Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 363–389
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Implementing componentwise Hastings algorithms☆

Richard A. Levine[a,*], Zhaoxia Yu[b], William G. Hanley[c], John J. Nitao[c]

[a]*Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA 92128, USA*
[b]*Department of Statistics, Rice University, P.O. Box 1892, Houston, TX 77251, USA*
[c]*Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94551, USA*

Received 1 February 2003; received in revised form 1 December 2003; accepted 1 December 2003

Abstract

Markov chain Monte Carlo (MCMC) routines have revolutionized the application of Monte Carlo methods in statistical application and statistical computing methodology. The Hastings sampler, encompassing both the Gibbs and Metropolis samplers as special cases, is the most commonly applied MCMC algorithm. The performance of the Hastings sampler relies heavily on the choice of sweep strategy, that is, the method by which the components or blocks of the random variable **X** of interest are visited and updated, and the choice of proposal distribution, that is the distribution from which candidate variates are drawn for the accept–reject rule in each iteration of the algorithm. We focus on the random sweep strategy, where the components of **X** are updated in a random order, and random proposal distributions, where the proposal distribution is characterized by a randomly generated parameter. We develop an adaptive Hastings sampler which learns from and adapts to random variates generated during the algorithm towards choosing the optimal random sweep strategy and proposal distribution for the problem at hand. As part of the development, we prove convergence of the random variates to the distribution of interest and discuss practical implementations of the methods. We illustrate the results presented by applying the adaptive componentwise Hastings samplers developed to sample multivariate Gaussian target distributions and Bayesian frailty models.

Keywords: Markov chain Monte Carlo; Metropolis algorithm; Gibbs sampler; Adaptive sweep strategies; Random proposal distributions

☆ Research supported in part by NSF FRG grant 0139948 and a grant from Lawrence Livermore National Laboratories, Livermore, California, USA.
* Corresponding author.
E-mail address: ralevine@sciences.sdsu.edu (R.A. Levine).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.02.002

1. Introduction

Markov chain Monte Carlo (MCMC) methods have had an extraordinary impact on the application and theoretical development of statistical methods. The Hastings sampler, which includes the Gibbs and Metropolis samplers as special cases, is the most popular implementation of MCMC. The Hastings sampler is an iterative accept–reject type routine which generates random variates $\mathbf{X} = \{X(1), \ldots, X(d)\}$ from a distribution $\pi(\mathbf{X})$ that is difficult to sample directly. The algorithm accomplishes this task by iteratively visiting each component or block of components of $\mathbf{X}$ each iteration and updating the component according to a prespecified acceptance rule. Under general regularity conditions, in the limit over the number of iterations, the random variates approach samples from the distribution $\pi(\mathbf{X})$.

The Metropolis and Gibbs samplers as introduced by Metropolis et al. (1953) and Geman and Geman (1984), respectively, suggest a *random sweep strategy* whereby the choice of components to visit at each iteration is randomly determined. Hastings (1970) also presents this strategy as part of his development of the general Hastings sampler. In the Gibbs sampler, Levine and Casella (2004) show that convergence properties of the algorithm may be significantly different depending on how often a component of $\mathbf{X}$ is visited during the algorithm. In particular, a "fair" updating scheme in which each component is visited with equal probability each iteration of the Gibbs sampler is not necessarily optimal. In this paper, we develop an adaptive Hastings sampler which chooses selection probabilities on the fly based on previous random variate generations. The algorithm may thus be automated, freeing the user from having to find the optimal set of selection probabilities.

The Hastings sampler is driven by candidate variates generated from a proposal distribution and accepted or rejected each iteration of the sampler. The choice of proposal distribution is well known to have a significant effect on convergence properties in the Hastings sampler (see Robert and Casella, 1999, Chapter 6). In particular, proposal distributions substantially flatter or more peaked than the distribution π may result in low acceptance rate and/or slow convergence of the algorithm to the distribution π. In this paper, we extend our adaptive Hastings sampler to the choice of proposal distribution through the use of the random proposal distribution of Besag et al. (1995). The random proposal distribution generalizes the adaptive rejection Metropolis sampling routine of Gilks et al. (1995). Our algorithm thus not only frees the user from choosing an appropriate proposal distribution by adaptively fitting the random proposal distribution, but provides the user with a general framework from which the proposal distribution is chosen.

As a motivational example, we consider the MCMC algorithm proposed by Clayton (1991), and further explored by Ibrahim et al. (2001, Section 4.1), for fitting a Bayesian gamma frailty model. In particular, suppose we are interested in failure time t_{ik} in relation to a set of p covariates $Z_{ik1}, \ldots, Z_{ikp}$, where $i = 1, \ldots, n$ denotes the cluster and $k = 1, \ldots, K_i$ denotes subjects within each cluster. Clayton (1991) considers a frailty model on the hazard function

$$\lambda_{ik}(\mathbf{t} \mid \mathbf{Z}_{ik}, w_i) = \lambda_0(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_{ik}) w_i,$$

$$\boldsymbol{\beta} \sim Normal\ (\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$$w_i \sim Gamma\ (\gamma^{-1}, \gamma^{-1}),$$

$$\gamma^{-1} \sim Gamma\ (\eta, \tau),$$

$$\Lambda_0 \sim Gamma\ Process(c\Lambda^*, c), \tag{1}$$

where $\lambda_0(\mathbf{t})$ is the unknown baseline hazard function, $\Lambda_0(\mathbf{t})$ is the unknown baseline cumulative hazard function, $\boldsymbol{\beta}$ is a p-vector of unknown regression parameters, w_i are independent and identically distributed frailty terms, the hierarchy presents the prior distributions on the model parameters, and the prior parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \eta, \tau, c, \Lambda^*)$ are assumed known, specified by the user/modeler.

The Markov chain induced by the Gibbs sampler for drawing Monte Carlo posterior inferences under this Bayesian frailty model suffers from an absorbing state and consequently strong autocorrelation in variance hyperparameter variates. Clayton (1991) thus proposes the following algorithm, combining the Imputation-Posterior (IP) algorithm of Tanner and Wong (1987) with the Hastings sampler.

Algorithm 1.1. Clayton sampler

(1) `Initialize` $\boldsymbol{\beta}^{(0)}, \mathbf{w}^{(0)}, \Lambda_0^{(0)}$.
(2) On the tth iteration
  a. `Draw` γ^* from $\gamma^{(t-1)}, \ldots, \gamma^{(t-B)}$.
  b. `Repeat` G times the following random variate generations: $j = 1, \ldots, G$

$$\boldsymbol{\beta}^{(j+(t-1)G)} \sim [\boldsymbol{\beta} \mid \mathbf{w}^{(j-1+(t-1)G)}, \Lambda_0^{(j-1+(t-1)G)}, \gamma^*,\ data]$$

$$\mathbf{w}^{(j+(t-1)G)} \sim [\mathbf{w} \mid \boldsymbol{\beta}^{(j+(t-1)G)}, \Lambda_0^{(j-1+(t-1)G)}, \gamma^*,\ data]$$

$$\Lambda_0^{(j+(t-1)G)} \sim [\Lambda_0 \mid \mathbf{w}^{(j+(t-1)G)}, \boldsymbol{\beta}^{(j+(t-1)G)}, \gamma^*,\ data].$$

  c. `Generate` $\gamma^{(t)} \sim [\gamma \mid \mathbf{w}^{(t \cdot G)}]$.
(3) `Repeat` step two until reaching equilibrium.

The choice of G is crucial to algorithm success: a small value of G results in a higher likelihood of the sampler wandering close to the absorbing state; a large value of G results in too few iterations between the IP steps (2b) and (2c) and slow convergence to the stationary distribution. However, Clayton (1991) proposes ad hoc specification of G following an extensive period of algorithm fine-tuning. Furthermore, sampling the conditional distributions on $\boldsymbol{\beta}$ and γ in step two is complicated by the model structure. Clayton (1991) proposes a Gaussian approximation to the $\boldsymbol{\beta}$ full conditional distribution and an accept–reject routine for sampling from the γ full conditional distribution. Each of these simulation algorithms requires substantial front-end work and coding to find appropriate approximating and proposal distributions for sampling these conditional distributions, ideally performed each iteration of the algorithm.

We propose to use the adaptive componentwise Hastings sampler to allow the computer to perform the algorithm fine-tuning, in the process determining optimal choice of G through the selection probabilities of the random sweep strategy, and optimal proposal distributions for sampling from the full conditional distributions on $\boldsymbol{\beta}$ and γ, through the random proposal distributions. The algorithm thus requires significantly less user coding time. Furthermore, though the computer is doing the front-end algorithm implementation work, we will show that the adaptive componentwise Hastings sampler is no more computationally expensive, in terms of sampler run time, than Algorithm 1.1.

As suggested by the Bayesian frailty model example, our motivation for developing the adaptive componentwise Hastings sampler is to minimize coding cost in implementing and fine-tuning the algorithm, a significant time-sink for the coder of Hastings samplers. We note that in this paper we present the general methodology of the adaptive componentwise Hastings sampler. Within this flexible framework, the user is afforded potentially three additional luxuries depending on implementation strategy. First, the adaptive scans have the potential to speed up mixing in the Hastings sampler needing fewer iterations for precise inferences based on the random variates generated. The adaptive routines proposed are not significantly slower than the standard implementation of the Hastings sampler thus providing a potentially substantial improvement in computational cost when using the adaptive schemes. Second, on a related point, the adaptive scheme is optimal in the sense of providing the best implementation of the Hastings sampler with respect to the objective function of choice (we will discuss, at least briefly, estimator precision, convergence rate, acceptance rates, and computational cost in subsequent sections) for comparing sweep strategies. Third, the adaptive routine may be automated both in terms of choice of candidate distribution, as mentioned above, and in choice of updating scheme. The routine is also modular allowing for much flexibility in choice of implementation and ease in code manipulations for users familiar with the inner workings of the MCMC sampling schemes.

The paper unfolds as follows. In Section 2, we formally define the componentwise Hastings algorithm, the basis of the algorithms constructed in the remainder of the paper. In Sections 3 and 4, we introduce the adaptive componentwise Hastings sampler to address the implementation questions of choosing selection probabilities and proposal distributions. This algorithm sets up a Hastings sampler that learns from and adapts to the random variates generated as the sampler proceeds to the distribution of interest π. As part of the development, we detail the necessary convergence theory of the adaptive Hastings samplers presented and relate the general scheme to the more specific cases of the Gibbs and Metropolis samplers. In Section 5, we illustrate the potential of our algorithms in the simple case of generating Gaussian random variates. In Section 6, we return to the Bayesian frailty model of Clayton, comparing our algorithms to Algorithm 1.1 in an animal carcinogenesis problem. Section 7 further discusses implementation of our algorithm relative to computational cost, convergence rate, and algorithm efficiency. All computations are performed in Matlab 6.5 on a 900 MHz Sparc SunBlade 2000 with 3.0 GB of RAM.

