Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

An Adaptive Exchange Algorithm for Sampling from Distributions with Intractable Normalizing Constants

Faming Liang, Ick Hoon Jin, Qifan Song & Jun S. Liu

To cite this article: Faming Liang, Ick Hoon Jin, Qifan Song & Jun S. Liu (2015): An Adaptive Exchange Algorithm for Sampling from Distributions with Intractable Normalizing Constants, Journal of the American Statistical Association, DOI: 10.1080/01621459.2015.1009072

To link to this article: http://dx.doi.org/10.1080/01621459.2015.1009072

An Adaptive Exchange Algorithm for Sampling from Distributions with Intractable Normalizing Constants

Faming Liang*, Ick Hoon Jin, Qifan Song, and Jun S. Liu

Abstract

Sampling from the posterior distribution for a model whose normalizing constant is intractable is a long-standing problem in statistical research. We propose a new algorithm, adaptive auxiliary variable exchange algorithm, or in short, adaptive exchange (AEX) algorithm, to tackle this problem. The new algorithm can be viewed as a MCMC extension of the exchange algorithm (Murray, Ghahramani and MacKay, 2006), which generates auxiliary variables via an importance sampling procedure from a Markov chain running in parallel. The convergence of the algorithm is established under mild conditions. Compared to the exchange algorithm, the new algorithm removes the requirement that the auxiliary variables must be drawn using a perfect sampler, and thus can be applied to many models for which the perfect sampler is not available or very expensive. Compared to the approximate exchange algorithms, such as the double Metropolis-Hastings sampler (Liang, 2010), the new algorithm overcomes their theoretical difficulty in convergence. The new algorithm is tested on the spatial autologistic and autonormal models. The numerical results indicate that the new algorithm is particularly useful for the problems for which the underlying system is strongly dependent.

Keywords: Autologistic Model; Exchange Algorithm; Intractable Normalizing Constant; Stochastic Approximation Monte Carlo.

*To whom correspondence should be addressed. Liang is Professor, Department of Biostatistics, University of Florida, Gainesville, FL 32611. Email: faliang@ufl.edu. Jin is Research Scientist, Center for Biostatistics, The Ohio State University, Columbus, Ohio 43221. Song is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47906. Liu is Professor, Department of Statistics, Harvard University, Science Center 715, 1 Oxford Street, Cambridge, MA 02138. Liang's research was partially supported by grants from the National Science Foundation DMS-1106494 and DMS-1317131. The authors thank the editor, associate editor and two referees for their constructive comments which have led to significant improvement of this paper.

1 Introduction

Sampling from the posterior distribution for a model whose normalizing constant is intractable is a long-standing problem in statistical research. Formally, this problem can be posed as follows: Suppose that we have a dataset $\boldsymbol{Y}$ generated from a model with the likelihood function given by

$$f(\boldsymbol{y}|\theta) = \frac{\varphi(\boldsymbol{y}, \theta)}{\kappa(\theta)}, \quad \boldsymbol{y} \in \mathcal{X},\ \theta = \Theta, \tag{1}$$

where θ denotes the parameter vector of the model, and $\kappa(\theta)$ is the normalizing constant which depends on θ and is unavailable in closed form. Examples of such models include the Ising and Potts models used in image analysis (Hurn et al., 2003), the autologistic and autonormal models used in spatial data analysis (Besag, 1974), and exponential random graph models used in social network analysis (see, e.g., Snijders et al., 2006), among others. Let $\pi(\theta)$ denote the prior density of θ. The posterior distribution of θ is then given by

$$\pi(\theta|\boldsymbol{y}) \propto \frac{1}{\kappa(\theta)}\varphi(\boldsymbol{y}, \theta)\pi(\theta). \tag{2}$$

Because $\kappa(\theta)$ is intractable, sampling from $\pi(\theta|\boldsymbol{y})$ has posed a great challenge on existing statistical methods.

It is known that the Metropolis-Hastings (MH) algorithm cannot be directly applied to sample from $\pi(\theta|\boldsymbol{y})$, as its acceptance probability would involve an unknown normalizing constant ratio $\kappa(\theta)/\kappa(\theta')$, where θ' denotes the proposed value. To tackle this problem, various methods have been proposed in the literature. These methods can be classified into two categories according to the strategies employed by them, the approximation-based methods and the auxiliary variable-based methods.

The methods in the first category are to approximate the likelihood function, the normalizing constant $\kappa(\theta)$, or the normalizing constant ratio $\kappa(\theta)/\kappa(\theta')$ using various approaches. Besag (1974) proposed the so-called maximum pseudo-likelihood estimator (MPLE) method, in which the likelihood function is approximated by a product of a series of conditional likelihood functions. Since

this approximation ignores certain dependence within the components of $\boldsymbol{y}$, the performance of MPLE is often unsatisfactory, especially when the dependence between the components of $\boldsymbol{y}$ is strong. Geyer and Thompson (1992) proposed an importance sampling-based method to approximate $\kappa(\theta)$. Let $\theta^{(0)}$ be an initial estimate of θ and let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m$ denote a set of random samples drawn from $f(\boldsymbol{x}|\theta^{(0)})$, which can be obtained via Markov chain Monte Carlo (MCMC) simulations. The log-likelihood function can then be approximated by

$$\log f_m(\boldsymbol{y}|\theta) = \log \varphi(\boldsymbol{y}, \theta) - \log\left(\frac{1}{m}\sum_{i=1}^{m} \varphi(\boldsymbol{x}_i, \theta)/\varphi(\boldsymbol{x}_i, \theta^{(0)})\right), \tag{3}$$

which approaches $\log f(\boldsymbol{y}|\theta)$ as $m \to \infty$, and the resulting estimator $\widehat{\theta} = \arg\max_\theta \log f_m(\boldsymbol{y}|\theta)$ is called the Monte Carlo maximum likelihood estimator (MCMLE). The performance of this method depends on the choice of $\theta^{(0)}$: If $\theta^{(0)}$ is near the true MLE, it can produce a good estimate of θ; otherwise, it may converge to a suboptimal solution. Liang (2007) proposed an alternative Monte Carlo method to approximate $\kappa(\theta)$, where $\kappa(\theta)$ is viewed as a marginal density function of the unnormalized distribution $\varphi(\boldsymbol{y}, \theta)$ and estimated using an adaptive kernel smoothing method with Monte Carlo draws. A similar method for approximating $\kappa(\theta)$ can be found in Atchade et al. (2013). Toward sampling from the posterior $\pi(\theta|\boldsymbol{y})$, Liang and Jin (2013) proposed the so-called Monte Carlo MH (MCMH) algorithm, which is to approximate the normalizing constant ratio $\kappa(\theta)/\kappa(\theta')$ at each iteration using samples simulated from either $f(\boldsymbol{x}|\theta)$ or $f(\boldsymbol{x}|\theta')$ through a finite run of Markov chain. Since the convergence of a finite run of Markov chain cannot be guaranteed, the algorithm is only approximately correct, at least, in theory. The Bayesian SAMC algorithm suggested in Jin and Liang (2014) and the marginal PMCMC algorithm suggested in Everitt (2012) suffer from a similar problem, a large number of samples needs to be simulated at each iteration to ensure its convergence.

The methods in the second category aim to have the normalizing constant ratio $\kappa(\theta)/\kappa(\theta')$ canceled in simulations by augmenting appropriate auxiliary variables to the target distribution and/or the proposal distribution. Along this direction, Møller et al. (2006) proposed an algorithm which

arguments both the target and proposal distributions, and Murray et al. (2006) proposed the so-called exchange algorithm which arguments only the proposal distribution. These two algorithms are usually termed as auxiliary variable MCMC algorithms in the literature. Although the underlying idea is very attractive, these algorithms require the auxiliary variables to be drawn using a perfect sampler (Propp and Wilson, 1996). Since perfect sampling can be very expensive or impossible for many models with intractable normalizing constants, the applications of these algorithms are highly hindered. To address this issue, Liang (2010) proposed the double Metropolis-Hastings (DMH) sampler, where each auxiliary variable is drawn through a short run of the MH algorithm initialized with the observation $\boldsymbol{y}$. As noted in Liang (2010), initializing the auxiliary MH chain with the observation $\boldsymbol{y}$ leads to improved convergence of the algorithm. Similar algorithms have been applied to social network analysis by Caimo and Friel (2011) and Everitt (2012). Since, in these algorithms, a finite MCMC run has to be used for generating auxiliary samples at each iteration, the resulting estimates are only approximately correct no matter how long these algorithms are run. A brief review of the exchange algorithm and the DMH algorithm is given in Section 2.

In this paper, we propose a new algorithm, the adaptive auxiliary variable exchange algorithm, or in short, the adaptive exchange (AEX) algorithm, for sampling from the posterior $\pi(\theta|\boldsymbol{y})$. AEX is an adaptive Monte Carlo version of the exchange algorithm, where the auxiliary variables are generated via an importance sampling procedure from a Markov chain running in parallel. The convergence of the algorithm is established under mild conditions. Compared to the existing auxiliary variable MCMC algorithms, AEX removes the requirement of perfect sampling and thus can be applied to many models for which perfect sampling is not available or very expensive. Compared to the DMH sampler, AEX overcomes its theoretical difficulty caused by inconvergence of finite MCMC runs. The new algorithm is tested on spatial autologistic models and autonormal models. The numerical results indicate that the new algorithm is particularly useful for the problems for which the underlying system is strongly dependent.

The remainder of this paper is organized as follows. In Section 2, we describe the AEX algorithm. In Section 3, we present some theoretical results on the convergence of AEX. In Section 4, we test AEX on a spatial autologistic model along with extensive comparisons with the exchange algorithm. In Section 5, we conclude the paper with a brief discussion.

2 The Adaptive Exchange Algorithm

In this section, we first give a brief review for the exchange and approximate exchange algorithms, and then describe the adaptive exchange algorithm.

2.1 The Exchange and Approximate Exchange Algorithms

Let θ_t denote the draw of θ at iteration t. One iteration of the exchange algorithm consists of the following steps:

Exchange Algorithm

1. Propose a candidate point θ' from a proposal distribution denoted by $q(\theta'|\theta_t)$.
2. Generate an auxiliary variable $\boldsymbol{x} \sim f(\boldsymbol{x}|\theta')$ using a perfect sampler (Propp and Wilson, 1996).
3. Set $\theta_{t+1} = \theta'$ with probability

(4)

and set $\theta_{t+1} = \theta_t$ with the remaining probability.

This algorithm is called the exchange algorithm because of the similarity of (4) with the acceptance probability of the swapping operation of exchange Monte Carlo (Geyer, 1991; Hukushima and Nemoto, 1996). The exchange algorithm is different from the conventional MH algorithm at that the proposal $q(\theta'|\theta_t)f(\boldsymbol{x}|\theta')$ consists of a randomization component which involves a random

draw of $\boldsymbol{x}$. To see why the exchange algorithm works, we define $s(\theta, \boldsymbol{x}, \theta') = \alpha(\theta, \boldsymbol{x}, \theta')q(\theta'|\theta)f(\boldsymbol{x}|\theta')$, and then it is easy to see that

$$\pi(\theta|\boldsymbol{y})\int_{\mathcal{X}} s(\theta, \boldsymbol{x}, \theta')d\boldsymbol{x} = \frac{1}{f(\boldsymbol{y})}\int_{\mathcal{X}} \min\{\pi(\theta')f(\boldsymbol{y}|\theta')q(\theta|\theta')f(\boldsymbol{x}|\theta), \pi(\theta)f(\boldsymbol{y}|\theta)q(\theta'|\theta)f(\boldsymbol{x}|\theta')\}d\boldsymbol{x},$$

which is symmetric about θ and θ'. This implies that

$$\pi(\theta|\boldsymbol{y})P(\theta, d\theta') = \pi(\theta'|\boldsymbol{y})P(\theta', d\theta),$$

where $P(\theta, d\theta')$ denotes the Markov transition kernel of the exchange algorithm; that is,

$$P(\theta, d\theta') = \int_{\mathcal{X}} s(\theta, \boldsymbol{x}, d\theta')d\boldsymbol{x} + \delta_\theta(d\theta')[1 - \int_{\Theta\times\mathcal{X}} s(\theta, \boldsymbol{x}, d\vartheta)d\boldsymbol{x}]. \tag{5}$$

Therefore, the exchange algorithm defines a valid Markov chain for simulating from $\pi(\theta|\boldsymbol{y})$.

As aforementioned, to ease sampling of auxiliary variables, the DMH sampler (Liang, 2010) proposed to draw the auxiliary variable $\boldsymbol{x}$ through a finite run of the MH algorithm initialized at the observation $\boldsymbol{y}$. Under the assumption of equilibrium, the acceptance probability for the candidate point θ' is reduced to (4). However, since the equilibrium can only be approximately reached for a large number of iterations, the DMH sampler is only approximately correct. It is worth mentioning that DMH can generally perform much better than its competitor, MPLE, in parameter estimation.

