Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

An Adaptive Interacting Wang–Landau Algorithm for Automatic Density Exploration

Luke BORNN, Pierre E. JACOB, Pierre DEL MORAL, and Arnaud DOUCET

While statisticians are well-accustomed to performing exploratory analysis in the modeling stage of an analysis, the notion of conducting preliminary general-purpose exploratory analysis in the Monte Carlo stage (or more generally, the model-fitting stage) of an analysis is an area that we feel deserves much further attention. Toward this aim, this article proposes a general-purpose algorithm for automatic density exploration. The proposed exploration algorithm combines and expands upon components from various adaptive Markov chain Monte Carlo methods, with the Wang–Landau algorithm at its heart. Additionally, the algorithm is run on interacting parallel chains—a feature that both decreases computational cost as well as stabilizes the algorithm, improving its ability to explore the density. Performance of this new parallel adaptive Wang–Landau algorithm is studied in several applications. Through a Bayesian variable selection example, we demonstrate the convergence gains obtained with interacting chains. The ability of the algorithm's adaptive proposal to induce mode-jumping is illustrated through a Bayesian mixture modeling application. Last, through a two-dimensional Ising model, the authors demonstrate the ability of the algorithm to overcome the high correlations encountered in spatial models. Supplemental materials are available online.

Key Words: Adaptive binning; Adaptive Markov chain Monte Carlo; Parallel chains; Wang-Landau algorithm.

1. INTRODUCTION

As improvements in technology introduce measuring devices capable of capturing ever more complex real-world phenomena, the accompanying models used to understand such phenomena grow accordingly.While linear models under the assumption of Gaussian noise were the hallmark of early twentieth century statistics, the past several decades have seen an explosion in statistical models that produce complex and high-dimensional density functions for which simple, analytical integration is impossible. This growth was largely fueled by renewed interest in Bayesian statistics accompanying the Markov chain Monte

Luke Bornn is Assistant Professor, Department of Statistics, Harvard University, Cambridge, MA 02138, USA (E-mail: *bornn@stat.harvard.edu*). Pierre E. Jacob is Ph.D. student, CEREMADE, Université Paris-Dauphine, France (E-mail: *pierre.jacob@ensae.fr*). Pierre Del Moral is Research Director, INRIA Bordeaux Sud-Ouest and University of Bordeaux, France (E-mail: *Pierre.Del-Moral@inria.fr*). Arnaud Doucet is Professor, Department of Statistics, Oxford University, Oxford OX1 3TG, UK (E-mail: *doucet@stats.ox.ac.uk*).

Journal of Computational and Graphical Statistics, Volume 22, Number 3, Pages 749–773
DOI: 10.1080/10618600.2012.723569

Carlo (MCMC) revolution in the 1990s. With the computational power to explore the posterior distributions arising from Bayesian models, MCMC allowed practitioners to build models of increasing size and nonlinearity.

As a core component of many of the MCMC algorithms discussed later, we briefly recall the Metropolis–Hastings (M–H) algorithm. With the goal of sampling from a density π, the algorithm generates a Markov chain $(x_t)_{t=0}^T$ with invariant distribution π. From a current state x_t, a new state x' is sampled using a proposal density $q_\eta(x_t, x')$ parameterized by η. The proposed state x' is accepted as the new state x_{t+1} of the chain with probability

$$\min\left(1, \frac{\pi(x')q_\eta(x', x_t)}{\pi(x_t)q_\eta(x_t, x')}\right)$$

and if it is rejected, the new state x_{t+1} is set to the previous state x_t. From this simple algorithmic description, it is straightforward to see that if x_t is in a local mode and the proposal density $q_\eta(x_t, x')$ has not been carefully chosen to propose samples from distant regions, the chain will become stuck in the current mode. This is due to the rejection of samples proposed outside the mode, underscoring the importance of ensuring $q_\eta(x_t, x')$ is intelligently designed.

Though standard MCMC algorithms such as the M–H algorithm and the Gibbs sampler have been studied thoroughly and the convergence to the target distribution is ensured under weak assumptions, many applications introduce distributions that cannot be sampled easily by these algorithms. Multiple reasons can lead to failure in practice, even if long-run convergence is guaranteed; the question then becomes whether or not the required number of iterations to accurately approximate the density is reasonable given the currently available computational power. Among these reasons, let us cite a few that will be illustrated in later examples: the probability density function might be highly multimodal, in which case the chain can get stuck in local modes. Alternatively or additionally, it might be defined on a high-dimensional state space with strong correlations between the components, in which case the proposal distributions (and in particular their large covariance matrices) are very difficult to tune manually. These issues lead to error and bias in the resulting inference, and may be detected through convergence monitoring techniques (see, e.g., Robert and Casella 2004). However, even when convergence is monitored, it is possible that entire modes of the posterior are missed. To address these issues, we turn to a burgeoning class of Monte Carlo methods, which we refer to as "exploratory algorithms."

In the following section, we discuss the traits that allow exploratory MCMC algorithms to perform inference in multimodal, high-dimensional distributions, connecting these traits to existing exploratory algorithms in the process. In Section 3, we detail one of these, the Wang–Landau algorithm, and propose several novel improvements that make it more adaptive, hence easier to use, and also improve convergence. Section 4 applies the proposed algorithm to variable selection, mixture modeling, and spatial imaging, before Section 5 concludes.

2. EXPLORATORY ALGORITHMS

As emphasized by Schmidler (2011), there are two distinct goals of existing adaptive algorithms. First, algorithms which adapt the proposal according to past samples are largely exploitative, in that they improve sampling of features already seen. However, modes or

features not yet seen by the sampler might be quite different from the previously explored region, and as such adaptation might prevent adequate exploration of alternate regions of the state space. As an attempted solution to this problem, Craiu, Rosenthal, and Yang (2009) suggested adapting regionally, with parallel chains used to perform the adaptation. Second, there exists a set of adaptive algorithms whose goal is to adapt in such a way as to encourage density exploration. These include, for instance, the equi-energy sampler (Kou, Zhou, and Wong 2006), parallel tempering (Swendsen and Wang 1986; Geyer 1991), and the Wang–Landau (Wang and Landau 2001a,b; Liang 2005; Atchadé and Liu 2010) algorithms among others. The algorithm developed here fits into the latter suite of tools, whose goal is to explore the target density, particularly distant and potentially unknown modes.

Although the aforementioned algorithms have proven efficient for specific challenging inference problems, they are not necessarily designed to be generic, and it is often difficult and time-consuming for practitioners to learn and code these algorithms merely to test a model. As such, while statisticians are accustomed to exploratory data analysis, we believe that there is room for generic exploratory Monte Carlo algorithms to learn the basic features of the distribution or model of interest, particularly the locations of modes and regions of high correlation. These generic algorithms would ideally be able to deal with discrete and continuous state spaces and any associated distribution of interest, and would require as few parameters to tune as possible, such that users can use them before embarking on time-consuming, tailor-made solutions designed to estimate expectations with high precision. In this way, one may perform inference and compare between a wide range of models without building custom-purpose Monte Carlo methods for each.

We first describe various ideas that have been used to explore highly multimodal densities, and then describe recent works aimed at automatically tuning algorithmic parameters of MCMC methods, making them able to handle various situations without requiring much case-specific work from the user.

2.1 ABILITY TO CROSS LOW-DENSITY BARRIERS

The fundamental problem of density exploration is settling into local modes, with an inability to cross low-density regions to find alternative modes. For densities π that are highly multimodal, or "rugged," one can employ tempering strategies, sampling instead from a distribution proportional to $\pi^{1/\tau}$ with temperature $\tau > 1$. Through tempering, the peaks and valleys of π are smoothed, allowing easier exploration. This is the fundamental idea behind parallel tempering, which employs multiple chains at different temperatures; samples are then swapped between chains, using highly tempered chains to assist in the exploration of the untempered chain (Geyer 1991). Marinari and Parisi (1992) subsequently proposed simulated tempering that dynamically moves a single chain up or down the temperature ladder. One may also fit tempering within a sequential Monte Carlo (SMC) approach, whereby samples are first obtained from a highly tempered distribution; these samples are transitioned through a sequence of distributions converging to π using importance sampling and moves with a Markov kernel (Neal 2001; Del Moral, Doucet, and Jasra 2006). However, using tempering strategies with complex densities, one must be careful of phase transitions where the density transforms considerably across a given temperature schedule.

A related class of algorithms works by partitioning the state space along the energy function $-\log \pi(x)$. The idea of slicing, or partitioning, along the energy function is the hallmark of several auxiliary variable sampling methods that iteratively sample $U \sim \mathcal{U}[0, \pi(X)]$ so that $X \sim \mathcal{U}\{X : \pi(X) \geq U\}$. This is the fundamental idea behind the Swendsen–Wang algorithm (Swendsen and Wang 1987; Edwards and Sokal 1988) and related algorithms (e.g., Besag and Green 1993; Higdon 1998; Neal 2003). The equi-energy sampler (Kou, Zhou, and Wong 2006; Baragatti, Grimaud, and Pommeret 2012), in contrast to the above auxiliary variable methods, begins by sampling from a highly tempered distribution; once convergence is reached, a new reduced-temperature chain is run with updates from a mixture of Metropolis moves and exchanges of the current state with the value of a previous chain in the same energy band. The process is continued until the temperature reaches 1 and the invariant distribution of the chain is the target of interest. As such, this algorithm works through a sequence of tempered distributions, using previous distributions to create intelligent mode-jumping moves along an equal-energy set.

In a similar vein, the Wang–Landau algorithm (Wang and Landau 2001a,b) also partitions the state space $\mathcal{X}$ along a reaction coordinate $\xi(x)$, typically the energy function: $\xi(x) = -\log \pi(x)$, resulting in a partition $(\mathcal{X}_i)_{i=1}^d$. The algorithm generates a time-inhomogeneous Markov chain that admits an invariant distribution $\tilde{\pi}_t$ at iteration t, instead of the target distribution π itself as, for example, in a standard M–H algorithm. The distribution $\tilde{\pi}_t$ is designed such that the generated chain equally visits the various regions $\mathcal{X}_i$ as $t \to \infty$. Because the Wang–Landau algorithm lies at the heart of our proposed algorithm, it is extensively described in Section 3.

It is worth discussing a similar, recently proposed algorithm that combines MCMC and free energy biasing (Chopin, Lelievre, and Stoltz 2012) and its SMC counterpart (Chopin and Jacob 2010). The central idea of the latter is to explore a sequence of distributions, successively biasing according to a reaction coordinate ξ in a similar manner. However, we have found the method to be largely dependent on selecting a well-chosen initial distribution π_0, as is usually the case with SMC methods for static inference. If the initial distribution is not chosen to be flatter than the target distribution, which is possibly the case since the regions of interest with respect to the target distribution are a priori unknown, the efficiency of the SMC methods relies mostly on the move steps, which are themselves M–H or Gibbs moves.