2. Hastings algorithm

MCMC routines may be broadly classified as Hastings samplers with the Metropolis and Gibbs samplers being special cases. The various implementations of the Hastings samplers differ in the means through which they iteratively sample the components of $\mathbf{X}$ and in turn induce a Markov chain with stationary distribution $\pi(\mathbf{X})$. This difference may be characterized by the random process or mechanism through which we choose the components of $\mathbf{X}$ to update each iteration of the MCMC algorithm.

The Hastings sampler is an accept–reject routine which induces a Markov chain by generating candidate random variates from a transition probability q. The algorithm is as follows.

Algorithm 2.1. Hastings sampler

(1) `Select` an initial point $\mathbf{X}^{(0)} = \{X^{(0)}(1), \ldots X^{(0)}(d)\} \sim \mu_0$ according to some initial distribution μ_0.
(2) On the tth iteration
   a. `Generate` $\mathbf{X}^* \sim q(\mathbf{X} \mid \mathbf{X}^{(t-1)})$.
   b. `Compute`

$$\rho(\mathbf{X}^{(t-1)}, \mathbf{X}^*) = \min\left\{1, \frac{\pi(\mathbf{X}^*)\, q(\mathbf{X}^{(t-1)} \mid \mathbf{X}^*)}{\pi(\mathbf{X}^{(t-1)})\, q(\mathbf{X}^* \mid \mathbf{X}^{(t-1)})}\right\}.$$

   c. `Take`

$$\mathbf{X}^{(t)} = \begin{cases} \mathbf{X}^* & \text{with probability } \rho(\mathbf{X}^{(t-1)},\ \mathbf{X}^*) \\ \mathbf{X}^{(t-1)} & \text{with probability } 1 - \rho(\mathbf{X}^{(t-1)},\ \mathbf{X}^*). \end{cases}$$

(3) `Repeat` step two until reaching equilibrium.

In words, at iteration t, the Hastings sampler in Algorithm 2.1 chooses a sample $\mathbf{X}^*$ from the proposal distribution q, which may depend on the last variate sampled, $\mathbf{X}^{(t-1)}$. The algorithm chooses to accept this new random variate, $\mathbf{X}^*$, if it, in a sense, is more likely to come from the stationary distribution than the old sample, as determined by the probability ρ. The Hastings sampler induces a Markov chain with stationary distribution $\pi(\mathbf{X})$, despite the arbitrary choice of proposal distribution q. In practice, we choose a q which is easy to sample yet close to the stationary distribution π. See Robert and Casella (1999) for details.

The original application of the Hastings algorithm in physics and image processing considers componentwise updates of $\mathbf{X}$ using component specific proposal distributions q_i, $i = 1, \ldots, d$. At iteration t, $q_i(\mathbf{X} \mid \mathbf{X}^{(t-1)})$ generates a sample for component i, $X^*(i)$, and leaves all other components the same so that $\mathbf{X}^*_{-i} = \mathbf{X}^{(t-1)}_{-i}$ (Besag et al., 1995).

Algorithm 2.2. Componentwise Hastings sampler

(1) `Initialization`
   a. `Select` an initial point $\mathbf{X}^{(0)} = \{X^{(0)}(1), \ldots X^{(0)}(d)\} \sim \mu_0$ according to some initial distribution μ_0.
   b. `Choose` selection probabilities $\boldsymbol{\alpha} = \{\alpha_1, \ldots \alpha_d\}$.

(2) On the tth iteration
   a. `Randomly choose` $i \in \{1, \ldots, d\}$ with probability α_i.
   b. `Generate` $\mathbf{X}^* \sim q_i(\mathbf{X} \mid \mathbf{X}^{(t-1)})$.
   c `Compute`

$$\rho_i(\mathbf{X}^{(t-1)}, \mathbf{X}^*) = \min\left\{1, \frac{\pi(X^*(i) \mid \mathbf{X}_{-i}^{(t-1)}) \cdot q_i(\mathbf{X}^{(t-1)} \mid \mathbf{X}^*)}{\pi(X^{(t-1)}(i) \mid \mathbf{X}_{-i}^{(t-1)}) \cdot q_i(\mathbf{X}^* \mid \mathbf{X}^{(t-1)})}\right\}.$$

   d. `Take`

$$\mathbf{X}_i^{(t)} = \begin{cases} \mathbf{X}^* & \text{with probability } \rho_i(\mathbf{X}^{(t-1)}, \mathbf{X}^*) \\ \mathbf{X}^{(t-1)} & \text{with probability } 1 - \rho_i(\mathbf{X}^{(t-1)}, \mathbf{X}^*). \end{cases}$$

(3) `Repeat` step two until reaching equilibrium.

Note that the componentwise Hastings sampler in Algorithm 2.2 randomly chooses the component to update each iteration. The transition function is a weighted sum of individual transition kernels $P(\mathbf{Y} \mid \mathbf{X}; \boldsymbol{\alpha}) = \sum_{i=1}^{d} \alpha_i P_i$ where

$$P_i = \rho_i(\mathbf{Y}, \mathbf{X}) q_i(\mathbf{Y} \mid \mathbf{X}) + \{1 - r_i(\mathbf{X})\}\delta_{Y(i)}(\mathbf{X}), \tag{2}$$

$r_i(\mathbf{X}) = \int \rho_i(\mathbf{X}, \mathbf{Y}) q_i(\mathbf{Y} \mid \mathbf{X}) \, \mathrm{d}Y(i)$, and δ_x denotes the Dirac mass in x. This result follows from step two of Algorithm 2.2 since any component may be updated during a given iteration. The weights, α_i, are the probabilities a component i is visited during a given iteration.

3. Adaptive scan Hastings algorithms

The componentwise Hastings sampler in Algorithm 2.2 is characterized by a set of selection probabilities $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$. These probabilities determine the percentage of visits to a specific site or component of the d-vector of interest $\mathbf{X} = \{X(1), \ldots, X(d)\}$ during a run of the sampler. Intuitively, the more difficult a given marginal is to understand or describe, the more often we should visit that component in our sweep strategy. Our goal in this section is to construct algorithms and methodologies for choosing selection probabilities in the componentwise Hastings sampler.

In this section, we introduce the adaptive scan Hastings sampler which chooses selection probabilities as the algorithm proceeds based on all random variates previously generated. The presentation is completely general to allow the user much flexibility in the application, particularly in the choice of distribution from which to choose the selection probabilities. Suggested implementation of these theoretical results will be introduced and discussed in Sections 5–7.

3.1. Adaptive sweep strategy

Assume we are interested in generating random variates $\mathbf{X}$ from a d-dimensional distribution $\pi(\mathbf{X})$. Consider sampling from the distribution $\pi(\mathbf{X})$ using the componentwise Hastings sampler. This sampler is characterized by the selection probabilities $\boldsymbol{\alpha}$. Rather than preselecting $\boldsymbol{\alpha}$ and leaving the values fixed throughout the algorithm, we consider an adaptive scheme which chooses $\boldsymbol{\alpha}$ each iteration of the sampler depending on the states visited during the routine. Therefore, both $\boldsymbol{\alpha}$ and $\mathbf{X}$ are updated every step of the sampler. Assume q_i is the proposal distribution for component $i = 1, \ldots, d$. As in Algorithm 2.2, these candidate distributions propose a replacement for $X(i)$, but leave the remaining components unchanged.

Algorithm 3.1. Adaptive componentwise Hastings sampler

(1) `Initialization`
   a. `Select` an initial point $\mathbf{X}^{(0)} = \{X^{(0)}(1), \ldots X^{(0)}(d)\} \sim \mu_0$ according to some initial distribution μ_0.
   b. `Choose` selection probabilities $\boldsymbol{\alpha}^{(0)} = \{\alpha_1^{(0)}, \ldots \alpha_d^{(0)}\}$.

(2) On the tth iteration
   a. `Choose` $\boldsymbol{\alpha}^{(t)}(\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \sim f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)})$.
   b. `Randomly choose` $i \in \{1, \ldots, d\}$ with probability $\alpha_i^{(t)}(\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)})$.
   c. `Generate` $\mathbf{X}^* \sim q_i(\mathbf{X} \mid \mathbf{X}^{(t-1)})$.
   d. `Compute`

$$\rho_i(\mathbf{X}^{(t-1)}, \mathbf{X}^*) = \min\left\{1, \frac{\pi(X^*(i) \mid \mathbf{X}_{-i}^{(t-1)}) \cdot q_i(\mathbf{X}^{(t-1)} \mid \mathbf{X}^*)}{\pi(X^{(t-1)}(i) \mid \mathbf{X}_{-i}^{(t-1)}) \cdot q_i(\mathbf{X}^* \mid \mathbf{X}^{(t-1)})}\right\}.$$

   e. `Take`

$$\mathbf{X}^{(t)} = \begin{cases} \mathbf{X}^* & \text{with probability } \rho_i(\mathbf{X}^{(t-1)}, \mathbf{X}^*) \\ \mathbf{X}^{(t-1)} & \text{with probability } 1 - \rho_i(\mathbf{X}^{(t-1)}, \mathbf{X}^*). \end{cases}$$

(3) `Repeat` step two until reaching equilibrium.

The scheme is called "adaptive" since the selection probabilities are constructed from a function $f_{\boldsymbol{\alpha}}$ which, as the chain traverses, is updated using all random variates generated. In the examples of Sections 5 and 6, we will apply and illustrate a variety of choices for the function $f_{\boldsymbol{\alpha}}$. In Section 7, we will discuss strategies for choosing $f_{\boldsymbol{\alpha}}$ in practice depending on user goals both for the sampler and subsequent statistical inferences.