2.2 The Adaptive Exchange Algorithm

To overcome the theoretical difficulty of DMH in convergence, we propose the adaptive exchange (AEX) algorithm. The basic idea of AEX can be loosely described as follows: AEX consists of two chains running in parallel. The first chain is auxiliary, which is run in the data space $\mathcal{X}$ ($\boldsymbol{y} \in \mathcal{X}$) and aims to draw samples from a family of distributions $f(\boldsymbol{x}|\theta^{(1)}), \ldots, f(\boldsymbol{x}|\theta^{(m)})$ defined on a set of pre-specified parameter values $\theta^{(1)}, \ldots, \theta^{(m)}$. The second chain is the target chain, which makes use of the auxiliary chain and aims to draw samples from the target posterior $\pi(\theta|\boldsymbol{y})$. The target chain is run in the parameter space Θ ($\theta \in \Theta$). For a candidate point θ', an auxiliary variable $\boldsymbol{x}$

is resampled from the past samples of the auxiliary chain via an importance sampling procedure. Here we assume that the neighboring distributions $f(\boldsymbol{x}|\theta^{(i)})$'s satisfy the following condition:

(A_0) The sample spaces of neighboring $f(\boldsymbol{x}|\theta^{(i)})$'s have a reasonable overlap and the parameters $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$ have covered the major part of the support of $\pi(\theta|\boldsymbol{y})$, e.g., $\int_{C_\theta} \pi(\theta|\boldsymbol{x})d\theta > 0.9999$, where C_θ denotes the convex hull formed by $\theta^{(1)}, \ldots, \theta^{(m)}$.

These assumptions ensure that the auxiliary chain can mix reasonably fast and thus the target chain can converge to the right posterior distribution $\pi(\theta|\boldsymbol{y})$ as the number of iterations becomes large. Actually, this is the key to the success of AEX. How to choose the auxiliary parameters $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$ will be described in Section 2.3.

To draw samples from the family of distributions $f(\boldsymbol{z}|\theta)$, $\theta \in \{\theta^{(1)}, \ldots, \theta^{(m)}\}$, we adopt the stochastic approximation Monte Carlo (SAMC) algorithm (Liang et al., 2007). SAMC ensures that each of the distributions, $f(\boldsymbol{z}|\theta^{(1)}), \ldots, f(\boldsymbol{z}|\theta^{(m)})$, can be drawn with a pre-specified frequency, while overcoming the local-trap problem that the simulation can get trapped at a single or few distributions. We note that some other MCMC algorithms, such as the reversible jump MCMC algorithm (Green, 1995) and evolutionary Monte Carlo (Liang and Wong, 2001), can also be used here to draw samples from the family of distributions, but they need to deal with the local-trap problem. When evolutionary Monte Carlo is used, $\theta^{(i)}$'s can be treated as different temperatures. To implement the SAMC algorithm, we define $\boldsymbol{p} = (p_1, \ldots, p_m)$ to be the desired sampling frequencies from respective distributions $f(\boldsymbol{z}|\theta^{(1)}), \ldots, f(\boldsymbol{z}|\theta^{(m)})$, where $0 < p_i < 1$ and $\sum_{i=1}^m p_i = 1$; and specify a positive, nonincreasing sequence $\{a_t\}$, the so-called gain factor sequence, which satisfies the condition:

(A_1) $\lim_{t\to\infty} a_t = 0, \quad \sum_{t=1}^{\infty} a_t = \infty, \quad \sum_{t=1}^{\infty} a_t^{\eta} < \infty$ for some $\eta \in (1, 2]$.

In this paper, we set $p_1 = \cdots = p_m = 1/m$ and

$$a_t = \frac{t_0}{\max(t_0, t)}, \quad t = 1, 2, \ldots, \tag{6}$$

for some known constant $t_0 > 1$. Let $w_t^{(i)}$ denote an abundance factor attached to the distribution $f(\boldsymbol{z}|\theta^{(i)})$ at iteration t, let $\boldsymbol{w}_t = (w_t^{(1)}, \ldots, w_t^{(m)})$, and let $\mathcal{W}$ denote the sample space of $\boldsymbol{w}_t$. Let $\{\mathcal{K}_s, s \geq 0\}$ be a sequence of compact subsets of $\mathcal{W}$ such that

$$\bigcup_{s\geq 0} \mathcal{K}_s = \mathcal{W}, \quad \text{and} \quad \mathcal{K}_s \subset \text{int}(\mathcal{K}_{s+1}), \quad s \geq 0, \tag{7}$$

where int(A) denotes the interior of set A. Let $\mathcal{X}_0$ be a subset of $\mathcal{X}$, and let $\mathbb{T} : \mathcal{X} \times \mathcal{W} \to \mathcal{X}_0 \times \mathcal{K}_0$ be a truncation function which is measurable and maps a point in $\mathcal{X} \times \mathcal{W}$ to a random point in $\mathcal{X}_0 \times \mathcal{K}_0$. Let σ_t denote the number of truncations performed until iteration t. Let $\boldsymbol{z}_t$ denote the samples generated by the auxiliary chain at iteration t, and let ϑ_t denote the parameter value associated with $\boldsymbol{z}_t$. Let S_t denote the set of auxiliary samples collected by iteration t. The AEX algorithm starts with a random point in $\mathcal{X}_0 \times \mathcal{K}_0$ and then iterates in the following steps:

Part 1: (Auxiliary Chain) Auxiliary Sample Collection via SAMC

1. (Sampling) Choose to update ϑ_t or $\boldsymbol{z}_t$ with pre-specified probabilities, e.g., 0.75 for updating ϑ_t and 0.25 for updating $\boldsymbol{z}_t$.

   (a) Update ϑ_t: draw ϑ' from the set $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$ according to a proposal distribution $T_1(\cdot|\vartheta_t)$, set $(\vartheta_{t+1}, \boldsymbol{z}_{t+1}) = (\vartheta', \boldsymbol{z}_t)$ with probability

   $$\min\left\{1, \frac{w_t^{(J(\vartheta_t))}}{w_t^{(J(\vartheta'))}} \frac{\varphi(\boldsymbol{z}_t, \vartheta')}{\varphi(\boldsymbol{z}_t, \vartheta_t)} \frac{T_1(\vartheta_t|\vartheta')}{T_1(\vartheta'|\vartheta_t)}\right\}, \tag{8}$$

   and set $(\vartheta_{t+1}, \boldsymbol{z}_{t+1}) = (\vartheta_t, \boldsymbol{z}_t)$ with the remaining probability, where $J(\vartheta_t)$ denotes the index of ϑ_t, i.e., $J(\vartheta_t) = j$ if $\vartheta_t = \theta^{(j)}$.

   (b) Update $\boldsymbol{z}_t$: draw $\boldsymbol{z}'$ according to a proposal distribution $T_2(\cdot|\boldsymbol{z}_t)$, set $(\boldsymbol{z}_{t+1}, \vartheta_{t+1}) = (\boldsymbol{z}', \vartheta_t)$ with probability

   $$\min\left\{1, \frac{\varphi(\boldsymbol{z}', \vartheta_t)}{\varphi(\boldsymbol{z}_t, \vartheta_t)} \frac{T_2(\boldsymbol{z}_t|\boldsymbol{z}')}{T_2(\boldsymbol{z}'|\boldsymbol{z}_t)}\right\},$$

   and set $(\boldsymbol{z}_{t+1}, \vartheta_{t+1}) = (\boldsymbol{z}_t, \vartheta_t)$ with the remaining probability.

2. (Abundance factor updating) Set

$$\log(w_{t+1/2}^{(i)}) = \log(w_t^{(i)}) + a_{t+1}(e_{t+1,i} - p_i), \quad i = 1, 2, \ldots, m,$$

where $e_{t+1,i} = 1$ if $\vartheta_{t+1} = \theta^{(i)}$ and 0 otherwise. If $\boldsymbol{w}_{t+1/2} \in \mathcal{K}_{\sigma_t}$, then set $(\boldsymbol{w}_{t+1}, \boldsymbol{z}_{t+1}) = (\boldsymbol{w}_{t+1/2}, \boldsymbol{z}_{t+1})$ and $\sigma_{t+1} = \sigma_t$; otherwise, set $(\boldsymbol{w}_{t+1}, \boldsymbol{z}_{t+1}) = \mathbb{T}(\boldsymbol{w}_t, \boldsymbol{z}_t)$ and $\sigma_{t+1} = \sigma_t + 1$.

3. (Auxiliary sample collection) Append the sample $(\boldsymbol{z}_{t+1}, \vartheta_{t+1}, w_{t+1}^{(J(\vartheta_{t+1}))})$ to the collection S_t. Denote the new collection by S_{t+1}, i.e., set $S_{t+1} = S_t \cup \{(\boldsymbol{z}_{t+1}, \vartheta_{t+1}, w_{t+1}^{(J(\vartheta_{t+1}))})\}$.

Part 2: (Target Chain) Adaptive Exchange

4. (Proposal) Propose a candidate point θ' from a proposal distribution $q(\theta'|\theta_t)$.

5. (Resampling) Resample an auxiliary variable $\boldsymbol{x}$ from the collection S_{t+1} via a dynamic importance sampling procedure; that is, setting $\boldsymbol{x} = \boldsymbol{z}_i$ with probability

$$P(\boldsymbol{x} = \boldsymbol{z}_i) = \frac{\sum_{j=1}^{|S_{t+1}|} w_j^{(J(\vartheta_j))} \varphi(\boldsymbol{z}_j, \theta') / \varphi(\boldsymbol{z}_j, \vartheta_j) I(\boldsymbol{z}_j = \boldsymbol{z}_i)}{\sum_{j=1}^{|S_{t+1}|} w_j^{(J(\vartheta_j))} \varphi(\boldsymbol{z}_j, \theta') / \varphi(\boldsymbol{z}_j, \vartheta_j)} \tag{9}$$

where $(\boldsymbol{z}_j, \vartheta_j, w_j^{(J(\vartheta_j))})$ denotes the j-th element of the set S_{t+1}, and $|S_{t+1}|$ denotes the size of S_{t+1}.

6. (Exchange) Set $\theta_{t+1} = \theta'$ with the probability

$$\alpha(\theta_t, \boldsymbol{x}, \theta') = \min\left\{1, \frac{\pi(\theta')\varphi(\boldsymbol{y}, \theta')}{\pi(\theta_t)\varphi(\boldsymbol{y}, \theta_t)} \frac{q(\theta_t|\theta')}{q(\theta'|\theta_t)} \frac{\varphi(\boldsymbol{x}, \theta_t)}{\varphi(\boldsymbol{x}, \theta')}\right\}, \tag{10}$$

and set $\theta_{t+1} = \theta_t$ with probability $1 - \alpha(\theta_t, \boldsymbol{x}, \theta')$.

For this algorithm, we have a few remarks:

- Part I of the algorithm is to use the SAMC algorithm to draw samples from the mixture distribution

$$f(\boldsymbol{z}|\theta^{(1)},\ldots,\theta^{(m)}) = \sum_{i=1}^{m} p_i f(\boldsymbol{z}|\theta^{(i)}) = \sum_{i=1}^{m} \frac{p_i}{\kappa(\theta^{(i)})}\psi(\boldsymbol{z},\theta^{(i)}),$$

meanwhile providing estimates for the normalizing constants $\kappa(\theta^{(1)}),\ldots,\kappa(\theta^{(m)})$. As shown in (14), $\boldsymbol{w}_t$ will converge to $(\kappa(\theta^{(1)})/p_1,\ldots,\kappa(\theta^{(m)})/p_m)$ (up to a multiplication factor) almost surely as $t \to \infty$. Part II is essential to run the exchange algorithm for simulation of the posterior $\pi(\theta|\boldsymbol{y})$, where the auxiliary variable is drawn via a dynamic importance sampling procedure. The dynamic importance function used at step t is proportional to

$$\sum_{i=1}^{m} \psi(z,\vartheta_t)/w_t^{(i)} I(\vartheta_t = \theta^{(i)}),$$

where $I(\cdot)$ is the indicator function. The validity of this procedure has been establish in Lemma 3.1. Since the underlying true proposal distribution for generating auxiliary variables in part II is changing from iteration to iteration, the new algorithm falls into the class of adaptive MCMC algorithms (for which the proposal distribution is changing from iteration to iteration). For this reason, we call the new algorithm the adaptive exchange algorithm.

- To prepare a good collection of auxiliary samples, one may first run the auxiliary chain for a large number of iterations, and then run the auxiliary and target chains in parallel. Certainly, this will improve the convergence of the target chain.

- The proposal $q(\theta'|\theta_t)$ can depend on $\boldsymbol{y}$; that is, it can be written in the form $q(\theta'|\theta_t,\boldsymbol{y})$. For simplicity, we notationally depress the dependence of the proposal on $\boldsymbol{y}$.

- On the choice of $\{a_t\}$ and convergence of AEX. In this paper, we set the gain factor in the form (6) with a free parameter t_0. As discussed in Liang et al. (2007), a large value of t_0 will force the sampler to reach all distributions $f(\boldsymbol{z}|\theta^{(i)})$'s quickly. Therefore, t_0 should be

set to a large number for a complex problem. In this paper, t_0 is set to 20,000 for the U.S. cancer mortality data example and 25,000 for all other examples. In general, the choice of t_0 should be associated with the choice of N, the total number of iterations of a run. The appropriateness of their choices can be diagnosed by checking the convergences of the auxiliary and target chains. The convergence of the auxiliary chain can be checked through an examination for the realized sampling frequencies $(\hat{p}_1, \ldots, \hat{p}_m)$, where $\hat{p}_i$ denotes the realized sampling frequency from the distribution $f(\boldsymbol{z}|\theta^{(i)})$. If $(\hat{p}_1, \ldots, \hat{p}_m)$ is not close to $(p_1, \ldots, p_m)$, the auxiliary chain should be diagnosed as non-convergent. In this case, the algorithm should be re-run with a larger value of N or a larger value of t_0 or both. Note that for the convergence diagnosis of the auxiliary chain, multiple runs are not necessary under the scenario considered in the paper, as it is known that each of the distributions $f(\boldsymbol{z}|\theta^{(i)})$'s is valid. However, to check the convergence of the target chain, multiple runs are still necessary.