2.2 ADAPTIVE PROPOSAL MECHANISM

Concurrent with the increasing popularity of exploratory methods, the issue of adaptively fine-tuning MCMC algorithms has also seen considerable growth since the foundational article of Haario, Saksman, and Tamminen (2001), including a series of conferences dedicated solely to the problem (namely, Adap'ski 1 through 3 among others); see the reviews of Andrieu and Thoms (2008) and Atchadé et al. (2011) for more details. While the de facto standard has historically been hand-tuning of MCMC algorithms, this new work finds interest in automated tuning, resulting in a new class of methods called adaptive MCMC.

The majority of the existing literature focuses on creating intelligent proposal distributions for an MCMC sampler. The principal idea is to exploit past samples to induce better moves across the state space by matching moments of the proposal and past samples,

or by encouraging a particular acceptance rate of the sampler. The raison d'être of these algorithms is that tuning MCMC algorithms by hand is both time-consuming and prone to inaccuracies. By automating the selection of the algorithm's parameters, practitioners might save considerable time in their analyses. This feature is pivotal in an automated density exploration algorithm. Due to its exploratory nature, it is likely that the practitioner might not have complete knowledge of even the scale of the density's support; as a result, having a proposal distribution that adapts to the density at hand is a crucial step in the automation process.

One must be careful in selecting the type of adaptation mechanism employed to encourage exploration, rather than simply exploiting previously explored modes. For instance, tuning a proposal covariance to a previously visited mode might prevent the algorithm from reaching as yet unexplored modes in the direction of the current mode's minor axis. Additionally, when combined with a progressively biased distribution as in the Wang–Landau algorithm, it is desirable to have a proposal which first samples what it sees well, then later grows in step size to better explore the flattened (biased) distribution.

3. PROPOSED ALGORITHM

We now develop our proposed algorithm. After recalling the Wang–Landau algorithm, which constitutes the core of our method, we describe three improvements: an adaptive binning strategy to automate the difficult task of partitioning the state space, the use of interacting parallel chains to improve the convergence speed and use of computational resources, and finally the use of adaptive proposal distributions to encourage exploration as well as to reduce the number of algorithmic parameters. We detail at the end of the section how to use the output of the algorithm, which we term parallel adaptive Wang–Landau (PAWL), to answer the statistical problem at hand.

3.1 THE WANG–LANDAU ALGORITHM

As previously mentioned, the Wang–Landau algorithm generates a time-inhomogeneous Markov chain that admits a distribution $\tilde{\pi}_t$ as the invariant distribution at iteration t. The biased distribution $\tilde{\pi}_t$ targeted by the algorithm at iteration t is based on the target distribution π, and modified such that (a) the generated chain visits all the sets $(\mathcal{X}_i)_{i=1}^d$ equally, that is, the proportion of visits in each set is converging to d^{-1} when t goes to infinity and (b) the restriction of the modified distribution $\tilde{\pi}_t$ to each set $\mathcal{X}_i$ coincides with the restriction of the target distribution π to this set, up to a multiplicative constant. The modification (a) is crucial, as inducing uniform exploration of the sets is the biasing mechanism that improves exploration; in fact similar strategies are used in other fields, including combinatorial optimization (Wei, Erenrich, and Selman 2004). Ideally the biased distribution $\tilde{\pi}$ would not depend on t, and would be available analytically as

$$\tilde{\pi}(x) = \pi(x) \times \frac{1}{d} \sum_{i=1}^{d} \frac{\mathcal{I}_{\mathcal{X}_i}(x)}{\psi(i)}, \tag{1}$$

where $\psi(i) = \int_{\mathcal{X}_i} \pi(x)dx$ and $\mathcal{I}_{\mathcal{X}_i}(x)$ is equal to 1 if $x \in \mathcal{X}_i$ and 0 otherwise. Checking that using $\tilde{\pi}$ as the invariant distribution of an MCMC algorithm would validate points

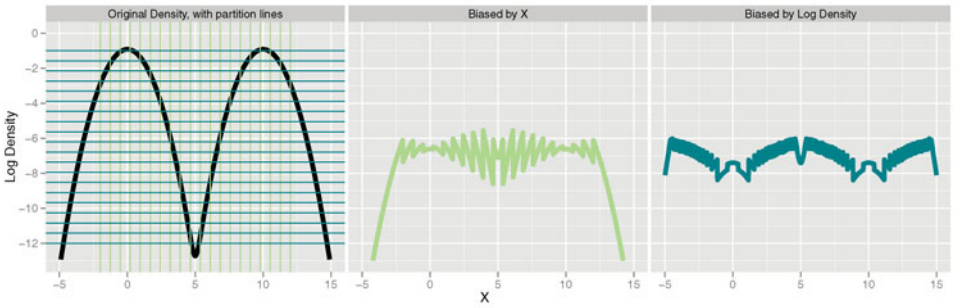

Figure 1. Probability density functions for a univariate distribution π and its biased version $\tilde{\pi}$ when partitioning the state space along the x-axis ($\xi(x) = x$, middle) and the log density ($\xi(x) = -\log \pi(x)$, right). The leftmost plot also shows the partitioning of the state space with $\xi(x)$; in both cases $d = 20$. The biasing is done such that the integral $\int_{\mathcal{X}_i} \tilde{\pi}(x)dx$ is the same for all $\mathcal{X}_i$ (areas under the curve for each set) and such that π and $\tilde{\pi}$ coincide on each set $\mathcal{X}_i$, up to a multiplicative constant. The online version of this figure is in color.

(a) and (b) is straightforward. Figure 1 illustrates a univariate target distribution π and its corresponding biased distribution $\tilde{\pi}$ under two different partitions of the state space.

In practical situations, however, the integrals $(\psi(i))_{i=1}^d$ are not available, hence we wish to plug estimates $(\theta(i))_{i=1}^d$ of $(\psi(i))_{i=1}^d$ into Equation (1). The Wang–Landau algorithm is an iterative algorithm that jointly generates a sequence of estimates $(\theta_t(i))_{t\geq 0}$ for all i and a Markov chain $(X_t)_t$, such that when t goes to infinity, $\theta_t(i)$ converges to $\psi(i)$ and consequently, the distribution of X_t converges to $\tilde{\pi}$. We denote by $\tilde{\pi}_{\theta_t}$ the biased distribution obtained by replacing $\psi(i)$ by its estimate $\theta_t(i)$ in Equation (1). Note that the normalizing constant of $\tilde{\pi}_{\theta_t}$ is now unknown. A simplified version of the Wang–Landau algorithm is given in Algorithm 1.

Algorithm 1 Simplified Wang–Landau Algorithm

1: Partition the state space into d regions $\{\mathcal{X}_1, \ldots, \mathcal{X}_d\}$ along a reaction coordinate $\xi(x)$.
2: First, $\forall i \in \{1, \ldots, d\}$ set $\theta(i) \leftarrow 1$.
3: Choose a decreasing sequence $\{\gamma_t\}$, typically $\gamma_t = 1/t$.
4: Sample X_0 from an initial distribution π_0.
5: **for** $t = 1$ to T **do**
6: Sample X_t from $P_{\theta_{t-1}}(X_{t-1}, \cdot)$, a transition kernel with invariant distribution $\tilde{\pi}_{\theta_{t-1}}(x)$.
7: Update the bias: $\log \theta_t(i) \leftarrow \log \theta_{t-1}(i) + \gamma_t(\mathcal{I}_{\mathcal{X}_i}(X_t) - d^{-1})$.
8: Normalize the bias: $\theta_t(i) \leftarrow \theta_t(i)/\sum_{i=1}^d \theta_t(i)$.
9: **end for**

The rationale behind the update of the bias is that if the chain is in the set $\mathcal{X}_i$, the probability of remaining in $\mathcal{X}_i$ should be reduced compared with the other sets through an increase in the associated bias $\theta_t(i)$. Therefore, the chain is pushed toward the sets that have been visited less during the previous iterations, improving the exploration of the state space so long as the partition $(\mathcal{X}_i)_{i=1}^d$ is well chosen. While this biasing mechanism adds cost to each iteration of the algorithm, the tradeoff is improved exploration. In Step 6, the

transition kernel is typically an M–H move, due to the lack of conjugacy brought about by biasing.

In this simplified form, the Wang–Landau algorithm reduces to standard stochastic approximation, where the term γ_t decreases at each iteration. The algorithm as given in Wang and Landau (2001a,b) uses a more sophisticated learning rate γ_t that does not decrease deterministically, but instead only when a certain criterion is met. This criterion, referred to as the "flat histogram" criterion, is met when for all $i \in \{1, \ldots, d\}$, $\nu(i)$ is close enough to d^{-1}, where we denote by $\nu(i)$ the proportion of visits of (X_t) in the set $\mathcal{X}_i$ since the last time the criterion was met. Hence, we introduce a real number c to control the distance between $\nu(i)$ and d^{-1}, and an integer k to count the number of criteria already met. We describe the generalized Wang–Landau algorithm in Algorithm 2.

Algorithm 2 Wang-Landau Algorithm

Partition the state space into d regions $\{\mathcal{X}_1, \ldots, \mathcal{X}_d\}$ along a reaction coordinate $\xi(x)$.
First, $\forall i \in \{1, \ldots, d\}$ set $\theta(i) \leftarrow 1$, $\nu(i) \leftarrow 0$.
Choose a decreasing sequence $\{\gamma_k\}$, typically $\gamma_k = 1/k$.
Sample X_0 from an initial distribution π_0.
for $t = 1$ to T **do**
  Sample X_t from $P_{\theta_{t-1}}(X_{t-1}, \cdot)$, a transition kernel with invariant distribution $\tilde{\pi}_{\theta_{t-1}}(x)$.
  Update the proportions: $\forall i \in \{1, \ldots, d\} \quad \nu(i) \leftarrow \frac{1}{t}[(t-1)\nu(i) + \mathcal{I}_{\mathcal{X}_i}(X_t)]$.
  if "flat histogram": $\max_{i \in [1,d]} |\nu(i) - d^{-1}| < c/d$ **then**
    Set $k \leftarrow k + 1$.
    Reset $\forall i \in \{1, \ldots, d\} \quad \nu(i) \leftarrow 0$.
  end if
  Update the bias: $\log \theta_t(i) \leftarrow \log \theta_{t-1}(i) + \gamma_k(\mathcal{I}_{\mathcal{X}_i}(X_t) - d^{-1})$.
  Normalize the bias: $\theta_t(i) \leftarrow \theta_t(i) / \sum_{i=1}^{d} \theta_t(i)$.
end for

When c is set to low values (e.g., $c = 0.1$ or 0.5), the algorithm must explore the various regions such that the frequency of visits to the region $\mathcal{X}_i$ is approximately d^{-1} before the learning rate γ_k is decreased. Also, the algorithm may be further generalized to target a desired frequency ϕ_i instead of the same frequency d^{-1} for every set; while such strategies may be useful, as demonstrated in the following section, for notational simplicity we focus on the case $\phi_i = d^{-1}$. As already mentioned, to answer the general question of exploring the support of a target density π, the default choice for the reaction coordinate is the energy function: $\xi(x) = -\log \pi(x)$, which has the benefit of being one-dimensional regardless of the dimension of the state space $\mathcal{X}$. However, for specific models other reaction coordinates have been used, such as one (or more) of the components x_j of x or a linear combination of components of x. In the applications in Section 4, we discuss the use of alternative reaction coordinates further.