Unfortunately, this routine for choosing $\boldsymbol{\alpha}$ induces a chain $\{\mathbf{X}^{(i)}\}_{i=1}^n$ which violates the Markov property. In particular, the distribution of $\mathbf{X}^{(t)}$ at a given iteration t of the

sampler is

$$\mu_t(A) \equiv P_{AHS}(\mathbf{X}^{(t)} \in A)$$
$$= \int_A \int_{(0,1)^d} \int \sum_{i=1}^{d} \alpha_i P_i(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}) f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)})$$
$$\mu_0(\mathrm{d}\mathbf{X}^{(0)}) \cdots \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)})\, \mathrm{d}\boldsymbol{\alpha}\, \mathrm{d}\mathbf{X}^{(t)} \tag{3}$$

for all measurable sets A. Here $P_i(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)})$ is as defined in (2).

Nonetheless, convergence in total variation norm is still obtained.

Theorem 3.1. *Suppose that*

a. $f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \to f(\boldsymbol{\alpha})$ *almost everywhere*; *and*
b. *the componentwise Hastings sampler with fixed selection probability distribution* $f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ *induces an ergodic Markov chain with stationary distribution* π.

Then $\|\mu_t - \pi\|_{TV} \to 0$ *as* $t \to \infty$.

The formal proof may be found in the Appendix. For purposes of intuition here, in an adaptive componentwise Hastings sampler, as t approaches infinity, the selection probabilities $\boldsymbol{\alpha}^{(t)}(\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)})$ are converging to $\boldsymbol{\alpha}$ almost surely. Hence, the adaptive chain $\{\mathbf{X}^{(i)}\}_{i=0}^{n}$ is coupling with a componentwise Hastings sampler with selection probabilities $\boldsymbol{\alpha}$. The ergodic theorem for the componentwise Hastings sampler of Algorithm 2.2 may thus be applied. The almost sure convergence assumed in condition (a) is with respect to the canonical probability over the distributions μ_j of $\mathbf{X}^{(j)}$, $j = 1, \ldots, t-1$.

3.2. Metropolis and Gibbs samplers

The original MCMC method of Metropolis et al. (1953) is a componentwise Hastings sampler with a symmetric q_i so that

$$\rho_i(\mathbf{X}^{(t-1)}, \mathbf{X}^*) = \min\{1,\ \pi(X^*(i) \mid \mathbf{X}_{-i}^{(t-1)}) / \pi(X^{(t-1)}(i) \mid \mathbf{X}_{-i}^{(t-1)})\}.$$

In fact, application of this *Metropolis* algorithm in physics uses random scan updating (e.g., see Besag, 2000). The random scan Gibbs sampler, the Gibbs sampler presented in Geman and Geman (1984), is a componentwise Hastings sampler with $q_i(\mathbf{X}^* \mid \mathbf{X}^{(t-1)}) = \pi(X^*(i) \mid \mathbf{X}_{-i})$, so $\rho_i(\mathbf{X}^{(t-1)},\ \mathbf{X}^*) = 1$; that is, we always accept. The adaptive componentwise Hastings sampler thus subsumes the Metropolis and Gibbs samplers. In particular, we may choose selection probabilities in either of these samplers adaptively using Algorithm 3.1.

4. Random proposal distributions

The adaptive componentwise Hastings algorithm adapts to the random variates generated by changing the sweep strategy through the selection probabilities $\boldsymbol{\alpha}$. As an

alternative or complementary approach, we may adapt the algorithm to the random variates generated through the proposal distribution q in Algorithms 2.2 and 3.1. Such adaptive schemes have been considered previously. The adaptive rejection Metropolis sampling (ARMS) of Gilks et al. (1995) adaptively updates a "pseudo-envelope" density of π using an accept–reject step (see Robert and Casella, 1999, for a clear description and details). Haario et al. (1999, 2001) adaptively update parameters in a Gaussian proposal distribution q. Holden (1998) adaptively updates proposal step functions based on nearest-neighbor evaluations of the stationary distribution π.

Besag et al. (1995) introduces random proposal distributions which, like the adaptive schemes of Haario et al. (1999, 2001), adaptively update parameters of the proposal distribution q. However, the scheme is not restricted to Gaussian proposal distributions. Furthermore, the random proposal distributions contain ARMS as a special case (Besag et al., 1995). We will show in this section how to adaptively update the random proposal distribution using ideas from the adaptive componentwise Hastings sampler of Algorithm 3.1.

As stated, random proposal distributions characterize the proposal distribution q through an unknown parameter γ. For each component $i = 1,\ldots,d$, we thus have a class of proposal distributions, $q_i(\mathbf{X} \mid \mathbf{Y}; \gamma)$ indexed by this parameter γ.

Algorithm 4.1. Adaptive componentwise Hastings sampler using random proposal distributions

(1) `Initialization`
  a. `Select` an initial point $\mathbf{X}^{(0)} = \{X^{(0)}(1),\ldots X^{(0)}(d)\} \sim \mu_0$ according to some initial distribution μ_0.
  b. `Choose` selection probabilities $\boldsymbol{\alpha}^{(0)} = \{\alpha_1^{(0)},\ldots \alpha_d^{(0)}\}$.
  c. `Select` parameter $\gamma^{(0)}$.

(2) On the tth iteration
  a. `Choose` $\gamma^{(t)}(\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) \sim g_\gamma(\gamma^{(t)} \mid \gamma^{(t-1)}, \mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)})$.
  [b. `Choose` $\boldsymbol{\alpha}^{(t)}(\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) \sim f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)})$.
  c. `Randomly choose` $i \in \{1,\ldots,d\}$ with probability $\alpha_i^{(t)}(\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)})$.
  d. `Generate` $\mathbf{X}^* \sim q_i(\mathbf{X} \mid \mathbf{X}^{(t-1)}; \gamma^{(t)})$.
  e. `Compute`

$$\rho_i(\mathbf{X}^{(t-1)}, \mathbf{X}^*) = \min\left\{1, \frac{\pi(X^*(i) \mid \mathbf{X}_{-i}^{(t-1)}) \cdot q_i(\mathbf{X}^{(t-1)} \mid \mathbf{X}^*, \gamma^{(t)})}{\pi(X^{(t-1)}(i) \mid \mathbf{X}_{-i}^{(t-1)}) \cdot q_i(\mathbf{X}^* \mid \mathbf{X}^{(t-1)}, \gamma^{(t)})}\right\}.$$

  f. `Take`

$$\mathbf{X}^{(t)} = \begin{cases} \mathbf{X}^* & \text{with probability } \rho_i(\mathbf{X}^{(t-1)},\ \mathbf{X}^*) \\ \mathbf{X}^{(t-1)} & \text{with probability } 1 - \rho_i(\mathbf{X}^{(t-1)},\ \mathbf{X}^*). \end{cases}$$

(3) `Repeat` step 2 until reaching equilibrium.

The distribution g_γ is chosen by the user for drawing the parameter γ each iteration of the sampler. Again, the chain induced by the sampler in Algorithm 4.1 is *not* Markov,

even if we eliminated step 2b and fixed $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(0)}$ for all iterations t. The distribution of $\mathbf{X}^{(t)}$ at a given iteration t of the sampler is similar to that of (3)

$$\begin{aligned}\mu_t(A) &\equiv P_{AHSRP}(\mathbf{X}^{(t)} \in A)\\ &= \int_A \int \int_{(0,1)^d} \int \sum_{i=1}^{d} \alpha_i P_i(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \gamma^{(t)}) f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)})\\ &\quad g_{\gamma}^{(t)}(\gamma \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \mu_0(\mathrm{d}\mathbf{X}^{(0)}) \cdots \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)})\, \mathrm{d}\boldsymbol{\alpha}\, \mathrm{d}\gamma\, \mathrm{d}\mathbf{X}^{(t)} \qquad (4)\end{aligned}$$

for all measurable sets A. Here $P_i(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \gamma)$ is as defined in (2) though the proposal distribution q_i is characterized by the parameter γ.

An argument similar to that in Section 3.1 gives us convergence in total variation norm. Details of the proof may be found in Appendix A.

Theorem 4.1. *Suppose that*

a. $f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \to f(\boldsymbol{\alpha})$ *almost everywhere*;
b. $g_{\gamma}^{(t)}(\gamma \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \to g_{\gamma}(\gamma)$ *almost everywhere*;
c. *the componentwise Hastings sampler with fixed selection probability distribution* $f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ *and fixed random proposal parameter distribution* $g_{\gamma}(\gamma)$ *induces an ergodic Markov chain with stationary distribution* π; *and*
d. *the proposal distribution* $q(\mathbf{Y} \mid \mathbf{X}; \gamma)$ *is chosen to be stochastically equicontinuous in* γ: for $\varepsilon > 0$, *there exists* $\delta > 0$ *such that* $|\gamma_1 - \gamma_2| < \delta$ *implies* $|q(\mathbf{Y} \mid \mathbf{X}, \gamma_1) - q(\mathbf{Y} \mid \mathbf{X}, \gamma_2)| < \varepsilon$ *almost everywhere on the support of* π.

Then $\|\mu_t - \pi\|_{TV} \to 0$ *as* $t \to \infty$.

In Section 5 we consider choosing an optimal set of selection probabilities and proposal parameters from fixed rules, rather than randomly sampling $\boldsymbol{\alpha}$ and γ each iteration from distributions $f_{\boldsymbol{\alpha}}$ and g_{γ}, respectively. The following theorem, proved analogously to Theorem 4.1, fits this framework in which concern is over convergence in $\boldsymbol{\alpha}^{(t)}$ and $\gamma^{(t)}$ rather than $f_{\boldsymbol{\alpha}}^{(t)}$ and $g_{\gamma}^{(t)}$.

Theorem 4.2. *Suppose that*

a. $\boldsymbol{\alpha}^{(t)}(\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \to \boldsymbol{\alpha}$ *almost everywhere for fixed* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in (0,1)^d$, $\sum_{i=1}^{d} \alpha_i = 1$;
b. $\gamma^{(t)}(\mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \to \gamma$ *almost everywhere for fixed* γ;
c. *the componentwise Hastings sampler with fixed selection probabilities* $\boldsymbol{\alpha}$ *and fixed random proposal parameters* γ *induces an ergodic Markov chain with stationary distribution* π; *and*
d. *the proposal distribution* $q(\mathbf{Y} \mid \mathbf{X}; \gamma)$ *is chosen to be stochastically equicontinuous in* γ: *for* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that* $|\gamma_1 - \gamma_2| < \delta$ *implies* $|q(\mathbf{Y} \mid \mathbf{X}, \gamma_1) - q(\mathbf{Y} \mid \mathbf{X}, \gamma_2)| < \varepsilon$ *almost everywhere on the support of* π.

Then $\|\mu_t - \pi\|_{TV} \to 0$ *as* $t \to \infty$.