2.3 Choice of Auxiliary Parameters

To choose the auxiliary parameters $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$, we suggest a fractional DMH algorithm-based procedure. Let θ_t denote the draw of θ at iteration t. One iteration of the fractional DMH algorithm can be described as follows:

Fractional DMH Algorithm

1. Propose a candidate point θ' from a proposal distribution denoted by $q(\theta'|\theta_t)$.

2. Generate an auxiliary variable $\boldsymbol{x}$ through a finite run of the MH algorithm which admits $f(\boldsymbol{x}|\theta')$ as the invariant distribution and is initialized at the observation $\boldsymbol{y}$.

3. Calculate the DMH ratio

$$r(\theta_t, \boldsymbol{x}, \theta') = \frac{\pi(\theta') f(\boldsymbol{y}|\theta')}{\pi(\theta_t) f(\boldsymbol{y}|\theta_t)} \cdot \frac{q(\theta_t|\theta')}{q(\theta'|\theta_t)} \cdot \frac{f(\boldsymbol{x}|\theta_t)}{f(\boldsymbol{x}|\theta')}.$$

4. Set $\theta_{t+1} = \theta'$ with probability $\alpha(\theta_t, \boldsymbol{x}, \theta') = \min\left\{1, [r(\theta_t, \boldsymbol{x}, \theta')]^{\zeta}\right\}$ and set $\theta_{t+1} = \theta_t$ with the remaining probability, where $0 < \zeta \leq 1$ is a pre-specified value.

If $\zeta = 1$, the algorithm is reduced to the DMH algorithm. If $0 < \zeta < 1$, the algorithm can result in an expanded support of the true posterior, as taking a fraction of $r(\theta_t, \boldsymbol{x}, \theta')$ reduces the rejection rate and thus can accept some samples with low posterior density values. Hence, when ζ is small, the samples $\{\theta_t, t = 1, 2, \ldots\}$ generated by this algorithm are expected to cover the major part of the support of the posterior $\pi(\theta|\boldsymbol{y})$. In this paper, we set $\zeta = 0.5$ for all examples.

Let $\{\theta_1, \ldots, \theta_n\}$ denote a set of samples generated by the fractional DMH algorithm. Then the auxiliary parameters $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$ can be selected from $\{\theta_1, \ldots, \theta_n\}$ in the following procedure.

Max-Min Procedure

1. Standardize each dimension of θ_t to the interval $[0, 1]$ by setting $\tilde{\theta}_{t,i} = (\theta_{t,i} - \min_j \theta_{t,j})/(\max \theta_{t,j} - \min \theta_{t,i})$ for $i = 1, \ldots, d$, where $\theta_{t,i}$ denotes the ith element of θ_t and d is the dimension of θ_t.

2. Calculate the distance matrix $D = (d_{st})$ of the standardized samples for $s = 1, \ldots, n$ and $t = 1, \ldots, n$, where $d_{st} = \sqrt{\sum_{i=1}^{d} (\tilde{\theta}_{si} - \tilde{\theta}_{ti})^2}$ denotes the distance between $\tilde{\theta}_s$ and $\tilde{\theta}_t$.

3. Randomly select one sample, say θ_k, as $\theta^{(1)}$. Set $A = \{k\}$, $A^c = \{1, 2, \ldots, k-1, k+1, \ldots, n\}$ and $j = 2$.

4. Select the next auxiliary parameter point $\theta^{(j)} = \theta_l$ such that the following conditions are satisfied: $l \in A^c$ and there exists $l' \in A$ such that $d_{ll'} = \max_{i \in A^c} \min_{s \in A} d_{is}$. Set $A = A \cup \{l\}$ and $A^c = A^c \setminus \{l\}$.

5. Repeat step 4 for $j = 3, \ldots, m$.

The standardization in step 1 ensures that each dimension of θ contributes equally to the distance function d_{st}, and thus SAMC can mix equally in all dimensions of θ. The *max-min* rule used

in step 4 ensures that the selected samples $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$ and the full samples $\{\theta_t, t = 1, \ldots, n\}$ have about the same convex hull when m is reasonably large, and thus $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$ are distributed over the support of $\{\theta_1, \ldots, \theta_n\}$, i.e., condition (A_0) is satisfied. Note that in the max-min rule, $\theta^{(2)}$ is always the furthest one from $\theta^{(1)}$ among the samples $\{\theta_1, \ldots, \theta_n\}$.

The number m can be determined by trial-and-error; that is, choosing m such that the auxiliary chain can converge reasonably fast, e.g., within $10^3 m \sim 10^4 m$ iterations. How to diagnose the convergence of the auxiliary chain has been discussed at the end of Section 2.2. Besides the auxiliary parameters and the gain factor sequence, the convergence of the auxiliary chain can depend on the proposal distributions used in simulations. In our simulations, we usually set $T_2(\cdot|\cdot)$ as the Gibbs sampler for updating auxiliary variables, and set $T_1(\cdot|\cdot)$ as the uniform distribution over a pre-specified nearest neighbor for each $\theta^{(i)}$. In this paper, we set $m = 100$ and set the neighboring size $m_0 = 10$ in all simulations. Different settings have been tried, such as $m = 50$ and 200 and $m_0 = 5$, the results are similar. Since the auxiliary parameters are essentially generated from the same posterior distribution, the neighboring distributions $f(\boldsymbol{y}|\theta^{(i)})$'s are always reasonably overlapped. Note that they all share, at least approximately, the same sample—the observation $\boldsymbol{y}$. This ensures a smooth transition of the auxiliary chain between different auxiliary parameters. More importantly, this property holds independently of the dimension of θ. Hence, AEX can work well for the problems for which the dimension of θ is high. Here we note that when choosing the values of m and m_0, the multimodality of the posterior $\pi(\theta|\boldsymbol{y})$ needs to be checked. If multiple modes exist, special cares need to be taken in choosing m and m_0 such that the resulting auxiliary chain is irreducible.

Finally, we note that there can be many variations for the auxiliary parameter selection procedure proposed above. For example, the fractional DMH algorithm can be replaced by any other algorithm that can result in an expanded support of the target posterior $\pi(\theta|\boldsymbol{y})$, e.g., the approximate Bayesian computation (ABC) algorithm (Beaumont et al., 2002). For the models for which the dimension of θ is low, say, the spatial autologistic model studied in Section 4, $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$ can be

set to some grid points around the MPLE of θ. In the *max-min* procedure, the standardization can be replaced by a regular normalization procedure, i.e., setting $\tilde{\theta}_i = \Sigma_n^{-1/2}(\theta_i - \overline{\theta}_n)$, where $\overline{\theta}_n$ and Σ_n denote, respectively, the sample mean and covariance matrix of $\{\theta_1, \ldots, \theta_n\}$.

3 Convergence of the Adaptive Exchange Algorithm

As aforementioned, AEX falls into the class of adaptive MCMC algorithms. Since the transition kernel of the exchange step may admit different stationary distributions for different iterations, the convergence theory developed for conventional adaptive MCMC algorithms, see e.g., Roberts and Rosenthal (2007) and Andrieu and Moulines (2006), is not applicable here. The ergodicity theory developed by Fort et al. (2011) for adaptive MCMC allows to change stationary distributions at different iterations and is indeed applicable to AEX. However, the strong law of large numbers established therein requires some strong conditions that cannot be verified for AEX. For this reason, we develop some theory for adaptive MCMC, including ergodicity and weak law of large numbers (WLLN), which are applicable to AEX.

The remainder of this section is organized as follows. In Section 3.1, we prove two theorems, ergodicity and weak law of large numbers, for general adaptive MCMC algorithms with changing stationary distributions. In Section 3.2, we consider the weak convergence of auxiliary variables drawn from the auxiliary chain. In Section 3.3, we establish the ergodicity and weak law of large numbers for AEX.

3.1 Convergence of Adaptive MCMC with Changing Stationary Distributions

To facilitate our study, we first define, following Roberts and Rosenthal (2007), some notations for adaptive Markov chains. Consider a state space $(\mathbb{X}, \mathcal{F})$, where $\mathcal{F} = \mathcal{B}(\mathbb{X})$ denotes the Borel set

defined on $\mathbb{X}$. Let $X_t \in \mathbb{X}$ denote the state of the Markov chain at iteration t, and let P_{γ_t} denote the transition kernel at iteration t, where γ_t is a realization of a $\mathbb{Y}$-valued random variable Γ_t. In simulations, γ_t is updated according to specified rules. Let $\mathcal{G}_t = \sigma(X_0, \ldots, X_t, \Gamma_0, \ldots, \Gamma_t)$ be the filtration generated by $\{(X_i, \Gamma_i)\}_{i=0}^t$. Thus,

$$P(X_{t+1} \in B | X_t = x, \Gamma_t = \gamma, \mathcal{G}_{t-1}) = P_\gamma(x, B), \quad x \in \mathbb{X}, \gamma \in \mathbb{Y}, B \in \mathcal{F}.$$

Let $P_\gamma^t(x, B) = P_\gamma(X_t \in B | X_0 = x)$ denote the t-step transition probability for the Markov chain with the fixed transition kernel P_γ and the initial condition $X_0 = x$. Let $P^t((x, \gamma), B) = P(X_t \in B | X_0 = x, \Gamma_0 = \gamma)$, $B \in \mathcal{F}$, denote the t-step transition probability for the adaptive Markov chain with the initial conditions $X_0 = x$ and $\Gamma_0 = \gamma$. Let

$$T(x, \gamma, t) = \|P^t((x, \gamma), \cdot) - \pi(\cdot)\| = \sup_{B \in \mathcal{F}} |P^t((x, \gamma), B) - \pi(B)|$$

denote the total variation distance between the distribution of the adaptive Markov chain at time t and the target distribution $\pi(\cdot)$. It is said the adaptive Markov chain ergodic if $\lim_{t \to \infty} T(x, \gamma, t) = 0$ for all $x \in \mathbb{X}$ and $\gamma \in \mathbb{Y}$.

Theorem 3.1 concerns the ergodicity of an adaptive chain with changing stationary distributions, whose proof is given in the supplementary material of the paper.

Theorem 3.1 *(Ergodicity) Consider an adaptive Markov chain defined on the state space $(\mathbb{X}, \mathcal{F})$ with the adaption index $\Gamma_t \in \mathbb{Y}$. The adaptive Markov chain is ergodic if the following conditions are satisfied:*

(a) (Stationarity) There exists a stationary distribution $\pi_{\gamma_t}(\cdot)$ for each transition kernel P_{γ_t}, where γ_t denotes a realization of the random variable Γ_t.

(b) (Asymptotic Simultaneous Uniform Ergodicity) For any $\epsilon' > 0$, there exists a measurable set $E_1(\epsilon')$ in the probability space such that $P(E_1(\epsilon')) \geq 1 - \epsilon'$ and on this set $E_1(\epsilon')$, for $\epsilon > 0$, there exist constants $K(\epsilon) > 0$ and $N(\epsilon) > 0$ such that

$$\sup_{x \in \mathbb{X}} \|P_{\Gamma_t}^n(x, \cdot) - \pi(\cdot)\| \leqslant \epsilon,$$

for all $t > K(\epsilon)$ *and* $n > N(\epsilon)$.

(c) (Diminishing Adaptation) $\lim_{t\to 0} D_t = 0$ *in probability, where*

$$D_t = \sup_{x\in\mathbb{X}} \|P_{\Gamma_{t+1}}(x,\cdot) - P_{\Gamma_t}(x,\cdot)\|.$$

Theorem 3.1 is a slight extension of Theorem 1 of Roberts and Rosenthal (2007), where the ergodicity is established for an adaptive Markov chain for which all transition kernels admit the same stationary distribution and are simultaneously uniformly ergodic. Note that the conditions (a) and (b) of Theorem 3.1 imply that on the set $E_1(\epsilon')$

$$\|\pi_{\Gamma_t}(\cdot) - \pi(\cdot)\| = \|\pi_{\Gamma_t}(\cdot)P^n_{\Gamma_t} - \pi(\cdot)\| \le \sup_x \|P^n_{\Gamma_t}(x,\cdot) - \pi(\cdot)\| \le \epsilon, \tag{11}$$

for $t > K(\epsilon)$ and $n > N(\epsilon)$; that is, as $t \to \infty$, $\pi_{\Gamma_t}(\cdot)$ converges to π with probability greater than $1 - \epsilon'$. Hence, in the limit case, Theorem 3.1 is reduced to the ergodicity theorem of Roberts and Rosenthal (2007). The proof of Theorem 3.1 is also based on the coupling theory as in Roberts and Rosenthal (2007).