We now propose improvements to the Wang–Landau algorithm to increase its flexibility and efficiency.

3.2 A NOVEL ADAPTIVE BINNING STRATEGY

The Wang–Landau and equi-energy sampler algorithms are known to perform well if the bins, or partitions of the one-dimensional reaction coordinate $\xi(x)$, are well chosen. However, depending on the problem it might be difficult to choose the bins to optimize sampler performance. A typical empirical approach to deal with this issue is to first run, for example, an adaptive MCMC algorithm to find at least one mode of the target distribution. The generated sample and the associated target density evaluations determine a first range of the target density values, which can be used to initialize the bins. At this point, the user can choose a wider range of target density values (e.g., by multiplying the range by 2), to allow for a wider exploration of the space. Within this initial range, one must still decide the number of bins.

Due to difficulties with selecting the bins, it has been suggested that one should adaptively compare adjacent bins, splitting a bin if the corresponding estimate θ is significantly larger than a neighboring value (Schmidler 2011). Because each ψ_i is a given bin's normalizing constant, we feel it is more important to maintain uniformity within a bin to allow easy within-bin movement. Our proposed approach to achieve this "flatness" is to look at the distribution of the realized reaction coordinate values within each bin. Figure 2 illustrates this distribution on an artificial histogram. The plot of Figure 2(a) shows a situation where, within one bin, the distribution might be strongly skewed toward one side. In this artificial example, very few points have visited the left side of the bin, which suggests that moving from this bin to the left neighboring bin might be difficult.

We propose to consider the ratio of the number of points on the left side of the middle (dashed line) over the number of points within the bin as a very simple proxy for the discrepancy of the chains within one bin (see, e.g., Niederreiter 1992 for much more sophisticated discrepancy measures). In a broad outline, if this ratio was around 50%, the within-bin histogram would be roughly uniform. On the contrary, the ratio corresponding

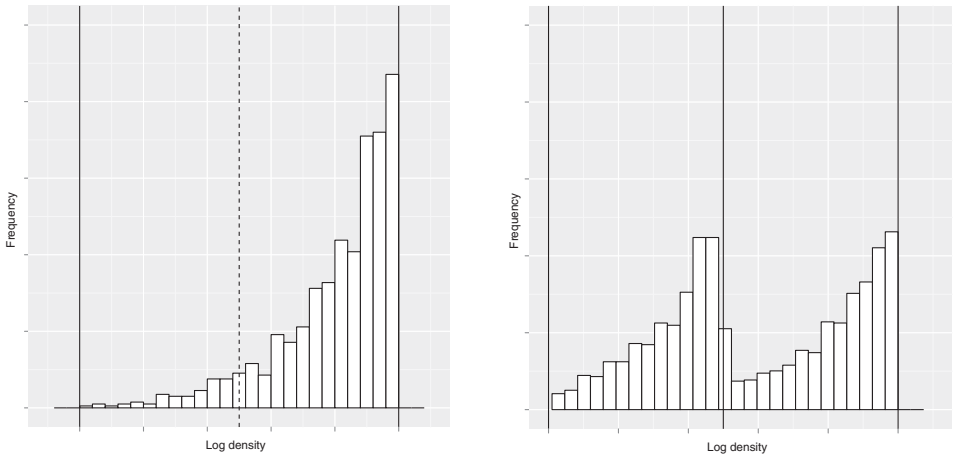

Figure 2. Artificial histograms of the log target density values associated with the chains generated by the algorithm, within a single bin (left) and within two bins created by splitting the former bin (right).

to Figure 2 (left) is around 7%. Our strategy is to split the bin if this ratio goes below a given threshold, say 25%; two new bins are created, corresponding to the left side and the right side of the former bin, and each bin is assigned a weight of $\theta/2$, where θ is the weight of the former bin. These provide starting values for the estimation of the weight of the new bins during the following iterations of the algorithm. Note also that the desired frequency of visits to each of the new bins, which was for instance equal to $1/d$ before the split, has to be specified as well. In the numerical experiments, we set the desired frequency of the new bins as one-half of the desired frequency of the former bin. Figure 2 (right) shows the distribution of samples within the two new bins. The resulting histogram is not uniform, yet exhibits a more even distribution within the bin—a feature that is expected to help the chain to move from this bin to the left neighboring bin. The threshold could be set closer to 50%, which would result in more splits and therefore more bins.

In practice, it is not necessary to check whether the bins have to be split at every iteration. Our strategy is to check every nth iteration, until the flat histogram criterion is met for the first time. When it is met, it means that the chains can move easily between the bins, and hence the bins can be kept unchanged for the remaining iterations. Finally, when implementing the automatic binning strategy for discrete distributions, one must ensure that a new bin corresponds to actual points in the state space. For example, if the bins are along the energy values and the state space is finite, there are certainly intervals of energy to which no state corresponds, and that would therefore never be reached. Section 4 demonstrates the proposed adaptive binning strategy in practice.

In addition to allowing for splitting of bins, it is also important to allow the range of bins to extend if new states are found outside of the particular range. That said, one must differentiate between the two extremes of the reaction coordinate. For example, if $\xi(x) = -\log \pi(x)$, then one might not wish to add more low-density (high-energy) bins, which would induce the sampler to explore further and further into the tails. However, if one finds a new high-density (low-energy) mode beyond the energy range previously seen, then the sampler might become stuck in this new mode. In this case, we propose to extend the first bin corresponding to the lowest level of energy to always include the lowest observed values. The adaptive partition $(\mathcal{X}_{i,t})_{i=1}^{d_t}$ of the state space takes the following form at time t:

$$\mathcal{X}_{1,t} = [e_{\min,t}, e_{1,t}],\ \mathcal{X}_{2,t} = [e_{1,t}, e_{2,t}],\ \ldots \mathcal{X}_{d_t,t} = [e_{d_t-1,t}, +\infty),$$

where $e_{\min,t} = \min_{t\geq 0}\{-\log \pi(X_t)\}$ and $(e_{1,t}, \ldots, e_{d_t-1,t})$ defines the limits of the inner bins at time t, which is the result of initial bin limits $(e_{1,0}, \ldots, e_{d_0-1,0})$, and possible splits between time 0 and time t. As such, if new low-energy values are found, the bin $\mathcal{X}_{1,t}$ is widened. If this results in unequal exploration across the reaction coordinate, then the adaptive bin-splitting mechanism will automatically split this newly widened bin.

3.3 PARALLEL INTERACTING CHAINS

We propose to generate multiple chains instead of a single one to improve computational scalability through parallelization as well as particle diversity. The use of interacting chains has become of much interest in recent years, with the multiple chains used to create diverse proposals (Casarin, Craiu, and Leisen 2011), to induce long-range equi-energy jumps (Kou, Zhou, and Wong 2006), and to generally improve sampler performance; see

Atchadé et al. (2011), Brockwell, Del Moral, and Doucet (2010), and Byrd (2010) for recent developments. The use of parallelization is not constrained to multiple chains, however, and has also been employed to speed up the generation of a single chain through prefetching (Brockwell 2006).

Let N be the desired number of chains. We follow Algorithm 2, with the following modifications. First, we generate N starting points $\boldsymbol{X}_0 = (X_0^{(1)}, \ldots, X_0^{(N)})$ independently from an initial distribution π_0 (Algorithm 2 line 4). Then, at iteration t, instead of moving one chain using the transition kernel $P_{\theta_{t-1}}$, we move the N chains using the same transition kernel, associated with the same bias θ_{t-1} (Algorithm 2 line 6). We emphasize that the bias θ is common to all chains, which makes the proposed method different from running Wang–Landau chains entirely in parallel. The proportions $\nu(i)$ are updated using all the chains, simply by replacing the indicator function $\mathcal{I}_{\mathcal{X}_i}(X_t)$ by the mean $N^{-1}\sum_{j=1}^{N}\mathcal{I}_{\mathcal{X}_i}(X_t^{(j)})$, that is, the proportion of chains currently in set $\mathcal{X}_i$ (Algorithm 2 line 7). Likewise, the update of the bias uses all the chains, again replacing the indicator function by the proportion of chains currently in a given set (Algorithm 2 line 12). We have therefore replaced indicator functions in the Wang–Landau algorithm by the law of the MCMC chain associated with the current parameter. Since this law is not accessible, we perform a mean field approximation at each time step. A similar expression has recently been employed by Liang and Wu (2011) for use in parallelizing the stochastic approximation Monte Carlo (SAMC) algorithm (Liang, Liu, and Carroll 2007; Liang 2009). Note that while we have designed the chains to communicate at each iteration, such frequent message passing can be costly, particularly on graphics processing units. In such situations, one could alter the algorithm such that the chains only communicate periodically.

Our results (see Section 4) show that N interacting chains run for T iterations can strongly outperform a single-chain run for $N \times T$ iterations, in terms of variance of the resulting estimates. Specifically, having a sample approximately distributed according to $\pi_{\theta_t}(x)$ instead of a single point at iteration t improves and stabilizes the subsequent estimate θ_{t+1}. We explore the tradeoff between N and T in more detail in Section 4.

Note that, while the original single-chain Wang–Landau algorithm was not straightforward to parallelize due to its iterative nature, the proposed algorithm can strongly benefit from multiple processing units: at a given iteration the N move steps can be done in parallel, as long as the results are consequently collected to update the bias before the next iteration. Therefore, if multiple processors are available, as, for example, in recent central processing units and in graphics processing units (see, e.g., Suchard and Rambaut 2009; Lee et al. 2010), the computational cost can be reduced much more than what was possible with the single-chain Wang–Landau algorithm. To summarize, the proposed use of interacting chains can both improve the convergence of the estimates, regardless of the number of available processors, and additionally benefit from multiple processors.