Conditions (a) and (b) of Theorem 4.2 assume that the selection probabilities and random proposal parameter at iteration t are a function of the variates $\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}$, though converging, respectively, to a fixed set of selection probabilities $\boldsymbol{\alpha}$ and parameter γ. The almost sure convergence assumed is again with respect to the canonical probability over the distributions μ_j of $\mathbf{X}^{(j)}$, $j=1,\ldots,t-1$.

5. Gaussian target distributions

In this section, we illustrate the adaptive componentwise Hastings sampler of Algorithm 3.1 on Gaussian target distributions in a similar spirit to Haario et al. (1999). The goal is to show, in a simple simulation problem, the potential of our adaptive routines on three fronts: (1) automated adaptive implementation of the componentwise Hastings sampler, (2) improved precision in post-processing of Monte Carlo samples, (3) improved convergence properties of the induced chain.

We consider two d-dimensional target distributions: Gaussian distribution with dispersion matrix $\boldsymbol{\Sigma}$, denoted $N_d(\mathbf{0},\ \boldsymbol{\Sigma})$, and the "banana-shaped" distribution being the composition of the d-dimensional Gaussian distribution $N_d(\mathbf{0},\ \boldsymbol{\Sigma})$ and the function $\eta_b=(x_1,x_2+bx_1^2-100b,x_3,\ldots,x_d)$. The banana-shaped distribution twists the d-variate Gaussian distribution by changing the second component according to some constant $b>0$. For the numerical study in this section, we will consider three-dimensional random variates ($d=3$) for simplicity in the following three cases. Let $diag(a_1,\ldots,a_d)$ denote a $d\times d$ diagonal matrix with diagonal elements $\{a_1,\ldots,a_n\}$ and $\mathbf{J}_d$ denote a $d\times d$ matrix of ones.

(1) Unimodal Gaussian distribution: $N_d(\mathbf{0},\ \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}=diag(100,10,1)-\mathbf{J}/8$.
(2) Non-linear "banana-shaped" distribution: $F_b(\mathbf{X})=F\circ\eta_b(\mathbf{X})$ with $b=0.03$, where F is the Gaussian distribution in case one.
(3) Bimodal Gaussian distribution: $0.5*N(\boldsymbol{\mu}_1,\ \boldsymbol{\Sigma}_1)+0.5*N(\boldsymbol{\mu}_2,\ \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\mu}_1=(-1.5,1.5,1.5)'$, $\boldsymbol{\mu}_2=(1.5,1.5,1.5)'$, and $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$ having variances (diagonal) $(10,5,1)$ and covariances $\sigma_{12}=\sigma_{23}=0.5$ and $\sigma_{13}=0.25$.

We consider estimating a function $h(\mathbf{X})$ where $\mathbf{X}$ follows the distribution specified in each of the three cases. We use the optimal, in terms of mixing of the induced Hastings chain, proposal distribution suggested by Gelman et al. (1996), namely $(2.4/\sqrt{3})N(\mathbf{0},\mathbf{I}_3)$ where $\mathbf{I}_d$ is a $d\times d$ identity matrix. We use the quasi-adaptive componentwise Hastings sampler suggested at the end of Section 7.1 where we first burn in the sampler for $50,000$ iterations under equal selection probabilities, $\boldsymbol{\alpha}=\{1/3,1/3,1/3\}$. We then run the adaptive componentwise Hastings algorithm for $60,000$ iterations, updating the selection probabilities every M iterations; M will be defined below.

For purposes of illustration, the choice of the selection probabilities is performed through the post-processing decision theoretic routine of Levine and Casella (2004). Briefly, suppose we generate samples $\{\mathbf{X}^{(i)}\}_{i=0}^{n}$ from our Hastings algorithm and wish to estimate $\mu=E_\pi\{h(\mathbf{X})\}$ for some function $h\in L^2(\pi)$. A natural Monte Carlo estimator for μ is the sample mean $\hat{\mu}=(1/t)\sum_i h(\mathbf{X})^{(i)}$. With respect to this estimator, the best

scanning strategy chooses selection probabilities such that the asymptotic risk

$$R(\boldsymbol{\alpha}, h) = \lim_{t\to\infty} t\ VAR_{\pi}(\hat{\mu}) = VAR_{\pi}\{h(\mathbf{X})\} + 2\sum_{i=1}^{\infty} cov_{\pi}\{h(\mathbf{X}^{(0)}),\ h(\mathbf{X}^{(i)})\} \tag{5}$$

is at a minimum. We are thinking about the sweep strategy, characterized by $\boldsymbol{\alpha}$, as the decision rule or estimator for a parameter of interest h.

Levine and Casella (2004) show that the covariance lags in (5) may be broken down further as explicit functions of $\boldsymbol{\alpha}$, developing a minimax strategy for choosing the selection probabilities. Since interest here lies in a specified function $h(\mathbf{X})$, we implement step (2a) of Algorithm 3.1 through the following substeps:

i `Estimate` $E(h(\mathbf{X}) \mid \mathbf{X}_{-i})$ for each $i = 1, 2, 3$ using all the previous samples

$$\hat{E}(h(\mathbf{X}) \mid \mathbf{X}_{-i}) = \frac{1}{t}\left\{\sum_{j=1}^{t} h(\mathbf{X}^{(j)})\right\}.$$

ii `Minimize` over $\boldsymbol{\alpha}$ the risk function

$$R(\boldsymbol{\alpha}, h) = \hat{VAR}_{\pi}\{h(\mathbf{X})\} + 2\hat{cov}_{\pi}\{h(\mathbf{X}^{(0)}), h(\mathbf{X}^{(1)})\} + 2\hat{cov}_{\pi}\{h(\mathbf{X}^{(0)}), h(\mathbf{X}^{(2)})\}$$

for the function $h(\mathbf{Z})$ of interest with respect to the constraint $\sum_{i=1}^{d} \alpha_i = 1$ to obtain new selection probabilities $\hat{\boldsymbol{\alpha}}$.

iii `Update` component i with probability $\hat{\alpha_i}$. Specifically, update $X_i^{(t)}$ from the conditional distribution $[X_i \mid \mathbf{X}_{-i}]$, leaving all other components of $\mathbf{X}$ the same (this is the componentwise Hastings sampler with selection probabilities $\hat{\boldsymbol{\alpha}}$).

iv `Repeat` steps i–iii until reach equilibrium.

The hats over the covariances and variances in step ii denote the use of the estimated conditional expectations from step i in calculation of the risk in step ii. The minimization in step ii is a cheap computation since the second order approximation to the asymptotic risk function is quadratic in $\boldsymbol{\alpha}$ and we are performing the minimization for a single function of interest $h(\mathbf{X})$. Of course, we may use additional terms in the expansion; but our experience shows that the second order expansion is sufficient and allows for quick computation of the selection probabilities.

Alternative criteria for choosing selection probabilities, based on post-processing Monte Carlo samples as performed here or other properties of the sampler, are available to the user depending on the goals and problem at hand. We will discuss these strategies and a means of approaching such methods in Section 7.

Table 1 presents the limiting selection probabilities from the adaptive componentwise Hastings algorithm. Note that the selection probabilities are significantly different than the commonly applied equal selection probability random scan Hastings sampler. In each case studied, an iteration of the random scan requires 0.02 s. Computation and minimization of the risk function requires 0.05 s. Therefore, the adaptive scan algorithm is not significantly slower than the standard random scan with fixed, pre-determined selection probabilities.

We discuss each of these three examples in turn. The performance evaluations consider the criteria suggested by Haario et al. (1999). In particular, over 100 repetitions

Table 1
Selection probabilities and risk (asymptotic variance) for the adaptive componentwise Hastings sampler for sampling from the three target distributions: $N_d(\mathbf{0},\ \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = diag(100, 10, 1) - \mathbf{J}/8$; $0.5 * N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.5 * N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\mu}_1 = (-1.5, 1.5, 1.5)'$, $\boldsymbol{\mu}_2 = (1.5, 1.5, 1.5)'$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ having variances (diagonal) $(10, 5, 1)$ and covariances $\sigma_{12} = \sigma_{23} = 0.5$ and $\sigma_{13} = 0.25$; and $F_b = F \circ \eta_b$, where F is the bimodal Gaussian distribution, $\eta_b = (x_1, x_2 + bx_1^2 - 100b, x_3)$, and $b = 0.03$. Last column presents the risk reduction relative to the random scan with equal selection probabilities

Example	α_1	α_2	α_3	Risk reduction
One-mode	0.70	0.15	0.15	46%
"Banana-shaped"	0.64	0.15	0.21	36%
Bi-mode	0.66	0.15	0.19	24%

of each of the simulation experiments, we present mean distance and standard deviation of the distances from the true value (MSE and Std MSE), mean error (abs error) and standard deviation (std) of the percentages of variates falling below the first and 68th percentiles, within the 68th and 95th percentiles, and within the 95th and 99th percentiles.

5.1. Unimodal Gaussian target distribution

We first consider estimating the mean of $\mathbf{X}$ so that $h(\mathbf{X}) = \{X(1) + X(2) + X(3)\}/3$. As shown in Table 1, the random scan with optimal selection probabilities with respect to the risk function criterion outperforms the random scan with equal selection probabilities, reducing the risk by 46%. Consequently, statistical inferences from the optimal random scan are more precise as seen in the performance criteria presented in Table 2. Notice that even in the random variate generation in this simple Gaussian case, the adaptive random scan outperforms the random scan with equal selection probabilities with respect to all criteria. Though not presented for brevity of presentation, the autocorrelation plot for simulations from the random scan with optimal and equal selection probabilities both indicate well mixing chains with approximately uncorrelated variates after lag five. However, the optimal random scan Gibbs sampler displays significantly smaller autocorrelations at lags two and three than the random scan with equal selection probabilities.

We next considered estimating the nonlinear functions in $\mathbf{X}$, $h(\mathbf{X}) = \{X(1)^2 + X(2) + X(3)\}/3$ and $h(\mathbf{X}) = \sqrt{|X(1) + X(2) + X(3)|}$. The results are similar to that in Table 2 and thus are not shown here for brevity of presentation. Of particular note is that though both chains mix very well, the optimal random scan mixes faster, in fact providing approximately uncorrelated variates as compared to the random scan with equal selection probabilities which provides approximately uncorrelated variates after a lag of five.