We note that under a slightly weaker condition of (b), Fort et al. (2011) established the same ergodicity result as Theorem 3.1. In this sense, Theorem 3.1 is redundant. Since part of its proof will be used in the proof of the next theorem, it is presented in the supplementary material. Fort et al. (2011) also established a strong law of large numbers for an adaptive Markov chain with changing stationary distributions, but under rather strong conditions. For example, it requires an explicit decreasing rate of D_t. However, figuring out the decreasing rate of D_t is very difficult, if not impossible, for AEX. To address this issue, we develop the next theorem, where D_t can converge to zero in probability at any rate. The proof of Theorem 3.2 can be found in the supplementary material of the paper.

Theorem 3.2 *(Weak Law of Large Numbers) Consider an adaptive Markov chain defined on the state space* $(\mathbb{X}, \mathcal{F})$. *Suppose that conditions (a), (b) and (c) of Theorem 3.1 hold. Let* $g(\cdot)$ *be a*

bounded measurable function. Then

$$\frac{1}{n}\sum_{t=1}^{n} g(X_t) \to \pi(g), \quad \textit{in probability},$$

as $n \to \infty$, *where* $\pi(g) = \int_{\mathbb{X}} g(x)\pi(dx)$.

3.2 Weak Convergence of Auxiliary SAMC Samples

To study the weak convergence of the auxiliary samples collected via SAMC, we first describe a general SAMC algorithm. Let

$$\pi(x) \propto \psi(x), \quad x \in \tilde{\mathcal{X}}, \tag{12}$$

denote the distribution that we want to draw samples from, where $\tilde{\mathcal{X}}$ denotes the sample space of $\pi(x)$. Let $E_1, \ldots, E_m$ denote m non-overlapping subregions, which form a partition of $\tilde{\mathcal{X}}$, i.e., $\tilde{\mathcal{X}} = \cup_{i=1}^{m} E_i$ and $E_i \cap E_j = \emptyset$ for $i \neq j$. Let $\boldsymbol{p} = (p_1, \ldots, p_m)$ denote the desired sampling frequencies of respective subregions. Let $\boldsymbol{w}_t = (w_t^{(1)}, \ldots, w_t^{(m)})$ denote the vector of abundance factors attached to the m subregions at iteration t, and let $\mathcal{W}$ denote the sample space of $\boldsymbol{w}_t$. Let $\{\mathcal{K}_s, s \geq 0\}$ be a sequence of compact subsets of $\mathcal{W}$ as defined in (7). Let $\tilde{\mathcal{X}}_0$ be a subset of $\tilde{\mathcal{X}}$, and let $\mathbb{T} : \tilde{\mathcal{X}} \times \mathcal{W} \to \tilde{\mathcal{X}}_0 \times \mathcal{K}_0$ be a truncation function which is measurable and maps a point in $\tilde{\mathcal{X}} \times \mathcal{W}$ to a random point in $\tilde{\mathcal{X}}_0 \times \mathcal{K}_0$. Let σ_t denote the number of truncations performed until iteration t. Let $\mathcal{S}$ denote the collection of the indices of the subregions from which a sample has been proposed; that is, $\mathcal{S}$ contains the indices of all subregions which are known to be non-empty. The algorithm starts with a random point drawn in $\tilde{\mathcal{X}}_0 \times \mathcal{K}_0$ and then iterates in the following steps:

A General SAMC Sampler

(a) (Sampling) Simulate a sample $x^{(t+1)}$ by a single MH update with the target distribution given by

$$\pi_t(x) \propto \sum_{i=1}^{m} \frac{\psi(x)}{w_t^{(i)}} I(x \in E_i).$$

(a.1) Generate y according to a proposal distribution $q(y|x_t)$. If $J(y) \notin \mathcal{S}$, then set $\mathcal{S} \leftarrow \mathcal{S} + \{J(y)\}$, where $J(y)$ denote the index of the subregion that the sample y belongs to.

(a.2) Calculate the ratio

$$r = \frac{w_t^{(J(x_t))}}{w_t^{(J(y))}} \frac{\psi(y)q(x^{(t)}|y)}{\psi(x^{(t)})q(y|x^{(t)})}. \tag{13}$$

(a.3) Accept the proposal with probability $\min(1, r)$. If it is accepted, set $x^{(t+1)} = y$; otherwise, set $x^{(t+1)} = x^{(t)}$.

(b) (Abundance factor updating) For all $i \in \mathcal{S}$, set

$$\log(w_{t+\frac{1}{2}}^{(i)}) = \log(w_t^{(i)}) + a_{t+1}(e_{t+1,i} - p_i),$$

where a_{t+1} is as defined in (A_1), and $e_{t+1,i} = 1$ if $x_{t+1} \in E_i$ and 0 otherwise. If $w_{t+1/2}^{(i)} \in \mathcal{K}_{\sigma_t}^{(i)}$ for all $i \in \mathcal{S}$, where $\mathcal{K}_{\sigma_t}^{(i)}$ denotes the ith dimensional space of $\mathcal{K}_{\sigma_t}$, then set $(\boldsymbol{w}_{t+1}, \boldsymbol{z}_{t+1}) = (\boldsymbol{w}_{t+1/2}, \boldsymbol{z}_{t+1})$ and $\sigma_{t+1} = \sigma_t$; otherwise, set $(\boldsymbol{w}_{t+1}, \boldsymbol{z}_{t+1}) = \mathbb{T}(\boldsymbol{w}_t, \boldsymbol{z}_t)$ and $\sigma_{t+1} = \sigma_t + 1$.

It is easy to see that the SAMC algorithm falls into the class of varying truncation stochastic approximation MCMC algorithms. The convergence of the varying truncation stochastic approximation MCMC algorithms has been studied in Andrieu et al. (2005) and Andrieu and Moulines (2006), where it is assumed that the Markov transition kernel $P_{\boldsymbol{w}}$, which is induced by (13) and depends on $\boldsymbol{w}$, satisfies some drift and minorisation conditions such that the resulting Markov chain is V-uniform ergodic. The function $V : \tilde{\mathcal{X}} \mapsto [1, \infty)$ is called the drift function. For SAMC, since the function $H(\boldsymbol{w}, x)$ is bounded and thus the mean field function and observation noise are bounded, we can set $V(x) \equiv 1$. Therefore, the resulting Markov chain is uniformly ergodic. For this reason, we assume that the Markov transition kernel $P_{\boldsymbol{w}}$ satisfies the Doeblin condition:

(A_2) (Doeblin condition) For any given $\boldsymbol{w} \in \mathcal{W}$, the Markov transition kernel $P_{\boldsymbol{w}}$ is irreducible and aperiodic. In addition, there exist an integer l, $0 < \delta < 1$, and a probability measure ν

such that for any compact subset $\mathcal{K} \subset \mathcal{W}$,

$$\inf_{\boldsymbol{w}\in\mathcal{K}} P^l_{\boldsymbol{w}}(x, A) \geq \delta\nu(A), \quad \forall x \in \tilde{\mathcal{X}},\ \forall A \in \mathcal{B}_{\tilde{\mathcal{X}}},$$

where $\mathcal{B}_{\tilde{\mathcal{X}}}$ denotes the Borel set of $\tilde{\mathcal{X}}$; that is, the whole support $\tilde{\mathcal{X}}$ is a *small* set for each kernel $P_{\boldsymbol{w}}$, $\boldsymbol{w} \in \mathcal{K}$.

To verify (A_2), one may assume that $\tilde{\mathcal{X}}$ is compact, $\psi(x)$ is bounded away from 0 and ∞ on $\tilde{\mathcal{X}}$, and the proposal distribution $q(y|x)$ satisfies the *local positive* condition:

(Q) There exists $\delta_q > 0$ and $\epsilon_q > 0$ such that, for every $x \in \mathcal{X}$, $|x - y| \leq \delta_q \Rightarrow q(y|x) \geq \epsilon_q$.

Then the condition (A_2) holds following from Theorem 2.2 of Roberts and Tweedie (1996), where it is shown that if the target distribution is bounded away from 0 and ∞ on every compact set of its support $\tilde{\mathcal{X}}$, then the MH chain with a proposal satisfying (Q) is irreducible and aperiodic, and every non-empty compact set is a *small* set. The proposals satisfying the local positive condition can also be easily designed for both continuous and discrete systems. For continuous systems, $q(y|x)$ can be set to a random walk Gaussian proposal, $y \sim N(x, \sigma^2 I_{d_x})$, where σ^2 can be calibrated to have a desired acceptance rate, e.g., $0.2 \sim 0.4$. For discrete systems, $q(y|x)$ can be set to a discrete distribution defined on a neighborhood of x. Besides the single-step MH move, the multiple-step MH move, the Gibbs sampler, and the Metropolis-within-Gibbs sampler can also be shown to satisfy (A_2) under appropriate conditions, see e.g. Rosenthal (1995) and Liang (2009) for the proofs. Note that to satisfy (A_2), $\tilde{\mathcal{X}}$ is not necessarily compact. Rosenthal (1995) gave one example for which the sample space is unbounded, yet the Markov chain is uniformly ergodic.

Under conditions (A_1) and (A_2), in a similar way to Liang et al. (2007), we can verify all the conditions given in Andrieu et al. (2005) for the convergence of varying truncation stochastic approximation MCMC algorithms. Hence, we claim that for SAMC, the number of truncations is almost surely finite and for all $i \in \mathcal{S}$,

$$\log(w_t^{(i)}) \to C + \log(\xi_i) - \log(p_i + \nu), \quad a.s., \tag{14}$$

as $t \to \infty$, where C denotes a constant, $\xi_i = \int_{E_i} \psi(x)dx$, and $\nu = \sum_{j\in\{i:\xi_i=0\}} p_j/(m - m_0)$ and m_0 is the cardinality of the set $\{i : \xi_i = 0\}$. The existence of empty subregions, i.e., the subregions with $\xi_i = 0$, is due to an inappropriate partition of the sample space, but SAMC does allow for the existence of empty subregions.

Further, we can show that a strong law of large numbers (SLLN) holds for SAMC: If $\psi(x)$ is bounded away from 0 and ∞ on $\mathcal{X}$ and $\tilde{\mathcal{X}}$ is compact, then for any bounded measurable function $G(x, \boldsymbol{w})$, where G is Lipschitz continuous with respect to $\boldsymbol{w}$, i.e. $|G(x, \boldsymbol{w}_1) - G(x, \boldsymbol{w}_2)| \le L\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$ for some $L > 0$, we have

$$\frac{1}{n}\sum_{t=1}^{n} G(X_t, \boldsymbol{w}_t) \to \pi_*(G), \quad \text{a.s.}, \quad \text{as } n \to \infty, \tag{15}$$

where $\pi_*(\cdot)$ denotes the limit distribution of $\pi_t(\cdot)$, and $\pi_*(G)$ denotes the expectation of $G(\cdot)$ with respect to $\pi_*(\cdot)$. A proof of (15) is given in the supplementary material (Lemma A.1). A similar result has been established in Atchade and Liu (2010) (Theorem 4.1) for a slightly different version of SAMC, where the gain factor sequence is self-adapted. However, verification of their conditions on stopping time is difficult.

For the auxiliary chain of AEX, we have $X_t = (\boldsymbol{z}_t, \vartheta_t)$, $\tilde{\mathcal{X}} = \mathcal{X}\times\{\theta^{(1)}, \ldots, \theta^{(m)}\}$ and $E_i = \mathcal{X}\times\{\theta^{(i)}\}$. In addition, ν, defined in (14), is equal to 0, as $\int_{\mathcal{X}} \varphi(\boldsymbol{x}, \theta^{(i)})d\boldsymbol{x} > 0$ for each i. Let N denote the total number of iterations of an AEX run. Then, it follows from (14) and (15) that

$$\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_t^{(i)}\frac{\varphi(\boldsymbol{z}_t, \theta')}{\varphi(\boldsymbol{z}_t, \vartheta_t)}I(\vartheta_t = \theta^{(i)})\right\} \to \sum_{i=1}^{m}\int_{\mathcal{X}}\frac{\kappa(\theta^{(i)})}{p_i}\frac{\varphi(\boldsymbol{z}, \theta')}{\varphi(\boldsymbol{z}, \theta^{(i)})}p_i f(\boldsymbol{z}|\theta^{(i)})d\boldsymbol{z} = m\kappa(\theta'), \quad a.s., \tag{16}$$

as $N \to \infty$, provided that $\mathcal{X}$ is compact and thus the ratio $\varphi(\boldsymbol{z}, \theta')/\varphi(\boldsymbol{z}, \theta^{(i)})$ is bounded for all $\boldsymbol{z} \in \mathcal{X}$. Note that as $t \to \infty$, the marginal distribution of $\boldsymbol{z}_t$ converges to the mixture distribution $p(\boldsymbol{z}) = \sum_{i=1}^{m} p_i f(\boldsymbol{z}|\theta^{(i)})$. Since here we have implicitly assumed that the distributions $f(\boldsymbol{z}|\theta^{(i)}), \ldots, f(\boldsymbol{z}|\theta^{(m)})$ share the same sample space $\mathcal{X}$, the condition (A_0) is not necessary for, but does benefit, the convergence of (16). Recall that the major role of the auxiliary parameters is to be used for construction of a good trial distribution for drawing the auxiliary variables used in the

target chain, and a good trial distribution can always improve the convergence of the importance sampling procedure. Similarly, for any Borel set $A \subset \mathcal{X}$, we have