Finally, an additional benefit of using N parallel chains is that they can start from various points, drawn from the initial distribution π_0; hence if π_0 is flat enough, the chains can start from different local modes, which improves de facto the exploration. However, we show in Section 4 that the chains still explore the space even if they start within the same mode, and hence the efficiency of the method does not rely on the choice of π_0, contrary to what we observed with SMC methods. Additionally, because the sampler is attempting to explore both the state space as well as the range of the reaction coordinate simultaneously,

our parallel formulation allows the sampler to borrow strength between chains, providing for exploration of the reaction coordinate without having to move a single chain across potentially large and high-dimensional state-spaces to traverse the reaction coordinate values.

3.4 ADAPTIVE PROPOSAL MECHANISM

As discussed earlier, it is important to automate the proposal mechanism to improve movement across the state space. A well-studied proxy for optimal movement is the algorithm's M–H acceptance rate. Too low an acceptance rate signifies the algorithm is attempting to make moves that are too large, and are therefore rejected. Too high an acceptance rate signifies the algorithm is barely moving. As such, we suggest adaptively tuning the proposal variance to encourage an acceptance rate of 0.234 as recommended in Roberts, Gelman, and Gilks (1997), although we have found settings in the range 0.1–0.5 to work well in all examples tested. The Robbins–Monro stochastic approximation update of the proposal standard deviation σ_t is

$$\sigma_{t+1} = \sigma_t + \rho_t \left(2\mathcal{I}(A > 0.234) - 1\right),$$

where t is the current iteration of the algorithm, ρ_t is a decreasing sequence (typically $\rho_t = 1/t$), and A is the acceptance rate (proportion of accepted moves) of the particles. Through this update, the proposal variance grows after samples are accepted, and shrinks when samples are rejected, encouraging exploration of the state-space.

Another approach to adaptively tuning the proposal distribution is to use the following mixture of Gaussian random walks:

$$X^* \sim w_1 \mathcal{N}\left(X_{t-1}, \frac{(2.38)^2}{p}\Sigma_t\right) + w_2 \mathcal{N}\left(X_{t-1}, \frac{(\sigma_I)^2}{p} I_p\right)$$

with $w_1 + w_2 = 1$, Σ_t being the empirical covariance of the chain history—an estimator of the covariance structure of the target—and I_p being the $p \times p$ identity matrix, where p is the dimension of the target space. The first component of this mixture makes the proposal adaptive and able to learn from the past, while the second component helps to explore the space. For instance, if the chain is stuck in a mode, the first component's variance might become small, yet the second component guarantees a chance to eventually escape the mode. Hence, the second component acts as a "safety net" and therefore its weight is small, typically $w_2 = 0.05$, and its standard deviation σ_I may be set large to improve mixing (Guan and Krone 2007).

In our context where parallel chains are run together, we use all the chains to estimate the empirical covariance Σ_t at each iteration. Note that the computation of this covariance does not require the storage of the whole history of the chain and can be done at constant cost, since recurrence formulas exist to compute the empirical covariance, as explained for instance in Welford (1962). The value 2.38^2 is justified by asymptotic optimality reflections on certain classes of models (see, e.g., Roberts, Gelman, and Gilks 1997; Roberts and Rosenthal 2009).

3.5 USING THE OUTPUT TO PERFORM INFERENCE

While the resulting samples from the proposed algorithm PAWL are not from π, but rather an approximation of the biased version (1), one can use importance sampling or advanced SMC ideas to transition the samples to π (see Chopin and Jacob 2010 for details). Alternatively, the samples from the exploratory algorithm can be used to seed a more traditional MCMC algorithm, as advocated by Schmidler (2011).

The pseudocode for PAWL, combining the parallel Wang–Landau algorithm with adaptive binning and proposal mechanisms, is given in the Appendix. Before proceeding to examples, it is important to reiterate the importance of the values $\psi(i) = \int_{\mathcal{X}_i} \pi(x)dx$. Specifically, certain choices of the reaction coordinate $\xi(x)$ result in $\psi(i)$ having inherent value. For example, it is possible in a model selection application to use the model order as $\xi(x)$, in which case the values $\psi(i)$ could be employed to calculate Bayes factors and other quantities of interest.

4. APPLICATIONS

We now demonstrate PAWL applied to three examples including variable selection, mixture modeling, and spatial imaging. A fourth pedagogical example is available as supplementary material. In each application, we walk through our proposed algorithm (described explicitly as Algorithm 3 in the Appendix), first running preliminary (adaptive) M–H MCMC to determine the initial range for the reaction coordinate ξ and initial values for the proposal parameters and starting state of the interacting Wang–Landau chains. This range is then increased to encourage exploration of low-density regions of the space, and an initial number of bins is specified. Once this groundwork is set, the same M–H algorithm run in the preliminary stage is embedded within the PAWL algorithm.

4.1 *g*-PRIOR VARIABLE SELECTION

We proceed by conducting variable selection on the pollution dataset of McDonald and Schwing (1973), wherein mortality is related to pollution levels through 15 independent variables including mean annual precipitation, population per household, and average annual relative humidity. Measured across 60 metropolitan areas, the response variable $\mathbf{y}$ is the age-adjusted mortality rate in the given metropolitan area. Our goal is to identify the pollution-related independent variables that best predict the response. With 15 variables, calculating the posterior probabilities of the 32,768 models exactly is possible but time-consuming. We have chosen this size of dataset to provide for difficult posterior exploration, yet allow a study of convergence of θ toward ψ.

With an eye toward model selection, we introduce the binary indicator variable $\gamma \in \{0, 1\}^p$, where $\gamma_j = 1$ means the variable $\boldsymbol{x}_j$ is included in the model. As such, γ can describe all of the 2^p possible models. Consider the normal likelihood

$$\mathbf{y}|\mu, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).$$

If $\mathbf{X}_\gamma$ is the model matrix that excludes all $\boldsymbol{x}_j$'s if $\gamma_j = 0$, we can employ the following prior distributions for $\boldsymbol{\beta}$ and σ^2 (Zellner 1986; Marin and Robert 2007):

$$\pi(\boldsymbol{\beta}_\gamma, \sigma^2|\gamma) \propto (\sigma^2)^{-(q_\gamma+1)/2-1} \exp\left[-\frac{1}{2g\sigma^2}\boldsymbol{\beta}_\gamma^T\left(\mathbf{X}_\gamma^T\mathbf{X}_\gamma\right)\boldsymbol{\beta}_\gamma\right],$$

where $q_\gamma = \mathbf{1}_n^T\gamma$ represents the number of variables in the model. While selecting g can be a difficult problem, we have chosen it to be very large ($g = \exp(20)$) to induce a sparse model, which is difficult to explore due to the small marginal probabilities of most variables. After integrating over the regression coefficients $\boldsymbol{\beta}$, the posterior density for γ is thus

$$\pi(\gamma|\mathbf{y}, \mathbf{X}) \propto (g+1)^{-(q_\gamma+1)/2}\left[\mathbf{y}^T\mathbf{y} - \frac{g}{g+1}\mathbf{y}^T\mathbf{X}_\gamma\left(\mathbf{X}_\gamma^T\mathbf{X}_\gamma\right)^{-1}\mathbf{X}_\gamma\mathbf{y}\right]^{-n/2}.$$

While we select the log energy function $-\log\pi(x)$ as the reaction coordinate $\xi(x)$ for our analysis, it is worth noting that many other options exist. For instance, it would be natural to consider the model saturation q_γ/p, which would ensure exploration across the different model sizes. However, we select $\xi(x) = -\log\pi(x)$ to emphasize the universality of using the energy function as the reaction coordinate.

We first run a preliminary M–H algorithm that flips a variable on/off at random, accepting or rejecting the flip based on the resulting posterior densities. Due to high correlation between variables, a better strategy might be to flip multiple variables at once; however, we restrain from exploring this to demonstrate PAWL's ability to make viable even poorly designed M–H algorithms. The preliminary algorithm run found values $377 < -\log\pi(x) < 410$, which we extend slightly to create 20 equally spaced bins in the range $[377, 450]$. It is worth reiterating that the resulting samples generated from PAWL are from a biased version of π; as such, importance sampling techniques could be used to recover π, or the samples obtained could be used to seed a more traditional MCMC algorithm.

Due to the size of the problem, we are able to enumerate all posterior values, and hence may calculate ψ exactly. As such, we begin by examining the effect of the number of particles N on the parallel Wang–Landau algorithm. To further focus on this aspect, we suppress adaptive binning and proposals for this example. Figure 3 shows the convergence of θ to Ψ for $N = 1, 10, 100$. We see that the algorithm's convergence improves with more particles. Using $N = 100$ particles, we now examine PAWL compared with the M–H algorithm (run on N chains) mentioned above on the unnormalized targets π, $\pi^{1/10}$, $\pi^{1/100}$. Consider Figure 4; on the target distribution π, the M–H algorithm becomes stuck in high-probability regions. However, on the tempered distributions, the algorithm explores the space more thoroughly, although not to the same level as PAWL. Specifically, PAWL explores a much wider range of models, including the highest probability models, whereas the tempered distributions do not. Here, the Wang–Landau algorithm as well as the M–H algorithm both use $N = 100$ chains for $T = 3500$ iterations, the former taking 253 ± 13 sec and the latter taking 247 ± 15 sec across 10 runs, indicating that the additional cost for PAWL is negligible.

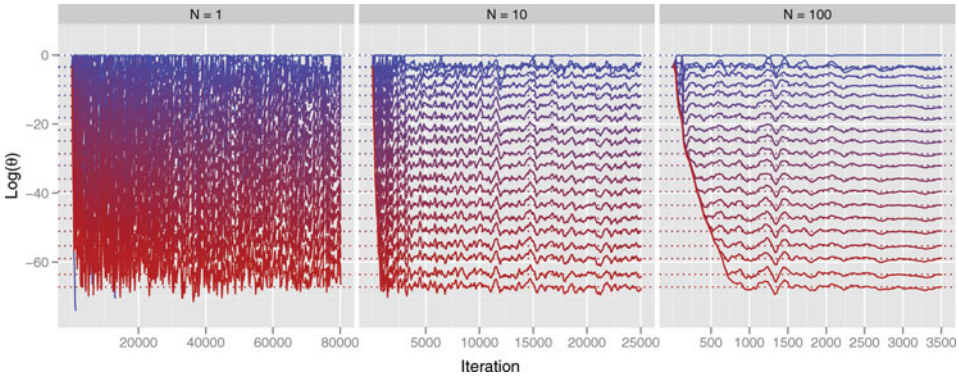

Figure 3. Variable selection example: convergence of Wang–Landau for $N = 1, 10, 100$. Iterations set such that each algorithm runs in 2 min (± 5 sec). θ for each bin is shown as solid lines. True values (ψ) are shown as dotted lines. The online version of this figure is in color.