Finally, we considered implementing the Hastings algorithm with a poor candidate distribution, namely a distribution with half the standard deviation of the candidate distribution recommended by Gelman et al. (1996). We return to estimating the mean $h(\mathbf{X}) = \{X(1) + X(2) + X(3)\}/3$. The relative performance of the random scan with

Table 2
Unimodal Gaussian target distribution: Summary statistics for comparison of random scan with optimal selection probabilities from the adaptive Hastings sampler and random scan Hastings sampler with equal selection probabilities. Random variates generated from the unimodal Gaussian target distribution and banana-shaped twisted Gaussian target distribution with function of interest $h(\mathbf{X}) = \{X(1) + X(2) + X(3)\}/3$

Criteria	Unimodal Opt sel probs	Equal sel probs	Banana Opt sel probs	Equal sel probs
MSE	0.2192	0.2671	1.2538	1.4179
Std MSE	0.1053	0.1734	2.2417	2.5698
std 68 prctile	0.0122	0.0138	0.0451	0.0514
abs error 68 prctile	0.0099	0.0111	0.0248	0.0290
std one prctile	0.0022	0.0027	0.0742	0.0797
abs error one prctile	0.0021	0.0021	0.9762	0.9762
std 68–95%	0.0106	0.0132	0.0456	0.0497
abs error 68–95%	0.0089	0.0106	0.0256	0.0285
std 95–99%	0.0048	0.0053	0.0291	0.0446
abs error 95–99%	0.0039	0.0042	0.0234	0.0314

optimal and equal selection probabilities is analogous to the other illustrations in this subsection and thus not presented for brevity. Of particular note is that both scans mix slowly, though, unlike the optimal random scan, the random scan with equal selection probabilities shows significant autocorrelation even after a lag of 1000.

The random scan with selection probabilities chosen by minimizing the asymptotic variance thus outperforms the random scan with equal selection probabilities in terms of mixing and estimator precision even in this simple Gaussian target distribution situation. The reason is that the optimal scan strategy samples highly variable components more often, focusing more effort on these coordinates than the equal selection probability strategy.

5.2. Banana-shaped Gaussian target distribution

We consider estimating the mean of $\mathbf{X}$ so that $h(\mathbf{X}) = \{X(1) + X(2) + X(3)\}/3$. As shown in Table 1, the random scan with optimal selection probabilities with respect to the risk function criterion outperforms the random scan with equal selection probabilities, reducing the risk by 36%. Consequently, statistical inferences from the optimal random scan are more precise as seen in the performance criteria presented in Table 2. Notice that even in the random variate generation in this simple Gaussian case, the adaptive random scan outperforms the random scan with equal selection probabilities with respect to all criteria.

Fig. 1 displays the autocorrelation plot for simulations from the random scan with optimal and equal selection probabilities. Both scan strategies mix slowly, though the random scan with equal selection probabilities mixes substantially slower, showing significant autocorrelation after a lag of 600 as compared to the optimal scan which displays approximately uncorrelated variates after a lag of 400.

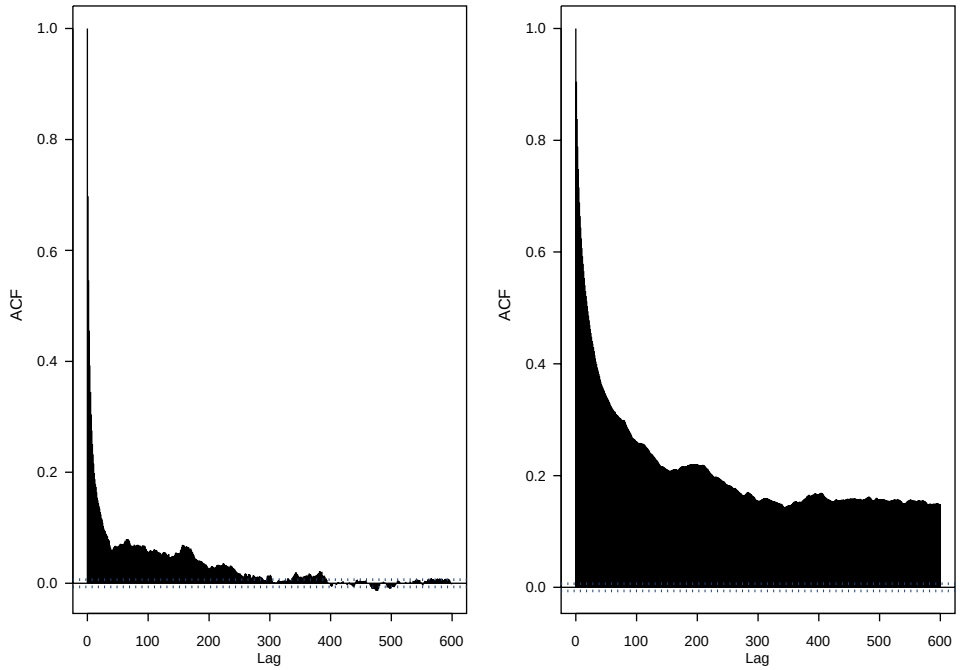

Fig. 1. Autocorrelation plots for the random scan with optimal selection probabilities (left panel) and equal selection probabilities (right panels) for the banana-shaped Gaussian target distribution for estimating $h(\mathbf{X}) = \{X(1) + X(2) + X(3)\}/3$.

5.3. Bimodal Gaussian target distribution

We first consider estimating the mean of $\mathbf{X}$ so that $h(\mathbf{X}) = \{X(1) + X(2) + X(3)\}/3$. As shown in Table 1, the random scan with optimal selection probabilities with respect to the risk function criterion outperforms the random scan with equal selection probabilities, reducing the risk by 24%. Fig. 2 displays the autocorrelation plot for simulations from the random scan with optimal and equal selection probabilities. The optimal scan mixes well. However, the random scan with equal selection probabilities mixes rather slowly, showing significant autocorrelation after a lag of 25.

In studying the bimodal Gaussian target distribution, we found it unnecessary to update the selection probabilities in every iteration of the adaptive random scan as additions of single random variates do not change the selection probabilities from iteration to iteration though add to the computational expense of the algorithm. We thus implemented the adaptive random scan updating the selection probabilities every 1000 iterations. The optimal random scan in this implementation performs equally as well as in the case of updating every iteration, though at less computational expense, mixing significantly faster than the random scan with equal selection probabilities.

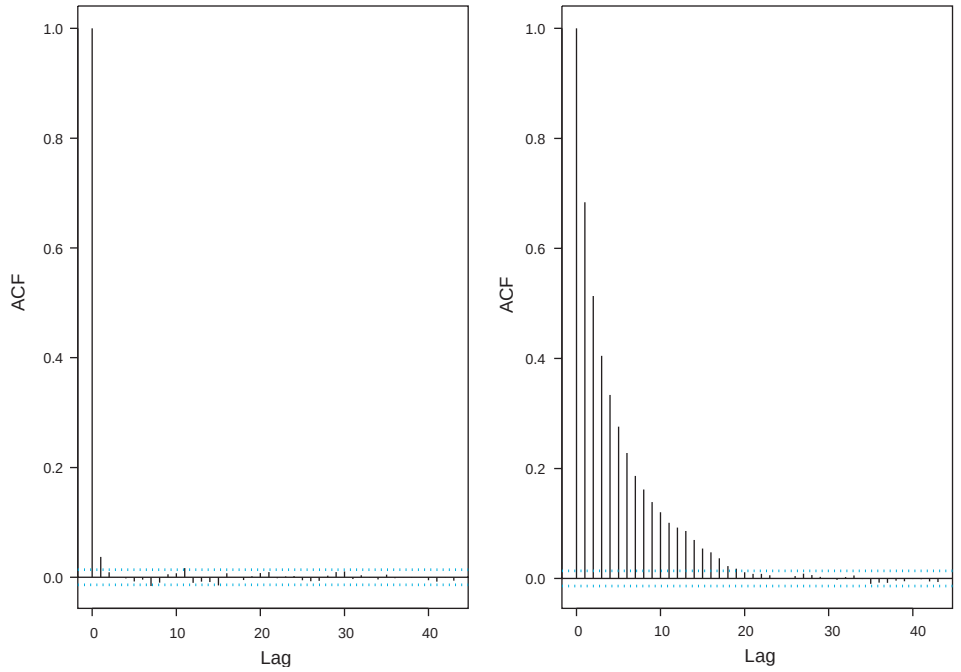

Fig. 2. Autocorrelation plots for the random scan with optimal selection probabilities (left panel) and equal selection probabilities (right panels) for the bimodal Gaussian target distribution for estimating $h(\mathbf{X}) = \{X(1) + X(2) + X(3)\}/3$.

6. Bayesian frailty modeling: animal carcinogenesis

For purposes of illustration and comparison, we apply our adaptive componentwise Hastings algorithm to the animal carcinogenesis data described in Mantel et al. (1977), considered by Clayton (1991). The experiment considered the effect of a putative carcinogen on tumor occurrence in 150 rats, randomly chosen from 50 litters (3 rats per litter). One rat randomly chosen from each litter was treated with the carcinogen, the other two rats from that litter served as controls. The data consist of time to tumor occurrence or censoring recorded to the nearest week (see Clayton, 1991, Table 1) and a single covariate, indicator of carcinogen exposure.

We fit the Bayesian frailty model (1) with noninformative priors, namely $\mu = \sigma = \eta = \tau = c = 0$. the application, we consider three adaptive routines proposed recently in the literature to illustrate the flexibility in implementing our algorithms. As per the motivational discussion of Section 1, these three approaches are used to overcome each of the three time-consuming tasks in implementing Algorithm 1.1.

First, in order to avoid an absorbing state in the space over which the chain traverses, Clayton (1991) proposed iterating between G repetitions of an MCMC sampler over the parameters $(\beta, \mathbf{w}, \Lambda_0)$ and an IP step for sampling γ. Clayton (1991), after extensive

fine-tuning, proposes fixing $G = 10$. We shall let the random selection probabilities $\boldsymbol{\alpha}$ choose G by, within each iteration, deciding with probability α_1 to repeat the MCMC sampling in step (2b) of Algorithm 1.1 and with probability $\alpha_2 = 1 - \alpha_1$ perform the IP step of choosing a γ variate. We decide to optimize the precision in estimating the regression coefficient β through a risk function proposed by Levine and Casella (2004), analogous to that discussed in Section 5. We note that this method satisfies Theorems 3.1 and 4.1, in particular forming a converging sequence of selection probabilities $\{\boldsymbol{\alpha}^{(t)}\}_t$.

