, solid lines delimit the intervals, and dashed lines indicate the locations of the parameter.

The inference result for β_2 is shown in Fig. 26 and it has a similar form to β_1. Both methods identify the parameters' location and range, and the proposed method obtains a slightly more precise estimation. This follows from the fact that, if implemented properly, the sampling approach of Bayesian may be comparable or better than the classical frequentist method.

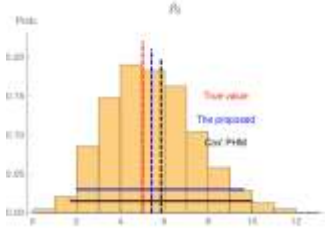

Fig. 26. Inference for β_2 for the ideal data.

Fig. 27 shows only the result of our method for the parameter λ of the underlying hazard, because PHM does not model this parameter. As a result, the PHM is termed as a semi-parametric method for lacking the capability. One of the benefits of sampling parameters for multivariate probability models is the exemption from traditional sophisticated and complicated algorithms. Therefore, once the convergence of the algorithm is assured, more insight into the model can be gained using the direct sampling approach.

The sample mean of λ (1.5) is not far from the true parameter value (which is 1), considering the small data size of 20. In

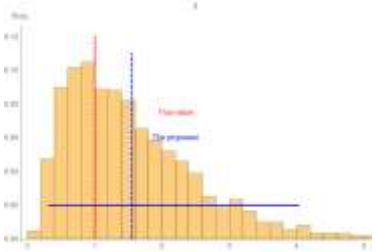

Fig. 27. Inference for λ for the ideal data.

addition, the 95% credible interval covers the true parameter value. This concludes that the sampling approach is capable to estimate whole parameters based on the ideal data. The estimated covariates may be more precise than that of the PHM algorithm, and the parameter estimation for the base-line hazard satisfies the inference requirement, that is, the credible interval covers the true parameter value with high probability.

Based on the estimated β and λ, we begin by considering the problem of predicting life time t for a particular variable x. If the prior of data is not overly restricted, ideal data will be drawn from the distribution in theory. The prediction based on the parameter estimated according to data will be precise. Though

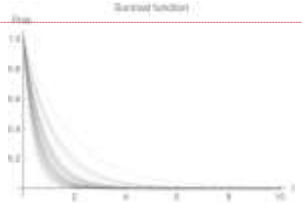

Fig. 28. Survival probability function ~~over~~ for a specific observed x based on 50 samples of the parameter.

the sampled parameters, it is routine to arrive at more sensible conclusions, such as survival time. Substituting a new x value and the parameters (β, λ) into survival function S(t), the curve of function with respect to t can be obtained. Since the predictive distribution over t, given a new input x, is easily evaluated, the statistical inference of survival time can be carried out. Fig. 28 illustrates survival curves according to 50 randomly selected sampled parameters, setting x_1 and x_2 to their median value (129.5 and -25.8). It is often advisable to anticipate the median survival time, which is the smallest time t such that S(t)<0.5 [4~~6~~5]. As the true value of λ is set to 1, the median time of the corresponding exponential distribution will

带格式的：字体: (默认) +标题 CS (Times New Roman), 复杂文种字体: +标题 CS (Times New Roman)

be 0.693147. Consider the above mentioned observation for x (129.5 and -25.8), the mean of the median survival time is found to be 0.40, which is somewhat near to the median of the exponential (i.e., 0.693147).

The RUL can be predicted in a similar way. For example, the probability of survival is 0.1, or 10% from the beginning to t_{MUL}. The histogram of t_{RUL} is illustrated in Fig. 29, with mean of 1.55 and 95% credible interval (0.59,2.96). The predicted RUL only depends on covariates (x_1 and x_2), thus the individual's properties determine inference of RUL, such as maximum, minimum, or interval.

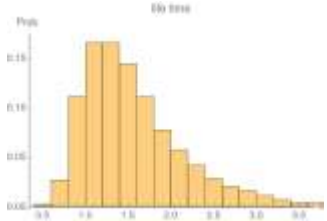

Fig. 29. The sample histogram of RUL for a specific observation.

If, on the contrary, the prior is too strict, the simulated data will misrepresent the ideal data for the corresponding value of the parameter, and the subsequent parameter estimation will be imprecise accordingly. Consider the prior defined over the interval of ±1, as shown in Fig. 30, the simulated data reside in a spatial region instead of a hyper-plane. Compare with the case of improper prior, the convergence is improved. The forthcoming problem due to the strict prior, however, is the disparity between simulated and ideal data (or parameter).