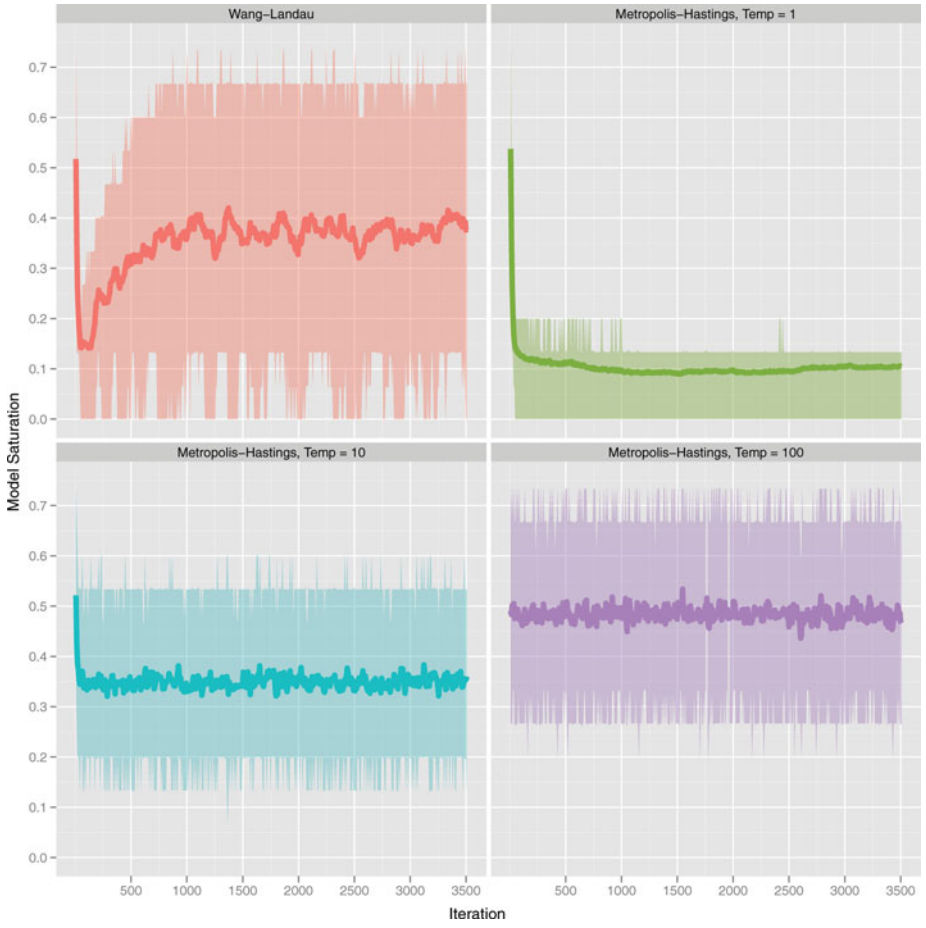

Figure 4. Variable selection example: exploration of model space by PAWL and Metropolis–Hastings (M–H) at three temperature settings. Model saturation (proportion of nonzero variables in model) as function of algorithm iterations. The solid lines are the mean of $N = 100$ chains, while the shaded regions are the middle 95% of the chains. The online version of this figure is in color.

4.2 MIXTURE MODELING

Mixture models provide a challenging case of multimodality, due partly to the complexity of the model and partly to a phenomenon called "label switching" (see, e.g., Frühwirth-Schnatter (2006) for a book covering Bayesian inference for these models, Diebolt and Robert (1994) and Richardson and Green (1997) for seminal articles using MCMC for mixture models, and Stephens (2000) and Jasra, Holmes, and Stephens (2005) on the label switching problem). Articles describing explorative MCMC algorithms often take these models as benchmarks, as, for example, population MCMC and SMC methods in Jasra, Stephens, and Holmes (2007), the Wang–Landau algorithm in Atchadé and Liu (2010), free energy methods in Chopin, Lelievre, and Stoltz (2012) and Chopin and Jacob (2010), and parallel tempering with equi-energy moves in Baragatti, Grimaud, and Pommeret (2012).

Consider a Bayesian Gaussian mixture model, that is, for $i = 1, \dots n$,

$$p(y_i|q, \mu, \lambda) = \sum_{k=1}^{K} q_k \, \varphi\big(y_i; \mu_k, \lambda_k^{-1}\big),$$

where φ is the probability density function of the Gaussian distribution, K is the number of components, q_k, μ_k, and λ_k are, respectively, the weight, the mean, and the precision of component k. The component index k is also called its label. Following Richardson and Green (1997), the prior is taken as, for $k = 1, \dots, K$,

$$\begin{aligned} \mu_k &\sim \mathrm{N}(M, \kappa^{-1}), & \lambda_k &\sim \mathrm{Gamma}(\alpha, \beta), \\ \beta &\sim \mathrm{Gamma}(g, h), & (q_1, \dots, q_{K-1}) &\sim \mathrm{Dirichlet}_K(1, \dots, 1), \end{aligned}$$

with, for example, $\kappa = 4/R^2$, $\alpha = 2$, $g = 0.2$, $h = 100g/\alpha R^2$, $M = \bar{y}$, $R = \mathrm{range}\,(y)$.

The invariance of the likelihood to permutations in the labeling of the components leads to the "label switching" problem: since there are $K!$ possible permutations of the labels, each mode has necessarily $K! - 1$ replicates. We emphasize that this model has been thoroughly studied and is hence well-understood from a modeling point of view, but it still induces a computationally challenging sampling problem for which difficulty can be artificially increased through the number of components K.

Note that in this parameterization β, the rate of the Gamma prior distribution of the precisions λ_k is estimated along with the parameters of interest $q_{1:K-1}$, $\mu_{1:K}$, and $\lambda_{1:K}$. Chopin, Lelievre, and Stoltz (2012) and Chopin and Jacob (2010) suggested that β can be used as a reaction coordinate, since a large value of β results in a small precision and hence in a flatter posterior distribution of the other parameters, which is easier to explore than the distribution associated with smaller values of β; we refer to these articles for further exploration of this choice of reaction coordinate, and instead default to $\xi(x) = -\log \pi(x)$.

We create a synthetic 100-sample from a Gaussian mixture with $k = 4$ components, weights $1/4$, means -3, 0, 3, 6, and variances 0.55^2 as in Jasra, Holmes, and Stephens (2005). The goal is to explore the highly multimodal posterior distribution of the 13-dimensional parameter $\theta = (w_{1:4}, \mu_{1:4}, \lambda_{1:4}, \beta)$, where w_k is the unnormalized weight: $q_k = w_k / \sum_{k=1}^{K} w_k$. Unnormalized weights may be handled straightforwardly in MCMC algorithms since they are defined on $\mathbb{R}^+$ and not on the K-simplex as with the q_k.

The proposed algorithm is compared with an SMC and a parallel adaptive Metropolis–Hastings (PAMH) algorithm, that we detail below. We admittedly use naive versions of these competitors, arguing that most improvements of these could be carried over to PAWL. For instance, a mixture of Markov kernels as suggested for the SMC algorithm in Section 3.2 of Jasra, Stephens, and Holmes (2007) can be used in the proposal distribution of PAWL; and since PAWL is a population MCMC algorithm, exchange and crossover moves could be used as well, as suggested for the Population SAMC algorithm in Liang and Wu (2011). To get a plausible range of values for the reaction coordinate of the proposed algorithm without user input, an initial adaptive MCMC algorithm is run with $N = 10$ chains and $T^{\text{init}} = 1000$ iterations. The initial points of these chains are drawn from the prior distribution of the parameters. This provides a range of log density values, from which we compute the 10% and 90% empirical quantiles, denoted by q_{10} and q_{90}, respectively. In a conservative spirit, the bins are chosen to equally divide the interval $[q_{10},\ q_{10} + 2(q_{90} - q_{10})]$ in 20 subsets. Hence, the algorithm is going to explore log density values in a range that is approximately twice as large as the values initially explored. Note that we use quantiles instead of minimum and maximum values to make the method more robust.

Next, PAWL itself is run for $T = 200{,}000$ iterations, starting from the terminal points of the N preliminary chains, resulting in a total number of $N(T + T_{\text{init}})$ target density evaluations. In this situation, even with only 100 data points, most of the computational cost goes into the evaluation of the target density. This confirms that algorithmic parameters such as the number of bins do not significantly affect the overall computational cost, at least as long as the target density is not extremely cheap to evaluate. The adaptive proposal is such that it targets an acceptance rate of 23.4%. Meanwhile, the PAMH algorithm using the same adaptive proposal is run with $N = 10$ chains and $T^* = 250{,}000$ iterations, hence relying on more target density evaluations for a comparable computational cost.

Finally, the SMC algorithm is run on a sequence of tempered distribution $(\pi_k)_{k=1}^K$, each density being defined by

$$\pi_k(x) \propto \pi^{\zeta_k}(x) p_0^{1-\zeta_k}(x),$$

where p_0 is an initial distribution (here taken to be the prior distribution), and $\zeta_k = k/K$. The number of steps K is set to 100 and the number of particles to 40,000. When the effective sample size goes below 90%, we perform a systematic resampling and five consecutive M–H moves. We use a random walk proposal distribution, in which variance is taken to be $c\hat{\Sigma}$, where $\hat{\Sigma}$ is the empirical covariance of the particles and c is set to 10%; see Jasra, Stephens, and Holmes (2007) for more details. The parameters are chosen to induce a computational cost comparable to the other methods. However, for the SMC sampler, the number of target density evaluations is a random number, since it depends on the random number of resampling steps: the computational cost is in general less predictable than using MCMC.