Second, rather than performing a Gaussian approximation to the β full conditional distribution, as proposed by Clayton (1991), we implement a Hastings sampler with random proposal distribution. At iteration t, the proposal distribution $q(\beta^* \mid \beta^{(t-1)}, c)$ is assumed $Normal(\beta^{(t-1)}, c^2)$ centered at the previously generated β variate with proposal parameter c. We take the coerced acceptance probability approach of Andrieu and Robert (2003). We find c^* such that $\rho_{c^*} = \rho$ where

$$\rho_c = \int \min\left\{1, \frac{\pi(\beta^*|\cdot)q(\beta^{(t-1)}|\beta^*; c)}{\pi(\beta^{(t-1)}|\cdot)q(\beta^*|\beta^{(t-1)}; c)}\right\} \pi(\beta^{(t-1)})q(\beta^*|\beta^{(t-1)}; c)\ \mathrm{d}\beta^*\ \mathrm{d}\beta^{(t-1)}$$

is the expected acceptance probability. We use the Robbins and Monro stochastic approximation algorithm (Robert and Casella, 1999, Chapter 5) to solve this equation, estimating the step size using the history of the chain. In our application, we force a 60% acceptance rate, ρ. We note that this method satisfies Theorem 4.2, in particular forming a converging sequence of proposal parameters $\{c^{(t)}\}_t$.

Third, rather than performing an intricate accept–reject routine for sampling $v = 1/\gamma$, as proposed by Clayton (1991), we implement a Hastings sampler with random proposal distribution. At iteration t, the proposal distribution $q(v \mid v^{(t-1)}, \sigma_v)$ is assumed $Truncated\ Normal(v^{(t-1)}, \sigma_v^2; 0)$ with location parameter being the previously generated v variate and proposal parameter being the scale parameter σ_v. We take the moment-matching approach of Haario et al. (2001), replacing σ_v with the method of moments estimator $\hat{\sigma}_v$ based on previously generated variates $\{v^{(i)}\}_{i=1}^{t-1}$. We note that this method satisfies Theorem 4.2, in particular forming a converging sequence of proposal parameters $\{\hat{\sigma}_v^{(t)}\}_t$.

Algorithm 4.1 is modified as follows to incorporate these three implementations.

Algorithm 6.1. Animal carcinogenesis: Adaptive componentwise Hastings sampler with random proposal distributions

(1) `Initialization`
   a. `Select` initial values for $\beta^{(0)}, \mathbf{w}^{(0)}, \Lambda_0^{(0)}$.
   b. `Choose` selection probabilities $\boldsymbol{\alpha}^{(0)} = \{\alpha_1^{(0)}, \ldots \alpha_d^{(0)}\}$.
   c. `Select` $c^{(0)}$ and $\hat{\sigma}_v$.

(2) On the tth iteration
   a. `Draw` γ^* from $\gamma^{(t-1)}, \ldots, \gamma^{(t-B)}$.
   b. `Choose` $\boldsymbol{\alpha}^{(t)} \leftarrow R^{(t)}(\boldsymbol{\alpha} \mid \beta^{(1:\sum G_{t-1})})$.

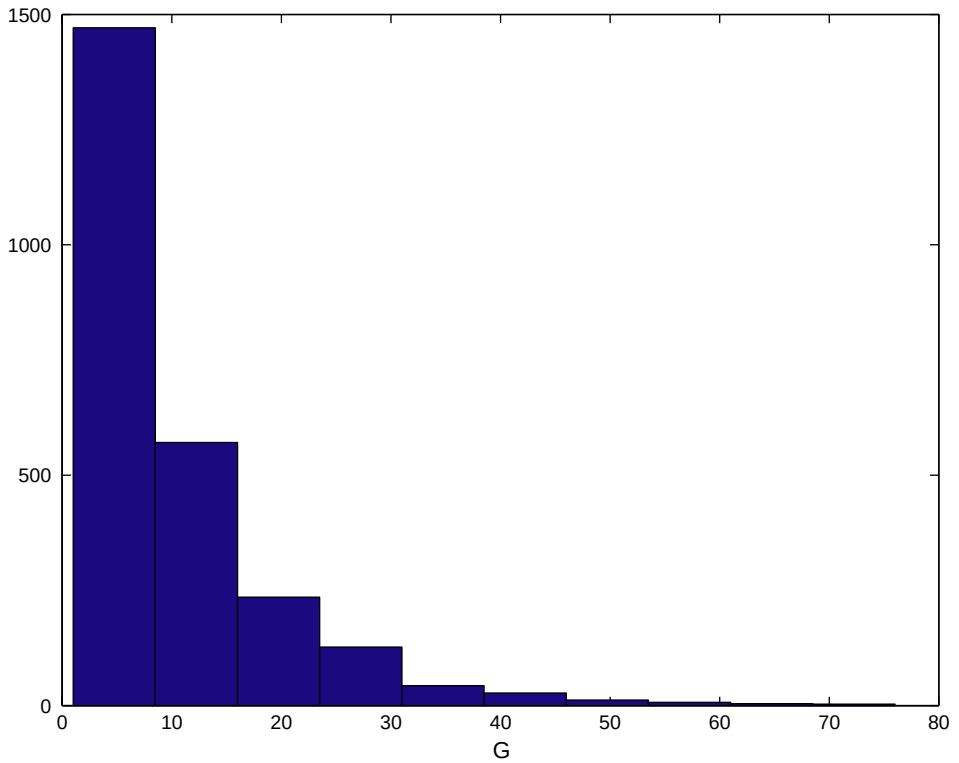

Fig. 3. Histogram of number of repetitions, G_t, of step (2c) each iteration t of Algorithm 6.1.

c. `Generate` a sample for $(\beta, \mathbf{w}, \Lambda_0)$ by iterating over the distributions

$$\beta^{(j+\sum G_{t-1})} \sim [\beta \mid \mathbf{w}^{(j-1+\sum G_{t-1})}, \Lambda_0^{(j-1+\sum G_{t-1})}, \gamma^*, data]$$

$$\mathbf{w}^{(j+\sum G_{t-1})} \sim [\mathbf{w} \mid \beta^{(j+\sum G_{t-1})}, \Lambda_0^{(j-1+\sum G_{t-1})}, \gamma^*, data]$$

$$\Lambda_0^{(j+\sum G_{t-1})} \sim [\Lambda_0 \mid \mathbf{w}^{(j+\sum G_{t-1})}, \beta^{(j+\sum G_{t-1})}, \gamma^*, data]$$

d. `Repeat` steps (2b) and (2c) with probability $\alpha_1^{(t)}$.
e. `Compute` $\hat{\sigma}_v^2$.
f. `Generate` $1/\gamma^{(t)} = v^{(t)} \sim [v \mid \mathbf{w}^{(t \cdot G_t)}]$ using Hastings sampler with *Truncated Normal* $(v^{(t-1)}, \hat{\sigma}_v^2; 0)$ proposal.

(3) `Repeat` steps two to seven until reaching equilibrium.

Though not explicit in Algorithm 6.1, the β variates are generated in step (2c) using a Hastings sampler with coerced acceptance probability and Gaussian proposal distribution.

Following Clayton (1991), we implement Algorithm 6.1 fit to the animal carcinogenesis data with a burn-in of 500 iterations and sampling of 2000 variates. After the

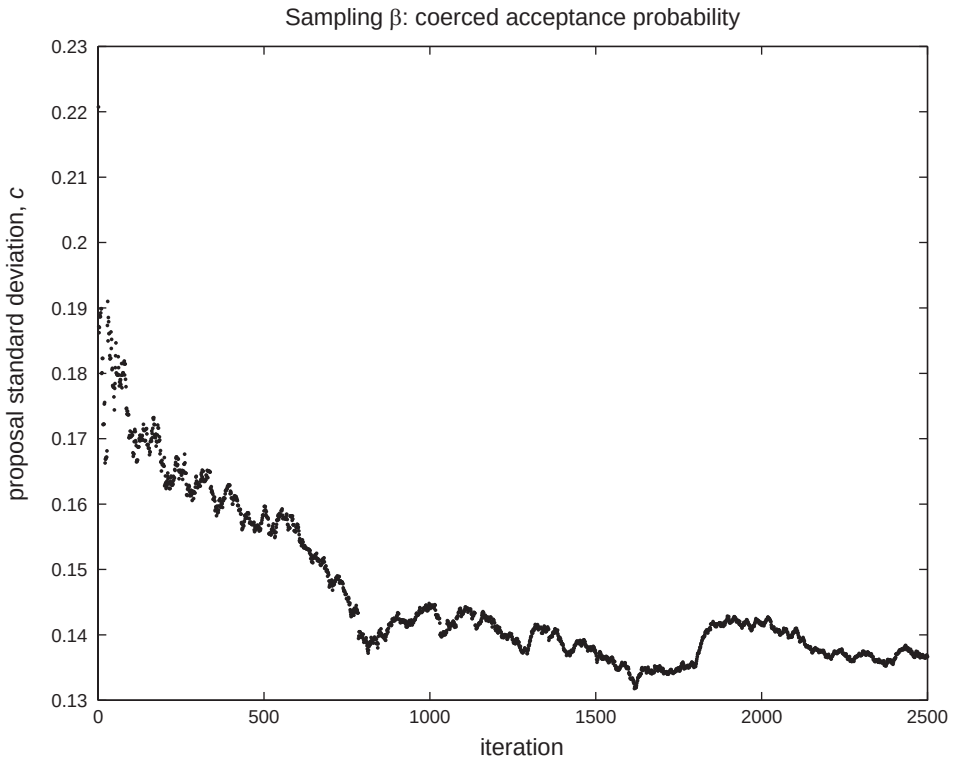

Fig. 4. Convergence plot for the Gaussian proposal parameter c of the Hastings sampler for generating β in step (2c) of Algorithm 6.1.

burn-in period, we fix $B = 100$. We may employ the automated routine of Levine and Casella (2001) to choose B, but for purposes of comparison here, we stay with the B selected by Clayton (1991). Carcinogen exposure is seen to be significantly related to tumor occurrence, the posterior mean of the β coefficient being 0.97 (standard deviation 0.10; hazard ratio 2.64 with 90% credible interval 1.52 to 4.76). The frailty parameter is significant indicating an association within clusters, posterior mean of γ being 0.67 (standard deviation 0.16 with 90% credible interval 0.13–1.51). These findings are of course analogous to those of Clayton (1991).