Fig. 30. Scatter plots of the non-ideal data with two angles of view.

Fig. 31 demonstrates the scatter plot of parameter based on the non-ideal data. Unlike the case of ideal data shown in Fig. 24, the points distributed across a 3-dimensional region. The type is obviously easier to sample, however, the sampled parameter will be very different from the true one.

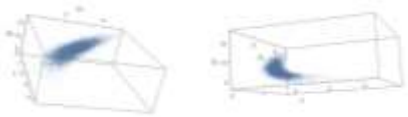

Fig. 31. Scatter plots of parameter for non-ideal data with two angles of views.

We can see in Fig. 32 the histogram of the sampled parameter based on the data which have prior distribution of uniform between -1 and 1. We know from Fig. 22 that the range of x

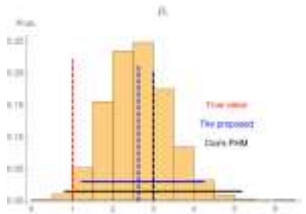

Fig. 32. Inference of β_1 for the non-ideal data.

should be at least several hundreds, thus the uniform prior of ±1 is too constricting. As shown in Fig. 32, both the sample mean of the proposed method and the MLE of the PHM algorithm are located away from the true value. Although the inference result of the proposed method is not precise, it is slightly better than that of the PHM.

In the same way, the relevant result for β_2 is shown in Fig. 33, which also reveals the slightly higher precision of the proposed method, although both methods are imprecise. The wider interval of the PHM may indicate both the strong non-normality and the large inference error. It means that the experiment based on real data and evaluated against PHM may derive wrong conclusion. Therefore, the parameters' true value, instead of the inference result of PHM, is used as the benchmark for comparing performance.

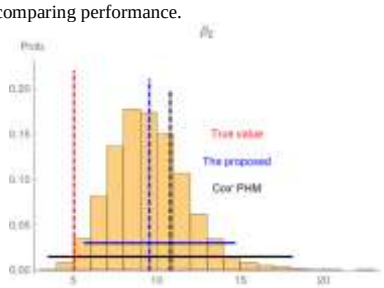

Fig. 33. Inference of β_2 for the non-ideal data.

The histogram of Fig. 34 shows that the 95% credible interval covers the true value, yet compared with Fig. 27, the sample mean is doubtlessly smaller than the true value. The phenomenon is opposite to β, which has larger sample mean than its true value in Fig. 32 and Fig. 33. As both results shown in Fig. 32 and Fig. 33 are inaccurate, it seems that the non-ideal data have a lesser effect on λ. The phenomenon might be explained according to (24): there is a directly coupling between x and β, therefore if x is inclined smaller then β may be larger to compensate the decreased x; the two effects can be partially canceled, leaving λ mainly unaffected.

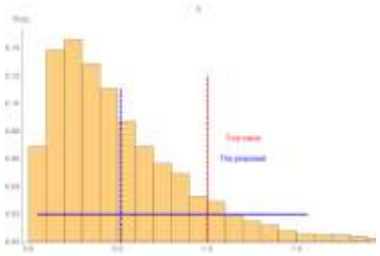

Fig. 34. Inference of λ for the non-ideal data.

As shown in Fig. 35, 50 samples are picked randomly from the sampling result and used to plot survival function. It is easy to expect that the estimated median survival time, which is 0.12 actually, is quite small. The value is far less than the median value of the exponential distribution, indicating a significant underestimate. For the unconstrained, the computed $x \cdot \beta$ is either small positive or small negative. Yet for the constrained prior, the sample median of x_1 and x_2 is evaluated to small positive. As both β_1 and β_2 are positive, treating median as observation, the computed $x \cdot \beta$ is also positive and slightly larger than the unconstrained case. As a result, it decays faster for the survival function based on the constrained data.

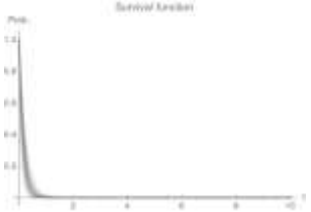

Fig. 35. Survival probability function based on the non-ideal data.

According to the experiment, it is conceivable that the predicted survival may be influenced by the prior of data, thus to obtain useful inference result, it may be necessary to consider carefully the influence of data preparation.