First, we look at graphical representations of the generated samples. Figure 5 shows the resulting points projected on the (μ_1, μ_2) plane, restricted on $[-5, 9]^2$. In this plane, there are 12 replicates of each mode, indicated by target symbols in Figure 5(a). These projections do not allow one to check that all the modes were visited since they project the 13-dimensional target space on a two-dimensional space. Figure 5(b) shows that the adaptive MCMC method clearly misses some of the modes, while visiting many others. Figure 5(c)

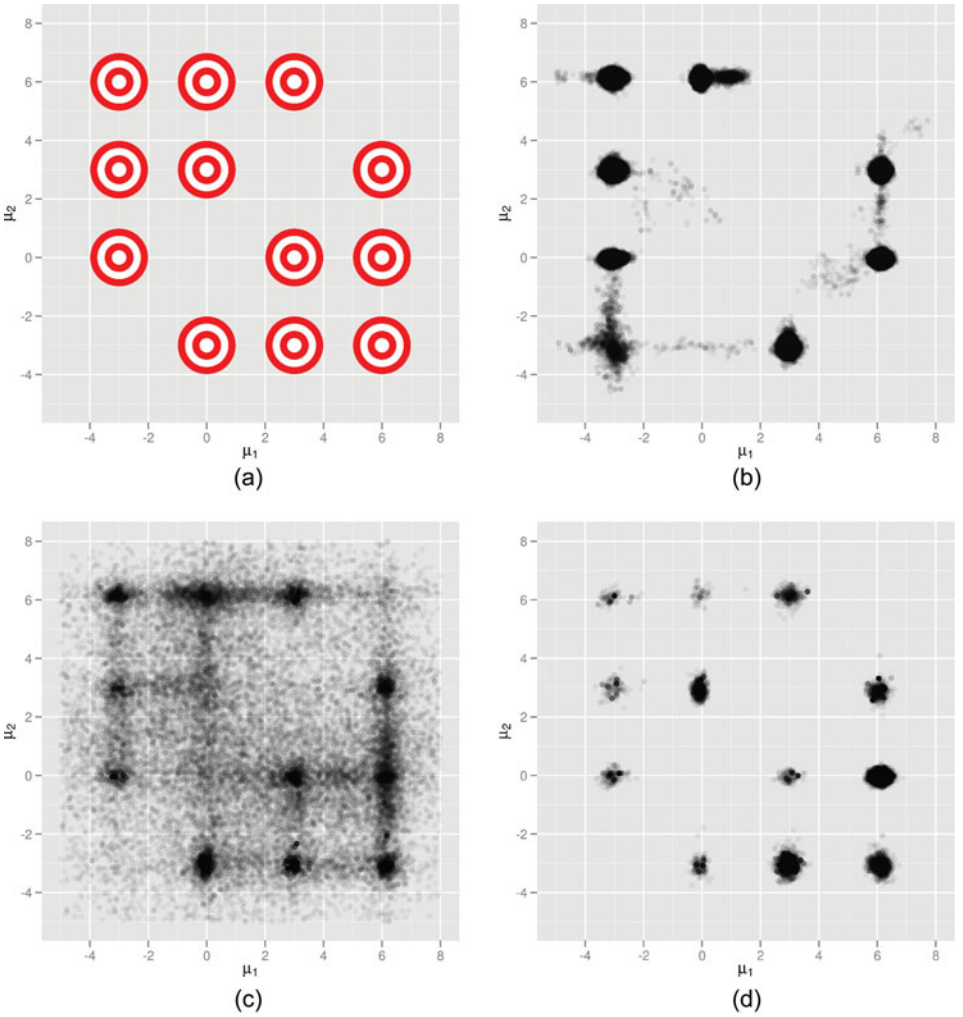

Figure 5. Mixture model example: exploration of the posterior distribution projected on the μ_1, μ_2 plane, using different algorithms. (a) Locations of the global modes of the posterior distribution projected on (μ_1, μ_2). (b) Projection of the chains generated by the parallel adaptive M–H algorithm on (μ_1, μ_2). (c) Projection of the chains generated by PAWL on (μ_1, μ_2). (d) Projection of the particles generated by the SMC algorithm on (μ_1, μ_2). The online version of this figure is in color.

shows how the chains generated by the modified Wang–Landau algorithm easily explore the space of interest, visiting both global and local modes. To recover the main modes from these chains, we use the final value of the bias, θ_T, as importance weights to correct for the bias induced by the algorithm; in Figure 5(c), the importance weights define the transparency of each point: the darker the point, the more weight it has. Finally, Figure 5(d) shows how the SMC sampler also put particles in each mode; again the transparency of the points is proportional to their weights.

We now turn to more quantitative measures of the error made on marginal quantities. Since the component means admit four identical modes around -3, 0, 3, and 6, we know

Table 1. Estimation of the means of the mixture components, for the proposed method (PAWL), parallel adaptive Metropolis–Hastings (PAMH), and sequential Monte Carlo (SMC), using the prior as initial distribution. Quantities averaged over 10 independent runs for each method

Method	μ_1	μ_2	μ_3	μ_4	Error	Time (sec)
PAWL	1.42 ± 0.99	1.42 ± 0.58	1.39 ± 0.90	1.75 ± 0.78	1.50 ± 0.59	209 ± 1
PAMH	1.58 ± 0.81	1.25 ± 0.72	1.04 ± 1.07	2.09 ± 1.00	1.75 ± 0.80	233 ± 1
SMC	1.00 ± 1.96	2.99 ± 1.38	0.92 ± 2.27	1.10 ± 2.11	3.89 ± 1.34	269 ± 7

that their overall mean is approximately equal to $\mu^* = 1.5$. We then compute the following error measurement:

$$\text{error} = \sqrt{\sum_{k=1}^{K} (\hat{\mu}_k - \mu^*)^2},$$

where $\hat{\mu}_k$ is the mean of the generated sample (taking the weights into account for PAWL and SMC). Table 1 shows the results averaged over 10 independent runs: the means of each component, the error defined above and the (wall-clock) run times obtained using the same CPU, along with their standard deviations. The results highlight that in this context PAWL gives more precise results than SMC, for the same or less computational cost; the comparison between parallel MCMC and SMC confirms the results obtained in Jasra, Stephens, and Holmes (2007). The small benefit of PAWL over PAMH can be explained by considering the symmetry of the posterior distribution: even if some modes are missed by PAMH as shown in Figure 5(b), the approximation of the posterior expectation might be accurate, though the corresponding variance will be higher.

Next, we consider a more realistic setting, where the initial distribution is not well spread over the parameter values: instead of taking the prior distribution itself, we use a similar distribution but with an hyperparameter κ equal to 1 instead of $4/R^2$, which for our simulated dataset is equal to 0.03. This higher precision makes the initial distribution concentrated on a few modes, instead of being fairly flat over the whole region of interest. We keep the prior unchanged, so that the posterior is left unchanged. For PAWL and PAMH, this means that the initial points of the chains are all close one to another; and likewise for the initial particles in the SMC sampler. The results are shown in Table 2, and illustrate the degeneracy of SMC when the initial distribution is not well-chosen; though this is not surprising, this is important in terms of exploratory algorithms when one does not have prior knowledge of the region of interest. Both parallel MCMC methods give similar results as with the previous, flatter initial distribution.

Table 2. Estimation of the means of the mixture components, for the proposed method (PAWL), PAMH, and SMC, using a concentrated initial distribution. Quantities averaged over 10 independent runs for each method

Method	μ_1	μ_2	μ_3	μ_4	Error	Time
PAWL	1.16 ± 0.75	2.04 ± 0.50	1.72 ± 0.80	1.07 ± 1.22	1.48 ± 1.10	210 ± 1
PAMH	1.37 ± 0.73	1.48 ± 1.39	1.71 ± 0.81	1.44 ± 1.11	1.75 ± 1.01	234 ± 1
SMC	0.35 ± 2.13	0.82 ± 1.55	3.19 ± 2.41	1.62 ± 1.85	4.17 ± 1.41	337 ± 8

Table 3. Estimation of the means of the mixture components, for the proposed method (PAWL), for different values of N, the number of chains, and T, the number of iterations. Quantities averaged over 10 independent runs for each set of parameters

Parameters	μ_1	μ_2	μ_3	μ_4	Error	Time
$N = 1$ $T = 5 \times 10^5$	0.37 ± 3.46	2.01 ± 3.27	2.53 ± 3.04	0.95 ± 3.46	6.39 ± 1.30	265 ± 40
$N = 10$ $T = 2 \times 10^5$	1.42 ± 0.99	1.42 ± 0.58	1.39 ± 0.90	1.75 ± 0.78	1.50 ± 0.59	209 ± 1
$N = 50$ $T = 5 \times 10^4$	1.51 ± 0.88	1.5 ± 0.9	1.65 ± 0.64	1.31 ± 0.31	1.22 ± 0.69	178 ± 2

Finally, we compare different algorithmic settings for the PAWL algorithm, changing the number of chains and the number of iterations. The results are shown in Table 3. First, we see that, even on a single CPU, the computing time is not exactly proportional to $N \times T$, the number of target density evaluation. Indeed, the computations are vectorized by iteration, and hence it is typically cheaper to compute one iteration of N chains than N iterations of one chain; although this would not hold for every model. We also see that the algorithm using only one chain failed to explore the modes, resulting in a huge final error. Finally, we see that with 50 chains and only 50,000 iterations, the algorithm provides results of approximately the same precision as with 10 chains and 200,000 iterations. This suggests that the algorithm might be particularly interesting if parallel processing units are available, since the computational cost would then be much reduced.

4.3 SPATIAL IMAGING

We finish our examples by identifying ice floes from polar satellite images as described in Banfield and Raftery (1992). Here, the image under consideration is a 200 by 200 grayscale satellite image, with focus on a particular 40 by 40 region (y, Figure 6); the goal is to identify

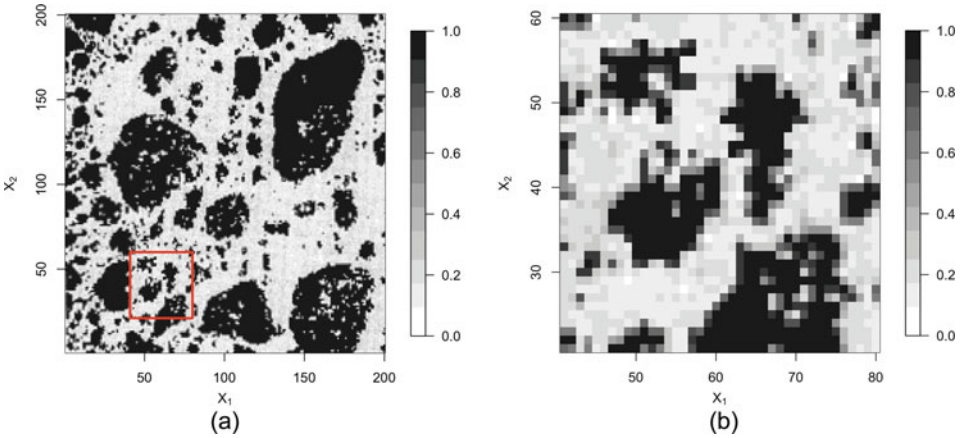

Figure 6. Spatial model example: (a) original ice floe image with highlighted region and (b) close-up of focused region. The online version of this figure is in color.

the presence and position of polar ice floes (x). Toward this goal, Higdon (1998) employed a Bayesian model. Basing the likelihood on similarity to the image and employing an Ising model prior, the resulting posterior distribution is

$$\log(\pi(x|y)) \propto \alpha \sum_i I[y_i = x_i] + \beta \sum_{i \sim j} I[x_i = x_j].$$

The first term, the likelihood, encourages states x that are similar to the original image y. The second term, the prior, favors states x for which neighboring pixels are equal. Here, neighborhood ($\sim$) is defined as the eight vertical, horizontal, and diagonal adjacencies of each (interior) pixel.