(17)

Putting (16) and (17) together, we have, as $N \to \infty$,

$$\frac{\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_t^{(i)}\frac{\varphi(\mathbf{Z}_t,\theta')}{\varphi(\mathbf{Z}_t,\vartheta_t)}I(\boldsymbol{z}_t \in A \,\&\, \vartheta_t = \theta^{(i)})\right\}}{\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_t^{(i)}\frac{\varphi(\mathbf{Z}_t,\theta')}{\varphi(\mathbf{Z}_t,\vartheta_t)}I(\vartheta_t = \theta^{(i)})\right\}} \to \int_A f(\boldsymbol{z}|\theta')d\boldsymbol{z}, \quad a.s.,$$

which, by Lebesgue's dominated convergence theorem, implies that

$$P(\boldsymbol{x} \in A|\theta') = E\left[P(\boldsymbol{x} \in A|\boldsymbol{z}_1,\vartheta_1,w_1;\ldots;\boldsymbol{z}_N,\vartheta_N,w_N;\theta')\right] \to \int_A f(\boldsymbol{z}|\theta')d\boldsymbol{z}, \tag{18}$$

where $(\boldsymbol{z}_1,\vartheta_1,w_1;\ldots;\boldsymbol{z}_N,\vartheta_N,w_N)$ denotes a set of samples generated by the auxiliary chain, and $\boldsymbol{x}$ denotes a sample resampled from $(\boldsymbol{z}_1,\vartheta_1,w_1;\ldots;\boldsymbol{z}_N,\vartheta_N,w_N)$. Furthermore, if the parameter space Θ is compact and the function $\varphi(\boldsymbol{z},\theta)$ is upper semi-continuous in θ for all $\boldsymbol{z} \in \mathcal{X}$, then the convergence in (18) is uniform over Θ. Refer to Ferguson (1996)(Theorem 16(a)) for a proof of uniform strong law of large numbers. In summary, we have the following lemma:

Lemma 3.1 *Assume that the conditions* (A_1) *and* (A_2) *are satisfied, and* $\mathcal{X}$ *is compact.* $\varphi(x,\theta)$ *is bounded away from 0 and* ∞ *on* $\mathcal{X}\times\Theta$. *Let* $\{\boldsymbol{z}_1,\vartheta_1,w_1;\ldots;\boldsymbol{z}_N,\vartheta_N,w_N\}$ *denote a set of samples generated by SAMC in an AEX run, and let* $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ *be distinct samples in* $\boldsymbol{z}_t$. *Resample a random variable/vector X from* $\{\boldsymbol{z}_1,\vartheta_1,w_1;\ldots;\boldsymbol{z}_N,\vartheta_N,w_N\}$ *such that*

$$P(X = \boldsymbol{x}_k|\theta') = \frac{\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_t^{(i)}\frac{\varphi(\mathbf{Z}_t,\theta')}{\varphi(\mathbf{Z}_t,\vartheta_t)}I(\boldsymbol{z}_t = \boldsymbol{x}_k \,\&\, \vartheta_t = \theta^{(i)})\right\}}{\sum_{t=1}^{N}\sum_{i=1}^{m}\left\{w_t^{(i)}\frac{\varphi(\mathbf{Z}_t,\theta')}{\varphi(\mathbf{Z}_t,\vartheta_t)}I(\vartheta_t = \theta^{(i)})\right\}}, \quad k = 1,\ldots,n,$$

then the distribution of X converges to $f(\cdot|\theta')$ *almost surely as* $N \to \infty$. *Furthermore, if* Θ *is compact and the unnormalized density function* $\varphi(\boldsymbol{z},\theta)$ *is upper semi-continuous in* θ *for all* $\boldsymbol{z} \in \mathcal{X}$, *then the convergence is uniform over* Θ.

As aforementioned, since $f(\boldsymbol{z}|\theta^{(i)})$'s share the same sample space $\mathcal{X}$, the condition (A_0) is not necessary for, but does benefit, the convergence stated in the lemma. For the efficiency of the algorithm, (A_0) is generally required to be satisfied.

3.3 Convergence of AEX

To study the convergence of AEX, we first note that $\{\theta_t\}$ forms an adaptive Markov chain with the transition kernel given by

$$\tilde{P}_l(\theta, d\theta') = \int_{\mathcal{X}} \alpha(\theta, \boldsymbol{x}, \theta') q(d\theta'|\theta) \nu_l(d\boldsymbol{x}|\theta') + \delta_\theta(d\theta')[1 - \int_{\Theta\times\mathcal{X}} \alpha(\theta, \boldsymbol{x}, \vartheta') q(d\vartheta'|\theta) \nu_l(d\boldsymbol{x}|\vartheta')], \qquad (19)$$

where $\alpha(\theta, \boldsymbol{x}, \theta')$ is defined in (10), l denotes the cardinality of the set of auxiliary variables collected from the auxiliary Markov chain, i.e., $l = |S_l|$, and $\nu_l(\boldsymbol{x}|\theta')$ denotes the true distribution of $\boldsymbol{x}$ resampled from S_l. For mathematical simplicity, we assume that the parameter space Θ (of θ) and the sample space $\mathcal{X}$ (of $\boldsymbol{x}$) are both compact. Although this makes our theory a little restrictive, it is quite common in the study of adaptive Markov chain theory, say e.g., Haario et al. (2001). Under this assumption, we have the following lemma, whose proof can be found in the supplementary material.

Lemma 3.2 *Assume that conditions* (A_1) *and* (A_2) *are satisfied, both* $\mathcal{X}$ *and* Θ *are compact, and the unnormalized density function* $\varphi(\boldsymbol{x}, \theta)$ *is continuously differentiable in* θ *for all* $\boldsymbol{x} \in \mathcal{X}$ *and bounded away from 0 and* ∞ *on* $\mathcal{X} \times \Theta$. *Furthermore, assume* $\pi(\theta)$ *and* $q(\theta'|\theta)$ *are continuously differentiable in* θ *and* θ' *as well. Define* $\tilde{D}_l = \sup_{\theta\in\Theta} ||\tilde{P}_l(\theta, \cdot) - P(\theta, \cdot)||$, *where* $P(\theta, \cdot)$, *as defined in (5), denotes the transition kernel of the exchange algorithm with perfect auxiliary variables. Then* $\tilde{D}_l \to 0$ *almost surely as* $l \to \infty$.

It is known that $P(\theta, d\theta')$ can induce a Markov chain which is irreducible, aperiodic and admits $\pi(\theta|\boldsymbol{y})$ as the invariant distribution, provided an appropriate proposal $q(\cdot|\cdot)$ been used therein. Define

$$\beta_l(\theta, \boldsymbol{x}, \theta') = \frac{\nu_l(\boldsymbol{x}|\theta')}{\nu_l(\boldsymbol{x}|\theta)} \frac{f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta')}, \quad \text{and}$$

$$r(\theta, \boldsymbol{x}, \theta') = \frac{\pi(\theta') f(\boldsymbol{y}|\theta')}{\pi(\theta) f(\boldsymbol{y}|\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)} \frac{f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x}|\theta')}, \quad r_v(\theta, \boldsymbol{x}, \theta') = \frac{\pi(\theta') f(\boldsymbol{y}|\theta')}{\pi(\theta) f(\boldsymbol{y}|\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)} \frac{\nu_l(\boldsymbol{x}|\theta)}{\nu_l(\boldsymbol{x}|\theta')}.$$

Then it is easy to see that

$$r(\theta, \boldsymbol{x}, \theta') = \beta_l(\theta, \boldsymbol{x}, \theta') r_v(\theta, \boldsymbol{x}, \theta'),$$

and the Markov chain defined by the acceptance rule $\min\{1, r_v(\theta, \boldsymbol{x}, \theta')\}$ is irreducible, aperiodic and admits the invariant distribution $\pi(\theta|\boldsymbol{y})$, provided an appropriate proposal $q(\cdot,\cdot)$ has been used therein. To ensure the convergence of the Markov chain, $\nu_l(\boldsymbol{x}|\theta)$ is not necessarily to have a support as large as $\mathcal{X}$. In fact, its support can be only a subset of $\mathcal{X}$. Let $P_v(\theta, d\theta')$ denote the transitional kernel of the Markov chain induced by the acceptance rule $\min\{1, r_v(\theta, \boldsymbol{x}, \theta')\}$. Lemma 3.3 shows that $\tilde{P}_l(\theta, d\theta')$, defined in (19), is also irreducible and aperiodic and admits an invariant distribution. This is done by showing that the accessible sets of P_v are included in those of $\tilde{P}_l$. The details are given in the supplementary material.

Lemma 3.3 *Assume (i) the conditions of Lemma 3.2 hold and (ii) P is irreducible and aperiodic and admits an invariant distribution. Then $\tilde{P}_l$, defined in (19), is irreducible and aperiodic, and hence there exists a stationary distribution $\tilde{\pi}_l(\theta|x)$ such that for any $\theta_0 \in \Theta$,*

$$\lim_{k\to\infty} ||\tilde{P}_l^k(\theta_0, \cdot) - \tilde{\pi}_l(\cdot|\boldsymbol{y})|| = 0.$$

If the proposal $q(\cdot,\cdot)$ satisfies the local positive condition (Q), then the ergodicity is uniform over Θ.

Lemma 3.4 concerns the simultaneous uniform ergodicity of the kernel $\tilde{P}_l$'s, whose proof can be found in the supplementary material.

Lemma 3.4 *Assume the conditions of Lemma 3.3 hold. If the proposal $q(\cdot,\cdot)$ satisfies the local positive condition, then for any $e > 0$ there exits a measurable set E_0 in the probability space such that $P(E_0) > 1 - e$ and on this set E_0, for any $\varepsilon > 0$, there exist $L(\varepsilon) \in \mathbb{N}$ and $K(\varepsilon) \in \mathbb{N}$ such that for any $l > L(\varepsilon)$ and $k > K(\varepsilon)$,*

$$||\tilde{P}_l^k(\theta_0, \cdot) - \pi(\cdot|\boldsymbol{y})|| \le \varepsilon, \quad \textit{for all } \theta_0 \in \Theta. \tag{20}$$

Define $D_l = \sup_{\theta\in\Theta} ||\tilde{P}_{l+1}(\theta, \cdot) - \tilde{P}_l(\theta, \cdot)||$. Since $D_l \le \sup_{\theta\in\Theta} ||\tilde{P}_{l+1}(\theta, \cdot) - P(\theta, \cdot)|| + \sup_{\theta\in\Theta} ||\tilde{P}_l(\theta, \cdot) - P(\theta, \cdot)||$, by Lemma 3.2, we have $\lim_{l\to\infty} D_l = 0$ almost surely. That is, $\tilde{P}_l$ satisfies the diminishing

adaptation condition of Theorem 3.1. Putting this together with Lemmas 3.3 and 3.4, we have the following theorem, which states that AEX is ergodic and the weak law of large numbers holds for the sample path average.

Theorem 3.3 *(Ergodicity and WLLN) Assume the conditions of Lemma 3.3 hold. If the proposal* $q(\cdot,\cdot)$ *satisfies the local positive condition and the unnormalized density function* $\varphi(\boldsymbol{x},\theta)$ *is upper semi-continuous in* θ *for all* $\boldsymbol{x} \in \mathcal{X}$*, then the adaptive exchange algorithm is ergodic and for any bounded measurable function g,*

$$\frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \to \pi(g|\boldsymbol{y}), \quad \textit{in probability},$$

as $n \to \infty$*, where* $\pi(g|\boldsymbol{y}) = \int_\Theta g(\theta)\pi(\theta|\boldsymbol{y})d\theta$.

4 Spatial Autologistic Model

The spatial autologistic model (Besag, 1974) has been widely used in spatial data analysis (e.g., Preisler, 1993; Sherman et al., 2006). Let $\boldsymbol{y} = \{y_i : i \in \mathcal{D}\}$ denote the observed binary data, where y_i is called a spin and $\mathcal{D}$ is the set of indices of the spins. Let $|\mathcal{D}|$ denote the total number of spins in $\mathcal{D}$, and let $n(i)$ denote the set of neighbors of spin i. The likelihood function of the model is given by

$$f(\boldsymbol{y}|\alpha,\beta) = \frac{1}{\kappa(\alpha,\beta)} \exp\left\{\alpha \sum_{i\in\mathcal{D}} y_i + \frac{\beta}{2}\sum_{i\in\mathcal{D}} y_i \left(\sum_{j\in n(i)} y_j\right)\right\}, \quad (\alpha,\beta) \in \Theta, \tag{21}$$

where the parameter α determines the overall proportion of $y_i = \pm 1$, the parameter β determines the intensity of interaction between y_i and its neighbors, and $\kappa(\alpha,\beta)$ is the intractable normalizing constant defined by

$$\kappa(\alpha,\beta) = \sum_{\text{for all possible } \boldsymbol{y}} \exp\left\{\alpha \sum_{i\in\mathcal{D}} y_i + \frac{\beta}{2}\sum_{i\in\mathcal{D}} y_i \left(\sum_{j\in n(i)} y_j\right)\right\}. \tag{22}$$

An exact evaluation of $\kappa(\alpha, \beta)$ is prohibited even for a moderate system. To conduct a Bayesian analysis for the model, a uniform prior on

$$(\alpha, \beta) \in \Theta = [-1, 1] \times [0, 1],$$

is assumed for the parameters, which restricts Θ to be a compact set. Since $\mathcal{D}$ is finite, the sample space $\mathcal{X}$ (of $\boldsymbol{y}$) is also finite.