V. Conclusion

This study has shown the inference of the parameter as well as the reliability and lifespan of reliability models using the modified Hamiltonian Monte Carlo sampler. The algorithm tunes all run-time parameter automatically, proves to be stable and effective, and only depends on the potential as well as its first and second derivative. To exempt from coding errors, we exclusively rely on a computer algebra system for computing the derivatives. The approach is rather general and can cope with complex models encountered in the reliability engineering. One major benefit of the approach of the automatic derivative is the flexibility and power. It may supersede the rule-based Gibbs samplers. The only necessity for the inference is to construct a potential function, which will considerably simplify the related researches. Compared with existing approaches, the proposed sampler is more accurate on most principal components. The improvement is verified with carefully designed quantitative evaluation. Experiments show that the method is suitable for multivariate, correlated, non-normal, and censored reliability models.

VI. Acknowledgements

This work was supported by the National Natural Science Fund of China [Grant Nos. 61373063, 61872188 and 62172225]; and Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University [Grant No. IIC1701].

References

[1] T. T. Thanh and B. Radim, "Improved new modified Weibull distribution: A Bayes study using Hamiltonian Monte Carlo simulation," *Journal of Risk and Reliability*, Part O, vol. 234, no. 3, pp. 496-511, Jun. 2020.

[2] M. R. Lyu, *Handbook of Software Reliability Engineering*, New York, NY, USA: McGraw-Hill, 1996, pp. 71–89.

[3] H. Pham, *Handbook of Reliability Engineering*, London, UK: Springer-Verlag, 2003, pp. 157.

[4] H. Langseth and L. Portinale, "Bayesian Networks in Reliability," *Reliability Engineering and System Safety*, vol. 92, no. 1, pp. 92–108, Jan. 2005.

[5] L. Bouillaut, O. Franois, Y. Putallaz, and C. Cieux, "A Hybrid Approach for the Evaluation of Rail Monitoring and Maintenance Strategies for the Grand Paris Express New Metro," *International Journal of Performability Engineering*, vol. 16, no.11, pp. 1685-1697, Nov. 2020.

[6] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, Cambridge, MA, USA: MIT Press, 2012, pp. 307.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York, NY, USA: Springer-Verlag, 2006, pp. 374.

[8] W. Zhang, X. Wang, D. Cabrera, and Y. Bai, "Product Quality Reliability Analysis based on ROUGH Bayesian Network," *International Journal of Performability Engineering*, vol. 16, no. 1, pp. 37-47, Jan. 2020.

[9] D. Nash and M. Hannah, "Using Monte-Carlo simulations and Bayesian Networks to quantify and demonstrate the impact of fertiliser best management practices," *Environmental Modelling and Software*, vol. 26, no. 9, pp. 1079–1088, Sep. 2011.

[10] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter, *The BUGS Book, A Practical Introduction to Bayesian Analysis*, FBoca Raton, FL, USA: CRC Press, 2012, pp. 253.

[11] Y. Liu and A. I. Abeyratne, *Practical Applications of Bayesian Reliability*, Hoboken, NJ, USA: Wiley, 2019, pp. 221–238.

[12] X. Si, W. Wang, C. Hu, and D. Zhou, "Remaining Useful Life Estimation–A Review on the Statistical Data Driven Approaches," *European Journal of Operational Research*, vol. 213, no. 1, pp. 1–14, Aug. 2011.

[13] B. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society: Series B*, vol. 34, no. 2, pp. 187–202, Jan. 1972.

[14] M. Aminisharifabad, Q. Yang and X. Wu, "A Deep Learning-Based Reliability Model for Complex Survival Data," *IEEE Trans. Reliability*, vol. 70, no. 1, pp. 73–81, Mar. 2021.

[15] J. Salvatier, T. V. Wiecki, C. Fonnesbeck and C. J. Fonnesbeck, "Probabilistic programming in Python using PyMC3," *PeerJ Computer Science 2: e55*, Oct. 2016.

[16] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, and A. Riddell, "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, vol. 76, no. 1, pp. 1–32, Jan. 2017.

[17] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter, "WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility," *Statistics and Computing*, vol. 10, pp. 325–337, Oct. 2000.

[18] A. Patil, D. Huard, C. J. Fonnesbeck, "PyMC: Bayesian Stochastic Modelling in Python," *Journal of Statistical Software*, vol. 35, no. 4, pp. 1–81, July. 2010.

[19] J. K. Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, 2nd edition*, Salt lake City, UT, USA: Academic Press, 2014, pp. 1–746.