Because the prior strongly prefers large blocks of identical pixels, an MCMC method that proposes to flip one pixel at a time will fail to explore the posterior, and hence Higdon (1998) suggested a partial decoupling technique specific to these types of models. However, to demonstrate PAWL's power and universality, we demonstrate its ability to make simple one-at-a-time M–H feasible in these models without more advanced decoupling methods.

First running a preliminary M–H algorithm of length 20,000, we use the range of explored energy values divided evenly across 10 bins. The algorithm subsequently splits bins six times (with splitting stopped once the algorithm reaches the extremes of the reaction coordinate values) resulting in 17 bins at the algorithm's conclusion. For both algorithms, we run 10 chains for 1,000,000 iterations with model parameters $\alpha = 1$, $\beta = 0.7$. Due to the flip-one-pixel approach, we suppress adaptive proposals for this example. In contrast to the mixture modeling example, in this example the target density is fairly straightforward to calculate, so it is a good worst-case comparison to demonstrate the additional time taken by the proposed algorithm. For this example, the M–H algorithm took 388 ± 21 sec across 10 runs, whereas PAWL required 478 ± 24 sec. Thus, in this case the Wang–Landau adds a 23% price to each iteration on average. However, as we will show, the exploration is significantly better, justifying the slight additional cost. Figure 7 shows a subset of the last 400,000 posterior realizations from one chain of each algorithm. We see that the proposed Wang–Landau algorithm encourages much more exploration of alternate modes. The corresponding average state explored over all 10 chains (after 400,000 burn-in) is

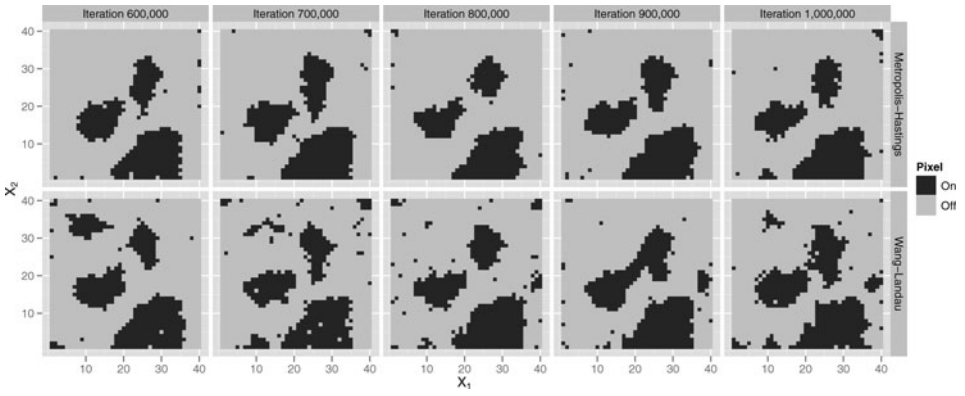

Figure 7. Spatial model example: states explored over 400,000 iterations for M–H (top) and proposed algorithm (bottom).

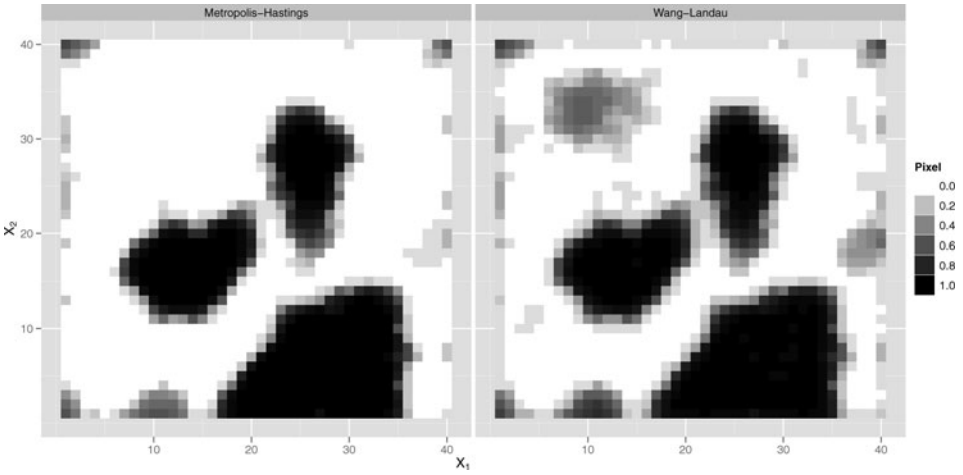

Figure 8. Spatial model example: average state explored with M–H (left) and PAWL after importance sampling (right).

shown in Figure 8. From this, we see that Wang–Landau induces exploration of the mode in the top left of the region in question, as well as a bridge between the central ice floes. In conclusion, while flip-one-pixel M–H is incapable of exploring the modes in the posterior caused by the presence/absence of large ice floes, the proposed algorithm encourages exploration of these modes, even in the presence of high between-pixel correlation. While Higdon (1998) developed a custom-tailored MCMC solution to overcome the inability of M–H to adequately explore the posterior density in Ising models, we employ PAWL—a general-purpose automatic density exploration algorithm—to achieve similar results.

5. DISCUSSION AND CONCLUSION

The proposed algorithm, PAWL, has at its core the Wang–Landau algorithm that, despite widespread use in the physics community, has only recently been introduced into the statistics literature. A well-known obstacle in implementing the Wang–Landau algorithm is selecting the bins through which to discretize the state-space; in response, we have developed a novel adaptive binning strategy. Additionally, we employ an adaptive proposal mechanism to further reduce the amount of user-defined parameters. Finally, to improve the convergence speed of the algorithm and to exploit modern computational power, we have developed a parallel interacting chain version of the algorithm, which proves efficient in stabilizing the algorithm. Through a host of examples, we have demonstrated the algorithm's ability to conduct density exploration over a wide range of distributions continuous and discrete. While a suite of custom-purpose MCMC tools exists in the literature for each of these models, the proposed algorithm handles each within the same unified framework.

As practitioners in fields ranging from image-processing to astronomy turn to increasingly complex models to represent intricate real-world phenomena, the computational tools to approximate these models must grow accordingly. In this article, we have proposed a general-purpose algorithm for automatic density exploration. Due to its fully adaptive

nature, we foresee its application as a black-box exploratory MCMC method aimed at practitioners of Bayesian methods. While statisticians are well-accustomed to performing exploratory analysis in the modeling stage of an analysis, the notion of conducting preliminary general-purpose exploratory analysis in the Monte Carlo stage (or more generally, the model-fitting stage) of an analysis is an area that we feel deserves much further attention. As models grow in complexity and endless model-specific Monte Carlo methods are proposed, it is valuable for the practitioner to have a universally applicable tool to throw at their problem before embarking on custom-tuned, hand-built Monte Carlo methods. Toward this aim, the authors have published an R package ("PAWL") to minimize user effort in applying the proposed algorithm to their specific problem.

APPENDIX: DETAILS OF PROPOSED ALGORITHM

Here, we detail PAWL, fusing together a Wang–Landau base with adaptive binning, interacting parallel chains, and an adaptive proposal mechanism. In comparison with the generalized Wang–Landau algorithm (Algorithm 2), when a flat histogram is reached, the distribution of particles within bins is tested to determine whether a given bin should be split. In addition, a suite of N interacting chains is employed, and hence the former chain $\boldsymbol{X}_t$ is now made of N chains: $\boldsymbol{X}_t = (X_t^{(1)}, \ldots, X_t^{(N)})$, each defined on the state space $\mathcal{X}$. All the N chains are used to update the bias θ_t, as described in Section 3.3.

The chains are moved using an adaptive mechanism determined by the M–H acceptance rate as explained in Section 3.4. While we present Algorithm 3 with adaptive proposal variance, it may also be implemented with an adaptive mixture proposal as described in Section 3.4. Note that when a bin is split, it is possible to set the desired frequency of the new bins to some reduced value, say each obtaining half the desired frequency of the original—in fact in the numerical experiments we do exactly that. However, for notational and pedagogical simplicity, we present here the algorithm where the desired frequency of each bin is equal to $1/d_t$ at iteration t.

A.1 ALGORITHM PSEUDOCODE

Algorithm 3 Proposed Density Exploration Algorithm

Run a preliminary exploration of the target, e.g., using adaptive MCMC, and determine an energy range.
Partition the state space into d_0 regions $\{\mathcal{X}_{1,0}, \ldots, \mathcal{X}_{d_0,0}\}$ along a reaction coordinate $\xi(x)$, the default choice being $\xi(x) = -\log \pi(x)$.
$\forall i \in \{1, \ldots, d_0\}$ set $\theta(i) \leftarrow 1$, $\nu(i) \leftarrow 0$.
Choose an initial proposal standard deviation σ_0.
Choose the frequency τ with which to check for a flat histogram.
Choose a decreasing sequence $\{\rho_t\}$, typically $\rho_t = 1/t$, to update the proposal standard deviation.
Choose a decreasing sequence $\{\gamma_k\}$, typically $\gamma_k = 1/k$, to update the bias.
Sample $\boldsymbol{X}_0 \sim \pi_0$, an initial distribution.
for $t = 1$ to T **do**
    Sample $\boldsymbol{X}_t$ from $P_{\theta_{t-1}}(\boldsymbol{X}_{t-1}, \cdot)$, a transition kernel with invariant distribution $\prod_{n=1}^{N} \tilde{\pi}_{\theta_{t-1}}(x)$, parameterized by the proposal standard deviation σ_{t-1}.
    Update the proposal standard deviation: $\sigma_t \leftarrow \sigma_{t-1} + \rho_t (2\mathcal{I}(A > 0.234) - 1)$, where A is the last acceptance rate.
    Set $d_t \leftarrow d_{t-1}$.