4.1 U.S. Cancer Mortality Data

United States cancer mortality maps have been compiled by Riggan et al. (1987) for investigating the possible association of cancer with unusual demographic, environmental, and industrial characteristics, or employment patterns. Figure 1(a) shows the mortality map for cancer of the liver and gallbladder (including bile ducts) cancers in white males during the decade 1950-1959, which indicates some apparent geographic clustering. See Sherman et al. (2006) for more descriptions of the data. Following Sherman et al. (2006), we modeled the data with a spatial autologistic model. The total number of spins is $|\mathcal{D}| = 2293$. A free boundary condition is assumed for the model, under which the boundary points have fewer neighboring points than the interior points. This assumption is natural to this dataset, as the lattice has an irregular shape.

The AEX algorithm was first applied to this example. It was run for 10 times independently. Each run consisted of three stages. The first stage was to choose the auxiliary parameters $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$. For this purpose, the fractional DMH algorithm was run for $N_1 = 5500$ iterations and 5000 samples of θ were collected after a burn-in period of 500 iterations, and then $m = 100$ auxiliary parameters were selected from the 5000 collected samples using the *max-min* procedure. The second stage was to build up an initial database of auxiliary variables. In this stage, only the auxiliary chain was run. The auxiliary chain started with the observed data $\boldsymbol{y}$ and $\boldsymbol{w}_0 = (1, \ldots, 1)$, and consisted of $N_2 = 1.1 \times 10^6$ iterations. The first 10^5 iterations were discarded for the burn-

in process, and then a database of 50000 auxiliary variables was collected from the remainder of the run at an equal time space of $s_0 = 20$ iterations. In the simulations, we set $\mathcal{X}_0 = \mathcal{X}$ and $\mathcal{K}_i = [0, 10^{100+10i}]$ for $i = 0, 1, 2, \ldots$. In the third stage, the auxiliary chain and the target chain were run simultaneously. The auxiliary chain was run for $N_3 = 4 \times 10^5$ iterations and the samples generated by it were continuously collected (at a time space of $s_0 = 20$ iterations) and added into the database, and the target chain was run for $20,000$ iterations. The target chain started with the average of the θ samples collected from the fractional DMH run, and was run for one iteration once a new auxiliary variable was added to the database. The CPU time for the whole run is 107s, which is measured on a single core of Intel® Xeon® CPU E5-2690(2.90Ghz) (the same computer was used for all simulations of this paper). The samples generated in the target chain were used for Bayesian inference of the model. For both the fractional DMH chain and the target chain, a Gaussian random walk proposal distribution with the covariance matrix $0.03^2 I_2$ was used, where I_2 denotes the 2-by-2 identity matrix. The overall acceptance rate of the target chain is 0.21, and that of the fractional DMH chain is 0.33. For the auxiliary chain, the proposal distribution $T_2(\cdot|\cdot)$ was set to a single cycle of Gibbs updates.

Figure 2(a) shows the scatter plot of the fractional DMH samples and the selected auxiliary parameters. The scatter plot of the selected auxiliary parameters is also shown in Figure 2(b). As expected, the selected auxiliary parameters are distributed over the space of fractional DMH samples, and the convex hulls formed by them are about the same. Figure 2(c) shows the histogram of the 100 auxiliary parameters achieved by the auxiliary chain in a run with the uniform desired sampling distribution. The flatness of the histogram implies that the auxiliary chain has converged.

A MCMC convergence diagnosis based on Gelman-Rubin's shrink factor (Gelman and Rubin, 1992), as shown in Figure 3, indicates that for this example AEX can converge very fast, usually within a few hundred iterations. Figure 3 was generated using the R package *coda* (Plummer et al., 2012) based on ten independent runs of AEX. As for conventional MCMC algorithms, the samples generated in the early stage of these runs can be discarded for the burn-in process. Table

1 summarizes the parameter estimation results for the model based on the ten independent runs.

To assess the validity of the AEX algorithm, the exchange algorithm was applied to this example with the same proposal distribution as used in the target chain of AEX. As previously mentioned, the exchange algorithm is an auxiliary variable MCMC algorithm that requires a perfect sampler for generating auxiliary variables, and it can sample correctly from the posterior distribution when the number of iterations is large. Hence, the estimates produced by the exchange algorithm can be used as the test standard for assessing the performance of AEX. For this example, the auxiliary variables were generated using the summary state algorithm (Childs et al., 2001), which is suitable for high-dimensional binary spaces. The exchange algorithm was run for 10 times independently. Each run consisted of 20,500 iterations, where the first 500 iterations were discarded for the burn-in process, and the remaining 20,000 iterations were used for estimation of θ. Therefore, in each run, the exchange and AEX algorithm produced the same number of samples. Each run of the exchange algorithm costs about 106s CPU time, and the overall acceptance rate was 0.21 which indicates the algorithm has been implemented efficiently. The numerical results are summarized in Table 1. The estimates produced by the AEX and exchange algorithms are also very close to those reported in the literature. Liang (2007) analyzed these data using a contour Monte Carlo method and produced the estimate (-0.3008, 0.1231). The contour Monte Carlo method approximates the normalizing constant function on a given region and then estimates the parameters based on the approximated normalizing constant function. As reported by Liang (2007), this algorithm took several hours of CPU time to approximate the normalizing constant function. Sherman et al. (2006) analyzed the data using the MCMLE method (Geyer and Thompson, 1992), and obtained the estimate (-0.304, 0.117), which is different from other Bayesian estimates. This may reflect the difference between the posterior mean and posterior mode.

For a thorough comparison, we have also applied the DMH and MCMH algorithm to this example. As aforementioned, the DMH sampler is the same with the exchange algorithm except that it draws the auxiliary variable $\boldsymbol{x}$ at each iteration through a finite run of the MH algorithm

initialized at the observation $\boldsymbol{y}$. In our implementation, we set the length of the finite MH run to be $K = 50$. DMH was run for 10 times. Each run consisted of 20,500 iterations and cost about 142s, where the samples generated in the last 20,000 iterations were used for parameter estimation. The results are summarized in Table 1.

The MCMH algorithm works in a different idea from DMH. It is not to cancel the unknown normalizing constant ratio using auxiliary variables, but to approximate the normalizing constant ratio using auxiliary variables. Let θ_t denote the sample of θ drawn at iteration t, and let $\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_K^{(t)}$ denote the auxiliary samples simulated from the distribution $f(\boldsymbol{x}|\theta_t)$. One iteration of the MCMH algorithm consists of the following steps:

Monte Carlo MH algorithm:

1. Draw θ' from a proposal distribution $q(\theta'|\theta_t)$.

2. Estimate the normalizing constant ratio $\kappa(\theta')/\kappa(\theta_t)$ by $\widehat{R}(\theta_t, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_K, \theta') = \frac{1}{K}\sum_{i=1}^{K} \frac{\psi(\boldsymbol{X}_i,\theta')}{\psi(\boldsymbol{X}_i,\theta_t)}$, and then calculate the ratio

$$\tilde{r}(\theta_t, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_K, \theta') = \frac{1}{\widehat{R}(\theta_t, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_K, \theta')} \frac{\psi(\boldsymbol{y}, \theta')\pi(\theta')q(\theta_t|\theta')}{\psi(\boldsymbol{y}, \theta_t)\pi(\theta_t)q(\theta'|\theta_t)}.$$

3. Set $\theta_{t+1} = \theta'$ with probability $\min\{1, \tilde{r}(\theta_t, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_K, \theta')\}$ and set $\theta_{t+1} = \theta_t$ with the remaining probability.

4. If the proposal is rejected in step 3, set $\boldsymbol{x}_i^{(t+1)} = \boldsymbol{x}_i^{(t)}$ for $i = 1, \ldots, K$. Otherwise, simulate samples $\boldsymbol{x}_1^{(t+1)}, \ldots, \boldsymbol{x}_K^{(t+1)}$ from $f(\boldsymbol{x}|\theta_{t+1})$ using the MH algorithm.

We note that the MCMH algorithm, an early version of this algorithm is presented in the book by Liang et al. (2010), is similar to the marginal PMCMC algorithm suggested by Everitt (2012). The MCMH algorithm is to use auxiliary variables to approximate the unknown normalizing constant ratio, while the marginal PMCMC algorithm, based on the work by Andrieu et al. (2010), is to use sequential Monte Carlo samples to approximate the unknown normalizing constant.

In our implementation of MCMH, we set $K = 100$ and the proposal $q(\cdot|\cdot)$ to be a Gaussian random walk with the covariance matrix $0.03^2 I_2$. The algorithm was run for 10 times independently. Each run consisted of 20500 iterations, where the first 500 iterations were for the burn-in process, and cost about 143s CPU time. The overall acceptance rate was about 0.35. The results are summarized in Table 1.

Table 1 shows that all the algorithms, AEX, DMH, MCMH and exchange, work well for this example. By treating the estimates of the exchange algorithm as the testing standard, it is easy to see that the estimates resulted from all the other three algorithms are unbiased. This is because the underlying system of this example dataset is only weakly dependent, and thus independent auxiliary samples can be easily drawn at each iteration by running a short Markov chain. In terms of efficiency, MCMH is the best among the three approximate MCMC algorithms. This is reasonable, as in MCMH the auxiliary sample are only drawn when a new sample of θ is accepted and all the auxiliary samples are used in simulating new samples of θ. For this example, MCMH has about the same efficiency as the exchange algorithm after taking account of their CPU times and standard deviations of the estimates. Compared to MCMH, AEX and DMH are less efficient in the use of auxiliary samples. AEX makes use of only a small proportion of the auxiliary samples and discards all the others. DMH is similar; it uses only the last sample generated by the short Markov chain at each iteration.

In Section 4.2, we present one example for which the underlying system is strongly dependent. Since, for such a system, independent auxiliary samples are difficult to draw with a short run of Markov chain, AEX outperforms the DMH and MCMH algorithms.

4.2 A Simulation Study

To assess the performance of the AEX algorithm for strongly dependent systems, we simulated one U.S. cancer mortality dataset using the summary state algorithm with $\alpha = 0$ and $\beta = 0.45$.

Since the lattice is irregular, the free boundary condition was again assumed in the simulation. Note that under the setting $\alpha = 0$, the autologistic model is reduced to the Ising model. Hence, the underlying system for this simulated dataset is strongly dependent by noting that the interaction parameter β is greater than the critical value (≈ 0.44) of the Ising model.

AEX was first applied to this example. It was run as for the last example except that the simulations in the second and third stages have been lengthened. For this example, we set $N_2 = 6 \times 10^6$, $s_0 = 50$, and $N_3 = 10^6$, where the first 10^6 iterations in the second stage were used for the burn-in process. AEX was run for 10 times independently, and each run costs about 327s. The numerical results are summarized in Table 2.

For comparison, we have also applied the exchange, DMH and MCMH to this example. The exchange algorithm was run for 10 times. Each run consisted of 1200 iterations, where the first 200 iterations were discarded for the burn-in process and the samples generated in the remaining iterations were used for parameter estimation. The proposal distribution is the same as that used for the last example. For the exchange algorithm, due to its difficulty in generating perfect samples, the CPU time of each run can be very long although consisting of only 1200 iterations. In addition, the CPU times of different runs can be much different. Childs et al. (2001) studied the behavior of the perfect sampler for the Ising model. They fitted an exponential law for the convergence time, and reported that the perfect sampler may not work well when the value of β is close to the critical value (≈ 0.44). Their finding is consistent with our results reported in Table 2.

In applying DMH to this example, we tried different values of K, $K = 10$, 100, 500 and 1000, to assess the effect of mixing of the short Markov chain on parameter estimation. For MCMH, we have also tried the same settings of K. Both DMH and MCMH were run for 10 times independently, and each run consisted of 20,500 iterations with the first 500 iterations discarded for the burn-in process. The results are summarized in Table 2. As shown in Table 2, for the same value of K, MCMH generally costs less CPU times than DMH, as MCMH only generates auxiliary variables when a new value of θ is accepted, while in DMH this needs to be done at each iteration. Table 2

shows that both the DMH and MCMH estimates of β have a trend of decreasing toward the true value 0.45. However, even with $K = 1000$, at which both DMH and MCMH cost more CPU times than AEX, their estimates are still far from the true value. In contrast, even AEX costs much less CPU time than these two algorithms (with $K = 1000$), its estimate of β is much closer to the true value and also the estimate of the exchange algorithm.

To have a fully exploration for the performance of different algorithms, we show in Figure 4 the box-plots of the β estimates resulted from 10 independent runs by the above four algorithms. It shows again the superiority of AEX for this example: AEX can work well for the problems for which the underlying system is strongly dependent, while DMH and MCMH cannot.