More importantly, our algorithm is as computationally costly as the Clayton Algorithm 1.1. Specifically, for cost comparison purposes, if we perform the adaptive routine on the selection probabilities (so all computational expense of adapting is built into the code) but force $G = 10$ for both Algorithms 1.1 and 6.1 (so both algorithms perform the same number of simulations from the full conditionals), then the Clayton algorithm requires 0.94 s per iteration and our adaptive algorithm requires 1.01 s per iteration.

Fig. 3 presents the distribution of repeated samplings of step (2c) in Algorithm 6.1, a Monte Carlo estimation of G. Interestingly, on average G is 9.93, close to the value $G = 10$ proposed by Clayton (1991). However, the standard deviation of G is 9.85,

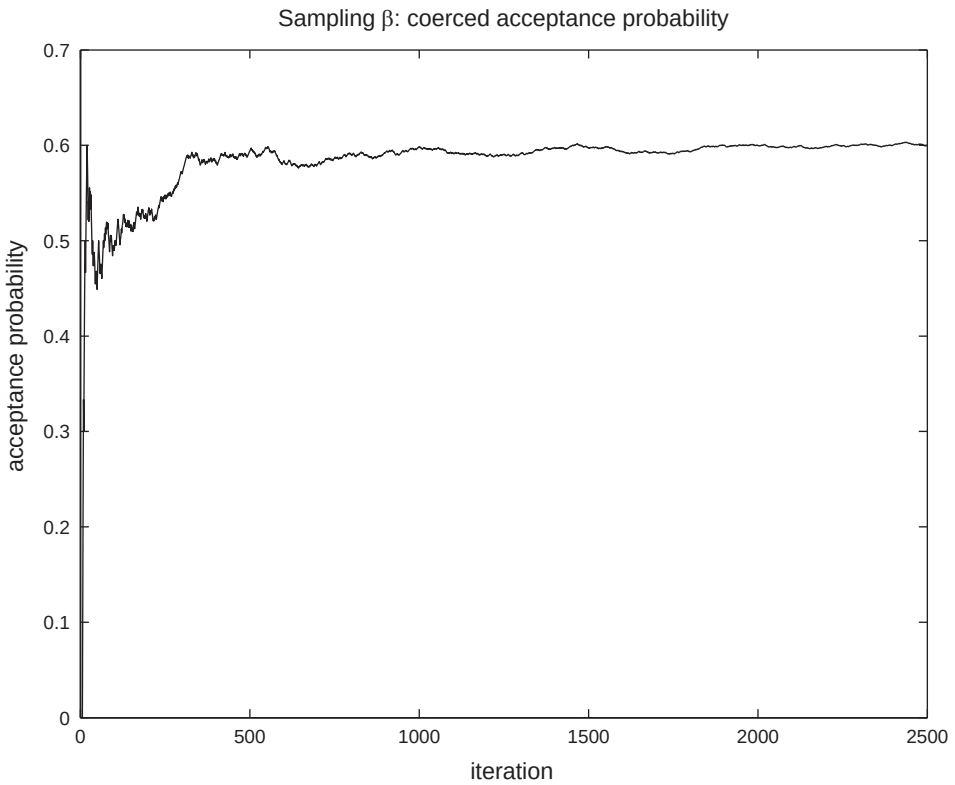

Fig. 5. Convergence plot of the acceptance probabilities in the Hastings sampler for generating β in step (2c) of Algorithm 6.1.

suggesting a fixed G throughout the implementation of the algorithm is not desirable, wasting computation at some iterations, but requiring more repetitions to avoid the absorbing states at some iterations as well. Figs. 4 and 5 present convergence plots of the proposal parameter c and the acceptance probability for the Hastings sampler used to generate $\boldsymbol{\beta}$. Note that the acceptance probabilities converge quickly to the 0.60 value. The standard deviation of the Gaussian proposal suggested in the limit is 0.14, slightly more disperse than the posterior distribution on β, with estimated standard deviation of 0.10. The algorithm recommends a proposal parameter σ_v^2 of 1.39, being the limiting point of the $\{\hat{\sigma}_v^2\}_t$ sequence.

7. Discussion

7.1. Alternative implementations

The adaptive componentwise Hastings sampler of Section 3.1 provides the most general strategy for updating the selection probabilities, being dependent on all

previous random variates generated by the sampler. A number of practical schemes are available which may perform almost as well as the adaptive componentwise Hastings sampler, in terms of convergence rate and precision of Monte Carlo estimates, but require substantially less computational expense.

- *Quasi-adaptive Hastings sampler*: The selection probabilities $\boldsymbol{\alpha}$ need not be updated every iteration of the Hastings sampler as suggested by Algorithm 3.1. Alternatively, we may update the selection probabilities every Mth iteration. These updates may be based on the last $M-1$ iterations so that step 2a may be written as
  2a. `Choose` $\boldsymbol{\alpha}^{(t)} \sim f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(t-M)}, \ldots, \mathbf{X}^{(t-1)})$.
  We call such an updating scheme "quasi-adaptive" since the scan is not fully adaptive, i.e., updating $\boldsymbol{\alpha}$ based on *all* previous samples. The advantage of the quasi-adaptive Hastings sampler over Algorithm 3.1 is that the sampling routine of step 2a, a potentially costly calculation, is performed less frequently. Furthermore, if $\boldsymbol{\alpha}$ is updated every iteration, it is less likely to change much from iteration to iteration since only one more random variate is added. Updating every M iterations may thus be more efficient. Analogous to the arguments of Section 3.1, the chain induced by the quasi-adaptive componentwise Hastings sampler converges to the stationary distribution of interest, we thus do not present the details here.
- *Burn-in*: We may run the adaptive componentwise Hastings sampler for B iterations, say, and then fix the selection probabilities at the value obtained after these B iterations, $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(B)}$. Again, this modification saves computational cost updating $\boldsymbol{\alpha}$ only B iterations rather than throughout the Hastings sampler. Depending on how close the user wishes to set the selections probabilities to the presumed limiting point, the burn-in B could be quite small.
- *Markov property*: If the distribution $f_{\boldsymbol{\alpha}}$ in step 2a of Algorithm 3.1 does not depend on previous iterations of the chain $\{\mathbf{X}^{(j)}\}_{j=0}^{T}$, the adaptive componentwise Hastings sampler retains the Markov property. For the Gibbs sampler, this property allows us to attain geometric convergence of the induced Markov chain to the stationary distribution (see Levine and Casella, 2004). For general Hastings samplers, there seems to be less of a theoretical gain other than the application of standard convergence results of Hastings algorithms as given in Robert and Casella (1999, Section 6.2.2).
- *Blocks*: Algorithms 2.2 and 3.1 suggest separately updating each univariate component of $\mathbf{X} = \{X(1), \ldots, X(d)\}$. However, we may update blocks of components, as in Besag et al. (1995), using the adaptive scheme. The gain is that $\boldsymbol{\alpha}$ is of a lower dimension perhaps allowing for faster computation of step 2a of Algorithm 3.1. Furthermore, blocking may lead to faster mixing of the chain (see, for example, Roberts and Sahu, 1997). Analogous to the arguments of Section 3.1, the chain induced by a "blocked" adaptive Hastings sampler converges to the stationary distribution of interest, we thus do not present the details here.
- *Selection probabilities*: the theoretical development of the adaptive scan focuses on sampling selection probabilities from a distribution $f_{\boldsymbol{\alpha}}$. However, we may choose selection probabilities with respect to an optimality criterion, say in Algorithm 3.1 step

2a. Choose $\boldsymbol{\alpha}^{(t)}(\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) \leftarrow R_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)})$.

We have taken this approach in Theorem 4.2, the proof of convergence being analogous to that of Theorem 4.1.

- *Random proposal distributions*: A number of variations on the random proposal distribution theme in Algorithm 4.1 are available. For example, we are not restricted to univariate parameters γ. The random proposal distribution may be characterized by a p-vector of parameters $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_p)$. We may also use different parameters γ_i for each proposal distribution q_i, $i = 1,\ldots,d$. Analogous arguments as those in Section 4 show that the induced chain still converges to the stationary distribution. The variations discussed above also apply to the distribution g_γ. We can mix and match these variations for choosing selection probabilities $\boldsymbol{\alpha}$ and random proposal distribution parameters γ.

Many other variations on the adaptive componentwise Hastings sampler theme are available to the practitioner. Our favorite implementation, which is used in Section 5, is to initialize the algorithm for B iterations using $\alpha_j = 1/d$ for all $j = 1,\ldots,d$. This burn-in sets up a sequence of variates $\mathbf{X}^{(0)},\ldots\mathbf{X}^{(B)}$ for the calculation of selection probabilities in step 2a of Algorithm 3.1. Following this initialization, we run a quasi-adaptive Hastings sampler, updating $\boldsymbol{\alpha}$ every M iterations, until the selection probabilities stabilize, say the L_2 norm of the selection probabilities change less than a prespecified value ε for five consecutive iterations. We then fix the selection probabilities on this limiting value for the remainder of the algorithm. As mentioned above, updating the selection probabilities every M iterations and then fixing the selection probabilities towards the end of the algorithm is computationally more efficient as the full adaptive routine does not change $\boldsymbol{\alpha}$ much iteration to iteration near the limiting selection probabilities.

7.2. Objective functions

In implementing MCMC samplers, we focus on two properties of the induced Markov chain: (1) rate of convergence to the stationary distribution and (2) MC estimator precision (Besag et al., 1995; Gelman et al., 1996). These properties are of course related, however due to the differing notions of optimality, they may suggest conflicting implementations of the sampler. In fact, Besag and Green (1993) and Mira (2001) suggest, after an appropriate burn-in period, switching from a sampler with good convergence properties to a sampler with good efficiency properties. The goal is to choose an MCMC sampler which converges to the target distribution quickly (fast convergence rate) and estimates parameters of interest via the generated variates with as small a variance as possible (high precision/efficiency).

Our adaptive componentwise Hastings sampler affords the user with the flexibility to implement either of these practices, with their preference of decision criteria, through the choice of selection probabilities via the function $f_{\boldsymbol{\alpha}}$ and proposal distribution parameter via the function g_γ. The practitioner is advised to choose criteria that satisfy the conditions of Theorems 4.1 and 4.2, which seem easily attained by any reasonable choice of objective function. Furthermore, the criteria should be computationally inexpensive relative to the sampler, particularly if they invoke optimization routines.