Update the proportions: $\forall i \in \{1, \ldots, d_t\} \quad \nu(i) \leftarrow \frac{1}{t}\left[(t-1)\nu(i) + N^{-1}\sum_{j=1}^{N} \mathcal{I}_{\mathcal{X}_{i,t}}(X_t^{(j)})\right]$.
Every τth iteration, check the distribution of samples within each bin, extending the range if necessary. For example, if $\xi(x) = -\log \pi(x)$ and a new minimum value of $\xi(x)$ was found, extend the first bin to include this value.
for $i \in \{1, \ldots, d_t\}$ **do**
  if bin i should be split **then**
    Create two sub-bins covering bin i, assign to each a weight equal to $\theta_t(i)/2$.
    Set $d_t \leftarrow d_t + 1$, extend ν.
  end if
end for
if "flat histogram": $\max_{i \in \{1,\ldots,d_t\}} |\nu(i) - {d_t}^{-1}| < c/d$ **then**
  Set $k \leftarrow k + 1$.
  Reset $\forall i \in \{1, \ldots, d_t\} \quad \nu(i) \leftarrow 0$.
end if
Update the bias: $\log \theta_t(i) \leftarrow \log \theta_{t-1}(i) + \gamma_k(N^{-1}\sum_{j=1}^{N} \mathcal{I}_{\mathcal{X}_{i,t}}(X_t^{(j)}) - d_t^{-1})$.
Normalize the bias: $\theta_t(i) \leftarrow \theta_t(i)/\sum_{i=1}^{d_t} \theta_t(i)$.
end for

Discussion of algorithm convergence is provided as supplementary material.

SUPPLEMENTARY MATERIALS

Tri-modal example: Additional continuous, trimodal example, which demonstrates bin-splitting mechanism and further suggestions for algorithm tuning. (SupplementaryMaterials.pdf)

Algorithm convergence: Discussion of related algorithms as well as the algorithm's convergence. (Supplementary Materials.pdf)

R Package: Code and data to reproduce all results and test the method more generally. (PAWL.zip)

ACKNOWLEDGMENTS

Luke Bornn is supported by grants from NSERC and The Michael Smith Foundation for Health Research. Pierre Jacob is supported by a Ph.D. fellowship from the AXA Research Fund. The authors also acknowledge the use of R and `ggplot2` in their analyses (R Development Core Team 2010; Wickham 2009). Finally, the authors are thankful for helpful conversations with Yves Atchadé, François Caron, Nicolas Chopin, Faming Liang, Eric Moulines, and Christian Robert.

[Received September 2011. Revised June 2012.]

REFERENCES

Andrieu, C., and Thoms, J. (2008), "A Tutorial on Adaptive MCMC," *Statistics and Computing*, 18, 343–373. [752]

Atchadé, Y., Fort, G., Moulines, E., and Priouret, P. (2011), *Adaptive Markov Chain Monte Carlo: Theory and Methods* (chapter 2), Cambridge, UK: Cambridge University Press, pp. 33–53. [752,758]

Atchadé, Y., and Liu, J. (2010), "The Wang-Landau Algorithm in General State Spaces: Applications and Convergence Analysis," *Statistica Sinica*, 20, 209–233. [751,763]

Banfield, J., and Raftery, A. (1992), "Ice Floe Identification in Satellite Images Using Mathematical Morphology and Clustering About Principal Curves," *Journal of the American Statistical Association*, 87, 7–16. [767]

Baragatti, M., Grimaud, A., and Pommeret, D. (2012), "Parallel Tempering With Equi-Energy Moves," *Statistics and Computing*, 1–17. [752,763]

Besag, J., and Green, P. (1993), "Spatial Statistics and Bayesian Computation," *Journal of the Royal Statistical Society,* Series B, 55, 25–37. [752]

Brockwell, A. (2006), "Parallel Markov Chain Monte Carlo Simulation by Pre-Fetching," *Journal of Computational and Graphical Statistics*, 15, 246–261. [758]

Brockwell, A., Del Moral, P., and Doucet, A. (2010), "Sequentially Interacting Markov Chain Monte Carlo Methods," *The Annals of Statistics*, 38, 3387–3411. [758]

Byrd, J. (2010), "Parallel Markov Chain Monte Carlo," unpublished Ph.D. thesis, University of Warwick. [758]

Casarin, R., Craiu, R., and Leisen, F. (2011), "Interacting Multiple Try Algorithms With Different Proposal Distributions," *Statistics and Computing*, 1–16. [757]

Chopin, N., and Jacob, P. (2010), "Free Energy Sequential Monte Carlo, Application to Mixture Modelling," in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford: Oxford University Press, 91–118. [752,760,763]

Chopin, N., Lelievre, T., and Stoltz, G. (2012), "Free Energy Methods for Bayesian Statistics: Efficient Exploration of Univariate Gaussian Mixture Posteriors," *Statistics and Computing*, 22, 897–916. [752,763]

Craiu, R. V., Rosenthal, J., and Yang, C. (2009), "Learn From Thy Neighbor: Parallel-Chain and Regional Adaptive MCMC," *Journal of the American Statistical Association*, 104, 1454–1466. [751]

Del Moral, P., Doucet, A., and Jasra, A. (2006), "Sequential Monte Carlo Samplers," *Journal of the Royal Statistical Society,* Series B, 68, 411–436. [751]

Diebolt, J., and Robert, C. P. (1994), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society,* Series B, 56, 363–375. [763]

Edwards, R., and Sokal, A. (1988), "Generalization of the Fortuin-Kasteleyn-Swendsen-Wang Representation and Monte Carlo Algorithm," *Physical Review D*, 38, 2009–2012. [752]

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, New York: Springer Verlag. [763]

Geyer, C. (1991), "Markov Chain Monte Carlo Maximum Likelihood," in *Proceedings of the 23rd Symposium on the Interface*, ed. E. Keramigas, Fairfax, VA: Interface Foundations, pp. 156–163. [751]

Guan, Y., and Krone, S. (2007), "Small-World MCMC and Convergence to Multi-Modal Distributions: From Slow Mixing to Fast Mixing," *The Annals of Applied Probability*, 17, 284–304. [759]

Haario, H., Saksman, E., and Tamminen, J. (2001), "An Adaptive Metropolis Algorithm," *Bernoulli*, 7, 223–242. [752]

Higdon, D. (1998), "Auxiliary Variable Methods for Markov Chain Monte Carlo With Applications," *Journal of the American Statistical Association*, 93, 585–595. [752,768]

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005), "MCMC and the Label Switching Problem in Bayesian Mixture Models," *Statistical Science*, 20, 50–67. [763]

Jasra, A., Stephens, D., and Holmes, C. (2007), "On Population-Based Simulation for Static Inference," *Statistics and Computing*, 17, 263–279. [763,764,766]

Kou, S., Zhou, Q., and Wong, W. (2006), "Equi-Energy Sampler With Applications in Statistical Inference and Statistical Mechanics," *The Annals of Statistics*, 34, 1581–1619. [751,752,757]

Lee, A., Yau, C., Giles, M., Doucet, A., and Holmes, C. (2010), "On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods," *Journal of Computational and Graphical Statistics*, 19, 769–789. [758]

Liang, F. (2005), "A Generalized Wang-Landau Algorithm for Monte Carlo Computation," *Journal of the American Statistical Association*, 100, 1311–1327. [751]

——— (2009), "Improving SAMC Using Smoothing Methods: Theory and Applications to Bayesian Model Selection Problems," *The Annals of Statistics*, 37, 2626–2654. [758]

Liang, F., Liu, C., and Carroll, R. (2007), "Stochastic Approximation in Monte Carlo Computation," *Journal of the American Statistical Association*, 102, 305–320. [758]

Liang, F., and Wu, M. (2011), "Population Stochastic Approximation MCMC Algorithm and its Weak Convergence," Technical Report 2011-201, Texas A& M University. [758,764]

Marin, J., and Robert, C. (2007), *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, New York: Springer-Verlag. [761]

Marinari, E., and Parisi, G. (1992), "Simulated Tempering: A New Monte Carlo Scheme," *EPL (Europhysics Letters)*, 19, 451–458. [751]

McDonald, G., and Schwing, R. (1973), "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics*, 15, 463–481. [760]

Neal, R. (2001), "Annealed Importance Sampling," *Statistics and Computing*, 11, 125–139. [751]

——— (2003), "Slice Sampling," *The Annals of Statistics*, 31, 705–741. [752]

Niederreiter, H. (1992), *Random Number Generation and Quasi-Monte Carlo Methods*, Philadelphia, PA: Society for Industrial Mathematics. [756]

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [771]

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components" (with discussion), *Journal of the Royal Statistical Society,* Series B, 59, 731–792. [763]

Robert, C., and Casella, G. (2004), *Monte Carlo Statistical Methods*, New York: Springer. [750]

Roberts, G., Gelman, A., and Gilks, W. (1997), "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms," *The Annals of Applied Probability*, 7, 110–120. [759]

Roberts, G., and Rosenthal, J. (2009), "Examples of Adaptive MCMC," *Journal of Computational and Graphical Statistics*, 18, 349–367. [759]

Schmidler, S. (2011), "Exploration vs. Exploitation in Adaptive MCMC," *Adap'ski Invited Presentation*. Available at *http://www.maths.bris.ac.uk/∼maxca/adapsklll/schmidler.html* [750,756,760]

Stephens, M. (2000), "Dealing With Label Switching in Mixture Models," *Journal of the Royal Statistical Society,* Series B, 62, 795–809. [763]

Suchard, M., and Rambaut, A. (2009), "Many-Core Algorithms for Statistical Phylogenetics," *Bioinformatics*, 25, 1370–1376. [758]

Swendsen, R., and Wang, J. (1986), "Replica Monte Carlo Simulation of Spin-Glasses," *Physical Review Letters*, 57, 2607–2609. [751]

——— (1987), "Nonuniversal Critical Dynamics in Monte Carlo Simulations," *Physical Review Letters*, 58, 86–88. [752]

Wang, F., and Landau, D. (2001a), "Determining the Density of States for Classical Statistical Models: A Random Walk Algorithm to Produce a Flat Histogram," *Physical Review E*, 64, 56101. [751,752,755]

——— (2001b), "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States," *Physical Review Letters*, 86, 2050–2053. [751,752,755]

Wei, W., Erenrich, J., and Selman, B. (2004), "Towards Efficient Sampling: Exploiting Random Walk Strategies," in *Proceedings of the National Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press, pp. 670–676. [753]

Welford, B. (1962), "Note on a Method for Calculating Corrected Sums of Squares and Products," *Technometrics*, 4, 419–420. [759]

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer. [771]

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With G-Prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno do Dinetti*, eds. P. Goel and A. Zellner, Amsterdam: North-Holland, pp. 233–243. [761]