5 Autonormal Model

5.1 Autonormal Model

To demonstrate the performance of the AEX algorithm on high dimensional models, we consider a second-order zero-mean Gaussian Markov random field $\boldsymbol{Y} = (Y_{ij})$ defined on an $M \times N$ lattice, whose conditional density function is given by

(23)

where $\beta_h, \beta_v, \beta_d$ and σ^2 are four parameters, $n_h(i, j) = \{(i, j-1), (i, j+1)\}$, $n_v(i, j) = \{(i-1, j), (i+1, j)\}$ and $n_d(i, j) = \{(i-1, j-1), (i-1, j+1), (i+1, j-1), (i+1, j+1)\}$ are neighbors of (i, j). This model is stationary when $|\beta_h| + |\beta_v| + 2|\beta_d| < 0.5$ (Balram and Moura, 1993). Let $\theta = (\beta_h, \beta_v, \beta_d, \sigma^2)$ denote the parameter vector. The joint likelihood function of this model can be written as

$$f(\boldsymbol{y}|\theta) = (2\pi\sigma^2)^{-MN/2}|B|^{1/2} \exp\{-\frac{1}{2\sigma^2}\boldsymbol{y}'B\boldsymbol{y}\},$$

where B is an $(MN \times MN)$-dimensional matrix, and $|B|$ is intractable except for some special cases (Besag and Moran, 1975).

To conduct a Bayesian analysis for the model, we assume the following priors:

$$\pi(\boldsymbol{\beta}) \propto I(|\beta_h| + |\beta_v| + 2|\beta_d| < 0.5), \qquad \pi(\sigma^2) \propto \frac{1}{\sigma^2}, \tag{24}$$

which $I(\cdot)$ is the indicator function. Under the free boundary condition for which the boundary pixels have fewer neighbors, we have the following posterior distribution

$$\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto (\sigma^2)^{-\frac{MN}{2}-1}|B|^{1/2} \exp\Big\{-\frac{MN}{2\sigma^2}\Big(S_y - 2\beta_h Y_h - 2\beta_v Y_v - 2\beta_d Y_d\Big)\Big\} I(|\beta_h| + |\beta_v| + 2|\beta_d| < 0.5), \tag{25}$$

where

Although σ^2 can be integrated out from the posterior, we do not suggest to do so. Working on the joint posterior will ease the generation of auxiliary variables in AEX.

5.2 Wheat Yield Data

The wheat yield data was collected on a 20×25 rectangular lattice (Table 6.1, Andrews and Herzberg, 1985). The data was shown in Figure 5(a), which indicates positive correlation between neighboring observations. This data has been analyzed by a number of authors, e.g., Besag (1974), Huang and Ogata (1999), and Gu and Zhu (2001). Following the previous authors, we subtracted the mean from the data and then fitted them by the autonormal model. In our analysis, the free boundary condition is assumed. This is natural, as the lattice for the real data is often irregular.

The AEX algorithm was applied to this example in a similar way to the cancer mortality data example. The algorithm was run for 10 times independently. Each run consisted of three stages, and their settings are the same as the runs for the simulated cancer mortality data except for the proposal distributions. For both the fractional DMH chain and the target chain, a Gaussian random walk proposal with the covariance matrix $0.01^2 I_4$ was used, where I_4 denotes the 4-by-4 identity matrix. For the auxiliary chain, the proposal $T_2(\cdot|\cdot)$ was set to a single cycle of Gibbs updates:

$$y_{ij}|\boldsymbol{y}_{(u,v)\in n(i,j)} \sim N\left(\beta_{ht}\sum_{(u,v)\in n_h(i,j)} y_{uv} + \beta_{vt}\sum_{(u,v)\in n_v(i,j)} y_{uv} + \beta_{dt}\sum_{(u,v)\in n_d(i,j)} y_{uv}, \sigma_t^2\right),$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, where $(\beta_{ht}, \beta_{vt}, \beta_{dt}, \sigma_t^2)$ denotes the value of θ at iteration t. Each run of AEX cost about 402s. Figure 5(b) shows the histogram of the 100 auxiliary parameters achieved by the auxiliary chain at the end of the second stage. It indicates that the auxiliary chain can mix very well for different auxiliary parameters. The parameter estimation results are summarized in Table 3.

For this example, the exchange algorithm is not applicable, as the perfect sampler is not available for the autonormal model. However, under the free boundary condition, the log-likelihood function of the model admits the following analytic form (Balram and Moura, 1993):

(26)

where S_y, Y_h, Y_v and Y_d are as defined in (25). The Bayesian inference for the model is then standard, with the priors as specified in (24). For comparison, the MH algorithm has been applied to simulate from the resulting analytic posterior distribution. It was run for 10 times with each run consisting of 20,500 iterations, where the first 500 iterations were discarded for the burn-in process. The proposal distribution adopted here was a Gaussian random walk with the same covariance matrix as that used in AEX. The overall acceptance rate is about 0.45. The resulting parameter estimates, the so-called true Bayes estimates, are summarized in Table 3. The comparison indicates that AEX works well for this example.

This example implies that AEX can work well for reasonably high dimensional problems. As aforementioned, the key to the success of AEX is to be able to generate auxiliary variables at a set of auxiliary parameters that cover the support of the true posterior. In AEX, all the auxiliary parameters are selected from the samples generated by the fractional DMH algorithm which can result in an expanded support of the true posterior. This, together with the *max-min* procedure, ensures that the selected auxiliary parameters are able to cover the support of the true posterior. Since the auxiliary parameters are essentially generated from the same posterior distribution, the neighboring distributions $f(\boldsymbol{x}|\theta^{(i)})$'s are reasonably overlapped by noting that they all share, at least,

the same sample—the observation $\boldsymbol{y}$. This ensures a smooth transition of the auxiliary chain between different auxiliary parameters and thus the success of AEX. More importantly, this property of AEX holds independent of the dimension of θ. For the wheat yield data example, the overall acceptance rate for the transitions between different auxiliary parameters is about 0.25.

6 Conclusion

We have proposed a new algorithm, the adaptive exchange algorithm or AEX in short, for sampling from distributions with intractable normalizing constants. The new algorithm can be viewed as a MCMC extension of the exchange algorithm, which generates auxiliary variables via an importance sampling procedure from a Markov chain running in parallel. The convergence of the new algorithm, including ergodicity and weak law of large numbers, has been established under mild conditions. Compared to the exchange algorithm, the new algorithm removes the requirement of perfect sampling, and thus can be applied to many models for which perfect sampling is not available or very expensive. The new algorithm has been tested on the spatial autologistic and autonormal models. The numerical results indicate that the new algorithm can outperform other approximate MCMC algorithms, such as the DMH and MCMH algorithms, for strongly dependent systems.

We have also applied the AEX algorithm to some more challenging problems, such as exponential random graph models for social networks. Our numerical results indicate that the AEX algorithm can outperform other approximate algorithms for large social networks. When the network is large, generating independent auxiliary networks is often difficult with a short Markov chain. Due to the space limit, these results will be reported elsewhere.

Our implementation of AEX is plain. Its efficiency can be improved in various respects. For example, in the current implementation, the auxiliary parameters are selected using a *max-min* procedure. In the future, a self-learning process may be introduced to the auxiliary Markov chain

toward an optimal selection of the auxiliary parameters. To improve the mixing of the target chain, a population-based MCMC algorithm may be used, e.g., adaptive direction sampling, parallel tempering or evolutionary Monte Carlo. We note that AEX can slow down with iterations, because it needs to resample from an increasingly large database.

Finally, we would like to compare, mainly conceptually, the AEX algorithm and the Russian Roulette sampling algorithm (Lyne et al., 2014). The latter is developed based on the pseudo-marginal MCMC approach of Andrieu and Roberts (2009), which requires an unbiased independent estimate of the intractable normalizing constant at each iteration. Hence, the algorithm can be extremely slow. For an Ising model example of 100 spins, where the normalizing constant is estimated using annealing importance sampling (Neal, 1998) at each iteration, the algorithm costs over 100 CPU minutes (on the same computer as used by AEX for the previous examples) for running 1000 iterations under the authors' default setting. The AEX algorithm also relies on the estimates of intractable normalizing constants, but only on a number of pre-specified points. Moreover, the estimates are obtained in an online manner and can be improved with iterations. Hence, the AEX algorithm can be much more efficient than the Russian Roulette sampling algorithm. This is evidenced by Table 2, where it is reported that AEX cost only 5.5 CPU minutes for running 20,000 iterations for a two-parameter spatial autologistic model of 2293 spins.

For large-scale problems that the dataset $\boldsymbol{Y}$ consists of a large number of observations, the difficulty arises from evaluation of the likelihood function. For the Russian Roulette sampling algorithm, its theory allows it to work with an unbiased estimate of the likelihood function. This is a strength of the Russian Roulette sampling algorithm, although achieving such an estimate at each iteration is again difficult. Sophisticated computational techniques, such as parallel computing and subsampling, may be needed there. For the AEX algorithm, to tackle this difficulty, it can be used with a split-and-merge strategy; that is, one can split the big dataset into many small subsets (e.g., through subsampling), perform AEX simulations for each subset data independently, and then combine the simulation results from each subset data to get the entire data-based inference. For

example, if we are merely interested in parameter estimation, the estimates from different subset data can be combined using a meta-analysis method (see e.g., Chang, 2011). If we are interested in posterior samples, the combination might be done via a random set intersection method, which is, to some extent, related to Dempster's rule of recombination operation in the Dempster-Shafer theory of belief functions (see e.g., Martin et al., 2010). This method has been explored in Lin (2014) for the case that the subset data are mutually independent. It would be of great interest to extend this method to the case that the subset data are generally dependent.

References

Andrews, D. and Herzberg, A. (1985), *Data*, New York: Springer.

Andrieu, C., Doucet, A., and Holenstein, R. (2010), "Particle Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society, Series B*, 72, 269–342.

Andrieu, C. and Moulines, E. (2006), "On the ergodicity properties of some adaptive MCMC algorithms," *Annals of Applied Probability*, 16, 1462–1505.

Andrieu, C., Moulines, E., and Priouret, P. (2005), "Stability of Stochastic Approximation Under Verifiable Conditions," *SIAM Journal of Control and Optimization*, 44, 283–312.

Andrieu, C. and Roberts, G. (2009), "The pseudo-marginal approach for efficient Monte Carlo computations," *Annals of Statistics*, 37, 697–725.

Atchade, Y. F., Lartillot, N., and Robert, C. P. (2013), "Bayesian computation for intractable normalizing constants," *Brazilian Journal of Statistics*, 27, 416–436.

Atchade, Y. F. and Liu, J. S. (2010), "The Wang-Landau algorithm in general state spaces: Applications and Convergence Analysis," *Statistica Sinica*, 20, 209–233.

Balram, N. and Moura, J. (1993), "Noncausal Gauss Markov random fields: parameter structure and estimation," *IEEE Trans. Inform. Theory*, 39, 1333–1355.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002), "Approximate Bayesian computation in population genetics," *Genetics*, 162, 2025–2035.

Besag, J. and Moran, P. (1975), "On the estimation and testing of spatial interaction in Gaussian lattice processes," *Biometrika*, 62, 555–562.

Besag, J. E. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Caimo, A. and Friel, N. (2011), "Bayesian inference for exponential random graph models," *Social Networks*, 33, 41–55.

Chang, M. (2011), *Modern Issues and Methods in Biostatistics*, New York: Springer.

Childs, A. M., Patterson, R. B., and MacKay, D. J. (2001), "Exact sampling from non-attractive distributions using summary states," *Physics Review E*, 63, 036113.

Everitt, R. G. (2012), "Bayesian parameter estimation for latent Markov random fields and social networks," *Journal of Computational and Graphical Statistics*, 21, 940–960.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman & Hall.

Fort, G., Moulines, E., and Priouret, P. (2011), "Convergence of adaptive and interacting Markov chain Monte Carlo algorithms," *Annals of Statistics*, 39, 3262–3289.

Gelman, A. and Rubin, D. (1992), "Inference from iterative simulation using multiple sequences (with discussion)," *Statistical Science*, 7, 457–511.

Geyer, C. J. (1991), "Markov chain Monte Carlo maximum likelihood," in *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, ed. Keramigas, E., Interface Foundation, Fairfax.

Geyer, C. J. and Thompson, E. A. (1992), "Constrained Monte Carlo maximum likelihood for dependent data," *Journal of the Royal Statistical Society, Series B*, 54, 657–699.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Gu, M. and Zhu, H. (2001), "Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation," *Journal of the Royal Statistical Society, Series B*, 63, 339–355.

Haario, H., Saksman, E., and Tamminen, J. (2001), "An adaptive Metropolis algorithm," *Bernoulli*, 7, 223–242.

Huang, F. and Ogata, Y. (1999), "Improvements of the maximum pesudo-likelihood estimators in various spatial statistical models," *Journal of Computational and Graphical Statistics*, 8, 510–530.

Hukushima, K. and Nemoto, K. (1996), "Exchange Monte Carlo method and application to spin glass simulations," *Journal of the Physical Society of Japan*, 65, 1604–1608.

Hurn, M. A., Husby, O. K., and Rue, H. (2003), *A Tutorial on Image Analysis*, vol. 173 of *Lecture Notes in Statistics*, Springer.

Jin, I. H. and Liang, F. (2014), "Use of SAMC for Bayesian Analysis of Statistical Models with Intractable Normalizing Constants," *Computational Statistics and Data Analysis*, 71, 402–416.