[20] C. Davidson-Pilon, *Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference*, Boston, MA, USA: Addison-Wesley, 2015, pp. 1–256.
[21] S. Wolfram, *The Mathematica Book, 4th edition*, London, UK: Cambridge University Press, 1999, pp. 1–1496.
[22] P. Cook and L. D. Broemeling, "Bayesian Statistics Using Mathematica," *The American Statistician*, vol. 49, no. 1, pp. 70–76, Jan. 1996.
[23] A. Xu and Y. Tang, "Bayesian Analysis of Pareto Reliability with Dependent Masked Data," *IEEE Trans. Reliability*, vol. 58, no. 4, pp. 583–588, Jan. 2009.
[24] A. Justel and D. Peña, "Gibbs Sampling Will Fail in Outlier Problems with Strong Masking," *Journal of Computational and Graphical Statistics*, vol. 5, no. 2, pp. 176–189, Jun. 1996.
[25] N. Al-Khairullah, T. Al-Baldawi, "Bayesian Computational Methods of the Logistic Regression Model," Journal of Physics: Conference Series; Bristol, vol. 1804, no. 1, Feb. 2021.
[26] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Physics Letters B*, vol. 195, no. 2, pp. 216–222, Sep. 1987.
[27] S. Prokhorenko, K. Kalke, Y. Nahas, and L. Bellaiche, "Large scale hybrid Monte Carlo simulations for structure and property prediction," *Npj Computational Materials*, vol. 4, no. 1, pp. 1–7, Dec. 2018.
[28] S. Livingstone, M. F. Faulkner, and G. Roberts, "Kinetic energy choice in Hamiltonian/hybrid Monte Carlo," *Biometrika*, vol. 106, no. 2, pp. 303–319, Jun. 2019.
[29] A. Beskos, N. Pillai, G. Roberts, J. Sanz-Serna, and A. Stuart, "Optimal tuning of the Hybrid Monte-Carlo Algorithm," *Bernoulli*, vol. 19, no. 5A, pp. 1501–1534, Nov. 2013.
[30] R. Gupta, G. W. Kilcup, and S. R. Sharpe, "Tuning the Hybrid Monte Carlo Algorithm," *Physical Review D Particles & Fields*, vol. 234, pp. 179–187, Jan. 2019.
[31] D. Luengo, L. Martino, M. Bugallo, V. Elvira, S. Särkkä, "A survey of Monte Carlo methods for parameter estimation," EURASIP Journal on Advances in Signal Processing, vol. no. 1, Dec. 2020.
[32] M. D. Hoffman and A. Gelman, "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, vol. 15, no. 47, pp. 1593–1623, 2014.
[33] M. Girolami and B. Calderhead, "Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods," *Journal of the Royal Statistical Society: Series B*, vol. 73, no. 2, pp. 123–214, Mar. 2011.
[34] W. Krauth, *Statistical Mechanics: Algorithms and Computations*, Oxford, UK: Oxford University Press, 2006, pp. 15–22.
[35] D. Mackay, Information Theory, *Inference and Learning Algorithms*, Cambridge, UK: Cambridge University Press, 2003, pp. 388.
[36] A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*, Boca Raton, FL, USA: CRC Press, 2011, pp. 125.
[37] J. S. Dickstein, M. Mudigonda, and M. DeWeese, "Hamiltonian Monte Carlo Without Detailed Balance," in *Int. Conf. Machine Learning*, BJ, China, 2014.
[38] M. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*, Oxford, UK: Oxford University Press, 2010, pp. 295–296.
[39] M. P. Kaminskiy and V. V. Krivtsov, "A Simple Procedure for Bayesian Estimation of the Weibull Distribution," *IEEE Trans. Reliability*, vol. 54, no. 4, pp. 612–616, Dec. 2005.
[40] M. Plummer, "JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling" in *3rd Int. Workshop on Distributed Statistical Computing*, Vienna, Austria, 2003.
[41] M. Modarres, M. Kaminskiy, and V. Krivtsov, *Reliability Engineering and Risk Analysis: A Practical Guide, Second Edition*, Boca Raton, FL, USA: CRC Press, 2009, pp. 106–108.
[42] M. S. Hamada, A. Wilson, C. S. Reese, and H. Martz, *Bayesian Reliability*, Secaucus, NJ, USA: Springer, 2008, pp. 13–14.
[43] J.P. Klein, H. C. V. Houwelingen, J. G. Ibrahim, and T. H. Scheike, *Handbook of Survival Analysis*, Boca Raton, FL, USA: CRC Press, 2013, pp. 9–10.
[44] D. Han and T. Bai, "Parameter Estimation Using EM Algorithm for Lifetimes from Step-Stress and Constant-Stress Accelerated Life Tests with Interval Monitoring," *IEEE Trans. Reliability*, to be published.
[45] D. Moore, *Applied Survival Analysis Using R*, Cham, CH: Springer, 2015, pp. 11–17.