Liang, F. (2007), "Continuous contour Monte Carlo for marginal density estimation with an application to a spatial statistical models," *Journal of Computational and Graphical Statistics*, 16, 608–632.

— (2009), "Improving SAMC using smoothing methods: theory and applications to Bayesian model selection problems," *Annals of Statistics*, 37, 2626–2654.

— (2010), "A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants," *Journal of Statistical Computing and Simulation*, 80, 1007–1022.

Liang, F. and Jin, I. H. (2013), "A Monte Carlo Metropolis-Hastings Algorithm for Sampling from Distributions with Intractable Normalizing Constants," *Neural Computation*, 25, 2199–2234.

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

— (2010), *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*, Wiley.

Liang, F. and Wong, W. H. (2001), "Real-parameter evolutionary sampling with applications in Bayesian Mixture Models," *Journal of the American Statistical Association*, 96, 653–666.

Lin, F. (2014), "Split-and-Merge Strategies for Big Data Analysis," *Ph.D Dissertation, Department of Statistics, Texas A&M University*.

Lyne, A., Girolami, M., Atchade, Y., Strathmann, H., and Simpson, D. (2014), "Playing Russian Roulette with Intractable Likelihoods," *arXv:1306.4032v2*.

Møller, J., Pettitt, A. N., Reeves, R. W., and Berthelsen, K. K. (2006), "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants," *Biometrika*, 93, 451–459.

Murray, I., Ghahramani, Z., and MacKay, D. J. (2006), "MCMC for doubly-intractable distributions," Proceedings of 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI).

Neal, R. (1998), "Annealed importance sampling," *Statistics and Computing*, 11, 125–139.

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., and Almond, R. (2012), "Coda: Output analysis and diagnostics for Markov Chain Monte Carlo simulations," *R Package*, http://cran.r–project.org.

Preisler, H. K. (1993), "Modeling spatial patterns of trees attacked by Bark-beetles," *Applied Statistics*, 42, 501–514.

Propp, J. G. and Wilson, D. B. (1996), "Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics," *Random Structures and Algorithms*, 9, 223–252.

Riggan, W. B., Creason, J. P., Nelson, W. C., Manton, K. G., Woodbury, M. A., Stallard, E.,

Pellom, A. C., and Beaubier, J. (1987), *U.S. Cancer Mortality Rates and Trends, 1950-1979. (Vol. IV: Maps)*, U.S. Government Printing Office.: U.S. Government Printing Office.

Roberts, G. O. and Rosenthal, J. S. (2007), "Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms," *Journal of Applied Probability*, 44, 458–475.

Roberts, G. O. and Tweedie, R. L. (1996), "Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and metropolis Algorithms," *Biometrika*, 83, 95–110.

Rosenthal, J. S. (1995), "Minorization conditions and convergence rate for Markov Chain Monte Carlo," *Journal of American Statistical Association*, 90, 558–566.

Sherman, M., Apanasovich, T. V., and Carroll, R. J. (2006), "On estimation in binary autologistic spatial models," *Journal of Statistical Computation and Simulation*, 76, 167–179.

Snijders, T. A., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New specification for exponential random graph models," *Sociological Methodology*, 36, 99–153.

Table 1: Parameter estimation for the autologistic model: the estimates and their standard deviations (given in the parenthesis) were calculated based on 10 independent runs. The row of "CPU(s)" reports the CPU time (in seconds) cost by a single run of each algorithm.

	AEX	DMH	MCMH	Exchange
α	-0.3017 (9.2e-4)	-0.3020 (6.8e-4)	-0.3016 (3.9e-4)	-0.3015 (4.7e-4)
β	0.1224 (4.6e-4)	0.1227 (3.8e-4)	0.1231 (1.9e-4)	0.1229 (2.2e-4)
CPU(s)	107	142	143	106

Table 2: Parameter estimation for the simulated cancer mortality data example: The estimates and their standard deviations (given in the parenthesis) were calculated based on 10 independent runs. The column of "CPU(s)" reports the CPU time (in seconds) cost by a single run of each algorithm. * Average CPU time over 10 runs with a standard deviation of 711s.

Algorithm	Setting	α	β	CPU(s)
AEX	—	2.1×10^{-4} (3.3×10^{-4})	0.4538 (7.2×10^{-4})	327
Exchange	—	8.5×10^{-4} (4.0×10^{-4})	0.4550 (9.6×10^{-4})	5899*
DMH	$K=10$	6.8×10^{-4} (9.4×10^{-5})	0.4713 (2.0×10^{-4})	28
	$K=100$	-4.3×10^{-4} (3.1×10^{-5})	0.4647 (2.5×10^{-4})	274
	$K=500$	-2.4×10^{-4} (2.4×10^{-5})	0.4618 (1.8×10^{-4})	1362
	$K=1000$	-2.0×10^{-4} (1.3×10^{-5})	0.4600 (1.6×10^{-4})	2721
MCMH	$K=10$	4.4×10^{-4} (2.0×10^{-4})	0.4734 (3.2×10^{-4})	18
	$K=100$	-2.9×10^{-4} (7.5×10^{-5})	0.4661 (1.8×10^{-4})	89
	$K=500$	2.1×10^{-4} (4.5×10^{-5})	0.4622 (1.8×10^{-4})	282
	$K=1000$	5.0×10^{-5} (3.0×10^{-4})	0.4618 (2.4×10^{-4})	533

Table 3: Computational results for the wheat yield data: The numbers in the parentheses denote the standard error of the estimates.

Algorithm	β_h	β_v	β_d	σ^2
True Bayes	0.1014(4.0e-4)	0.3560(3.7e-4)	0.0061(1.4e-4)	0.1233(2.8e-4)
AEX	0.1022(8.0e-4)	0.3559(8.5e-4)	0.0066(1.7e-4)	0.1232(5.3e-4)

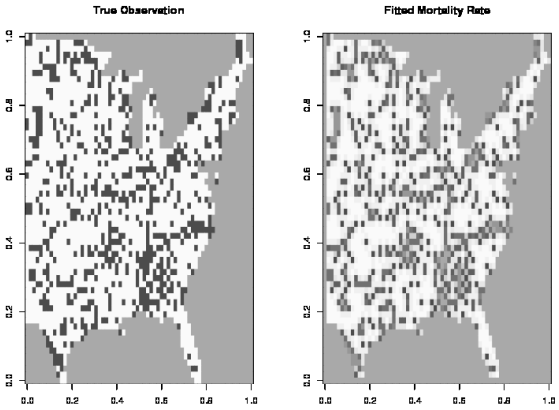

Figure 1: US cancer mortality data. Left: The mortality map of liver and gallbladder cancers (including bile ducts) for white males during the decade 1950-1959. Black squares denote counties of high cancer mortality rate, and white squares denote counties of low cancer mortality rate. Right: Estimated cancer mortality rates using the autologistic model with the model parameters being replaced by their approximate Bayesian estimates. Gray level of the corresponding square represents the cancer mortality rate of each county.

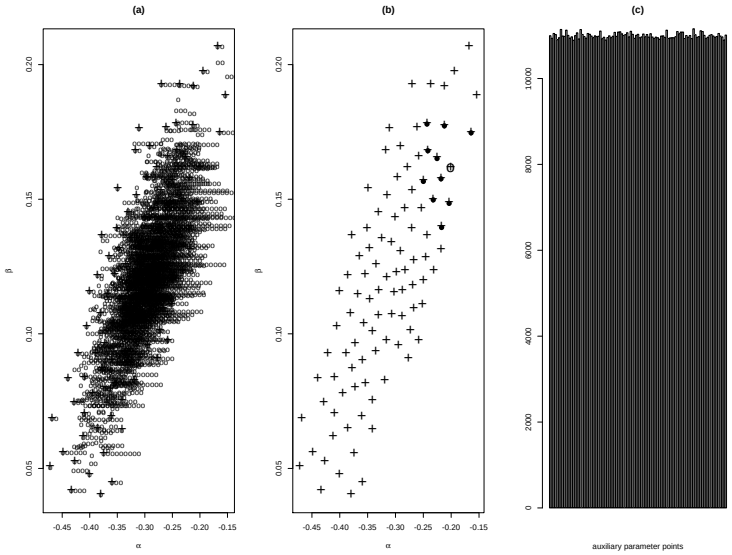

Figure 2: Auxiliary parameter set construction for the U.S. Cancer Mortality example: (a) the scatter plot of the fractional DMH samples ("o") and the selected auxiliary parameters ("+"); (b) the neighborhood (solid dots) of a selected auxiliary parameter point (big circle); and (c) the histogram of the 100 auxiliary parameters, where each bar corresponds to an auxiliary parameter.

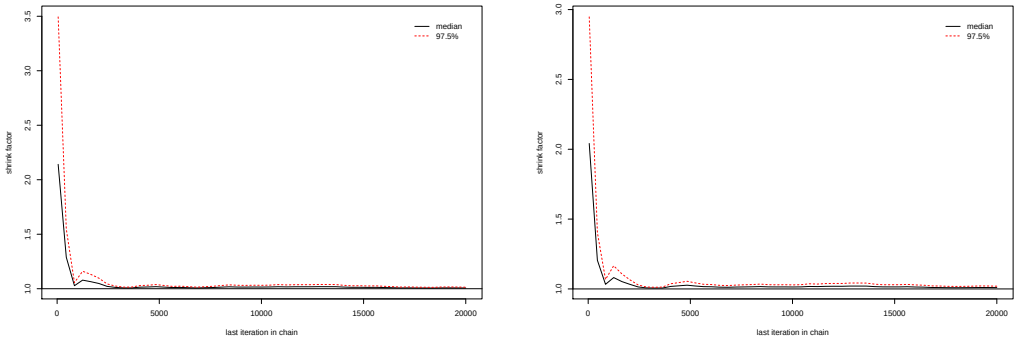

Figure 3: Convergence diagnostic of the adaptive exchange algorithm for the U.S. Cancer Mortality example: (a) Gelman-Rubin's shrink factor for the samples of α generated in 10 runs; (b) Gelman-Rubin's shrink factor for the samples of β generated in 10 runs.

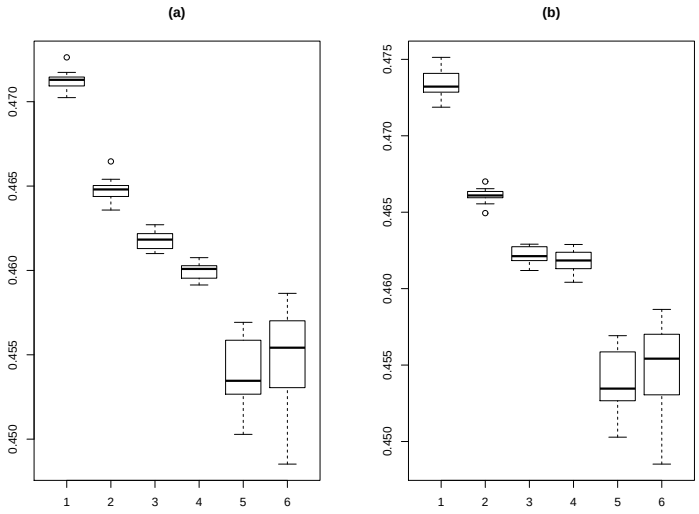

Figure 4: (a) The box-plots from left to right are for the algorithms: (1) DMH with $K = 10$; (2) DMH with $K = 100$; (3) DMH with $K = 500$; (4) DMH with $K = 1000$; (5) AEX; and (6) Exchange. (b) The box-plots from left to right are for the algorithms: (1) MCMH with $K = 10$; (2) MCMH with $K = 100$; (3) MCMH with $K = 500$; (4) MCMH with $K = 1000$; (5) AEX; and (6) Exchange.

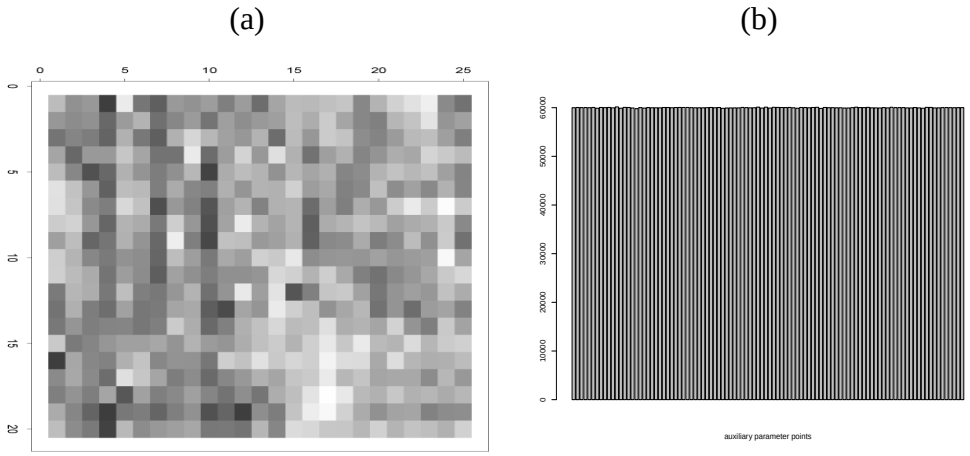

Figure 5: (a) Image of the wheat yield data, where the black and white squares denote high and low yield areas, respectively. (b) Histogram of the 100 auxiliary parameters achieved by the auxiliary chain.