In this paper, we focus on estimator efficiency for specifying $f_{\boldsymbol{\alpha}}$, drawing on the minimax optimality results of Levine and Casella (2004). The user of course is given the option to choose any measure of efficiency through the asymptotic variance of the desired estimates, integrated autocorrelation time, or whichever function $f_{\boldsymbol{\alpha}}$ appropriate for the problem at hand. Alternatively, we may focus attention on sampler convergence, where the convergence rate is a natural measure of performance of the Hastings sampler. However, analytical computation of convergence rates for MCMC samplers is notoriously difficult. Gaussian approximations in the sense of Roberts and Sahu (2001) may provide a means of appropriately specifying the convergence rate towards this end. We consider these options elsewhere.

Convergence and efficiency of the Hastings sampler is also affected by the choice of the proposal distribution. In this paper we focus on selecting proposal distributions by coercing the acceptance probability and matching the moments with that of the target distribution. Andrieu and Robert (2003) show that these two methods are closely related to sampler efficiency in the post-processing of generated variates. In choosing g_γ we may consider an asymptotic variance approach similar to that of Levine and Casella (2004) or relate the proposal parameter to the convergence rate.

Overall, our algorithms provide extensive flexibility in Hastings sampler implementations, relieving the user of much front-end work and algorithm fine-tuning. If decision criteria are chosen carefully to minimize computational cost, these algorithms will complement other implementation strategies such as reparameterization and variate blocking/clustering to improve speed of convergence to stationarity and algorithm efficiency in terms of Monte Carlo estimator precision.

Appendix A. Convergence proofs

We prove the convergence results for the case of adaptation of the selection probabilities $\boldsymbol{\alpha}$ as in Theorem 3.1 and adaptation of the proposal distribution parameters γ as in Theorem 4.2. This case is the most involved, convergence proofs for all other permutations of the algorithms following an analogous line of reasoning. The conditions under consideration thus

a. $f_{\boldsymbol{\alpha}}^{(t)}(\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) \to f_{\boldsymbol{\alpha}}$ almost everywhere,
b. $\gamma^{(t)}(\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) \to \gamma$ almost everywhere for fixed γ,
c. the componentwise Hastings sampler with fixed selection probabilities $\boldsymbol{\alpha}$ and fixed random proposal parameters γ induces an ergodic Markov chain with stationary distribution π, and
d. the proposal distribution $q(\mathbf{Y}\,|\,\mathbf{X};\gamma)$ is chosen to be stochastically equicontinuous in γ: for $\varepsilon > 0$, there exists $\delta > 0$ such that $|\gamma_1 - \gamma_2| < \delta$ implies $|q(\mathbf{Y}\,|\,\mathbf{X},\gamma_1) - q(\mathbf{Y}\,|\,\mathbf{X},\gamma_2)| < \varepsilon$ almost everywhere on the support of π.

Let $\eta_1,\ \eta_2,\ \varepsilon_1,\varepsilon_2 > 0$. By Egoroff's theorem and condition (a) there exists a T_1 such that

$$|f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha}|\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| < \varepsilon_1$$

for all $t \geqslant T$ except on a set of measure less than η_1. Similarly, by condition (b) there exists a T_2 such that

$$|\gamma^{(t)}(\gamma|\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) - \gamma| < \varepsilon_2$$

for all $t \geqslant T_2$ except on a set of measure less than η_2.

Consider a transition from state $\mathbf{X}$ to $\mathbf{Y}$ of the adaptive componentwise Hastings sampler. Following Eq. (2), the transition kernel is then

$$P_i(\mathbf{Y}\,|\,\mathbf{X};\boldsymbol{\alpha},\gamma) = \min\left\{1, \frac{\pi(Y(i)\,|\,\mathbf{X}_{-i})}{\pi(X(i)\,|\,\mathbf{Y}_{-i})}\,\frac{q_i(\mathbf{X}\,|\,\mathbf{Y};\gamma)}{q_i(\mathbf{Y}\,|\,\mathbf{X};\gamma)}\right\} q_i(\mathbf{Y}\,|\,\mathbf{X};\gamma) + \{1 - r_i(\mathbf{X})\}\delta_{\mathbf{Y}}(\mathbf{X}),$$

where $r_i(\mathbf{X}^{(t-1)}) = \int \rho_i(\mathbf{X},\mathbf{Y}) q_i(\mathbf{Y}\,|\,\mathbf{X};\gamma)\,\mathrm{d}Y(i)$.

Note that $P_i(\mathbf{Y}\,|\,\mathbf{X};\boldsymbol{\alpha},\gamma^{(t)})$ is a function of $\gamma^{(t)}$ through $q_i(\mathbf{Y}\,|\,\mathbf{X};\gamma^{(t)})$. Let $\eta_3, \varepsilon_1/3 > 0$. By Egoroff's theorem and condition (d), there exists T_3 such that

$$|P_i(\mathbf{Y}\,|\,\mathbf{X};\boldsymbol{\alpha},\gamma^{(t)}) - P_i(\mathbf{Y}\,|\,\mathbf{X};\boldsymbol{\alpha},\gamma)| < \varepsilon_1$$

for all $t \geqslant T_3$ except on a set of measure less than η_3.

Consider the following string of equations.

$$\begin{aligned}
&\left| \int_{(0,1)^d} \int \sum_{i=1}^{d} \alpha_i P(\mathbf{X}^{(t)}\,|\,\mathbf{X}^{(t-1)};\gamma^{(t)}) f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha}|\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) \right. \\
&\quad \times \mu_0(\mathrm{d}\mathbf{X}^{(0)}) \cdots \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)})\,\mathrm{d}\boldsymbol{\alpha} \\
&\quad \left. - \int_{(0,1)^d} \int P(\mathbf{X}^{(t)}\,|\,\mathbf{X}^{(t-1)};\gamma) f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \mu_0(\mathrm{d}\mathbf{X}^{(0)}) \cdot \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)})\,\mathrm{d}\boldsymbol{\alpha} \right| \\
&< \left| \int_{(0,1)^d} \int \sum_{i} \alpha_i \left\{ P(\mathbf{X}^{(t)}\,|\,\mathbf{X}^{(t-1)};\gamma^{(t)}) f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha}\,|\,\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) \right. \right. \\
&\quad + P(\mathbf{X}^{(t)}\,|\,\mathbf{X}^{(t-1)};\gamma) f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) - P(\mathbf{X}^{(t)}\,|\,\mathbf{X}^{(t-1)};\gamma^{(t)}) f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \\
&\quad + P(\mathbf{X}^{(t)}\,|\,\mathbf{X}^{(t-1)};\gamma^{(t)}) f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) - 2P(\mathbf{X}^{(t)}\,|\,\mathbf{X}^{(t-1)};\gamma) f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \\
&\quad + P(\mathbf{X}^{(t)}\,|\,\mathbf{X}^{(t-1)};\gamma^{(t)}) f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha}\,|\,\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) \\
&\quad \left. + P(\mathbf{X}^{(t)}\,|\,\mathbf{X}^{(t-1)};\gamma^{(t)}) f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha}\,|\,\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(t-1)}) \right\} \\
&\quad \left. \times \mu_0(\mathrm{d}\mathbf{X}^{(0)}) \cdots \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)})\,\mathrm{d}\boldsymbol{\alpha} \right|
\end{aligned}$$

$$
\begin{aligned}
&< \int_{(0,1)^d} \int \sum_i \alpha_i \{ |P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \gamma^{(t)}) - P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \gamma)| \\
&\cdot |f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| \\
&\quad + |P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \gamma^{(t)}) - P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \gamma)| |f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| \\
&\quad + |P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \gamma)| f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| \} \\
&\quad \times \mu_0(\mathrm{d}\mathbf{X}^{(0)}) \cdots \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)}) \, \mathrm{d}\boldsymbol{\alpha} \\
&< \varepsilon_1^2 + 2\varepsilon_1 \qquad \text{(A.1)}
\end{aligned}
$$

by the inequalities implied by conditions (a), (b), and (d).

We may now show convergence in total variation norm. Recall the distribution of $\mathbf{X}^{(t)}$ during the adaptive componentwise Hastings sampler, denoted μ_t, as defined in Eqs. (3) and (4). In the componentwise Hastings sampler where the selection probability distribution $f_{\boldsymbol{\alpha}}$ and proposal parameters γ are not updated in each iteration, we will denote the distribution of $\mathbf{X}^{(t)}$ by $\mu_{t;\boldsymbol{\alpha},\gamma}$, defined analogously to μ_t though replacing $f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)})$ with $f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ and $\gamma^{(t)}$ with γ.

By the triangle inequality we have that

$$\|\mu_t - \pi\|_{TV} \leqslant \|\mu_t - \mu_{t;\boldsymbol{\alpha},\gamma}\|_{TV} + \|\mu_{t;\boldsymbol{\alpha},\gamma} - \pi\|_{TV}$$

for fixed selection probability distribution $f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ and fixed proposal parameter γ.

Let μ_0 be the initial distribution of initial variate $\mathbf{X}^{(0)}$. By condition (c), there exists T such that for all $t \geqslant T$,

$$\|\mu_{t;\boldsymbol{\alpha},\gamma} - \pi\|_{TV} < \varepsilon/2.$$

Furthermore, we have that for measurable sets A,

$$
\begin{aligned}
&\|\mu_t - \mu_{t;\boldsymbol{\alpha}.\gamma}\|_{TV} \\
&\quad = \sup_A \left| \int_A \int_{(0,1)^d} \int \sum_i \alpha_i P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \boldsymbol{\alpha}, \gamma^{(t)}) f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \right. \\
&\qquad \times \mu_0(\mathrm{d}\mathbf{X}^{(0)}) \cdots \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)}) \, \mathrm{d}\boldsymbol{\alpha} \, \mathrm{d}\mathbf{X}^{(t)} \\
&\quad - \int_A \int_{(0,1)^d} \int \sum_i \alpha_i P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \boldsymbol{\alpha}, \gamma) f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \; \mu_0(\mathrm{d}\mathbf{X}^{(0)}) \\
&\qquad \left. \times \cdots \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)}) \, \mathrm{d}\boldsymbol{\alpha} \, \mathrm{d}\mathbf{X}^{(t)} \right|
\end{aligned}
$$

$$
\begin{aligned}
&< \sup_A \int_A \int_{(0,1)^d} \int \sum_i \alpha_i \left\{ \left| P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \boldsymbol{\alpha}, \gamma^{(t)}) - P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \boldsymbol{\alpha}, \gamma) \right| \right. \\
&\quad \cdot |f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| \\
&\quad + |f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| |P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \boldsymbol{\alpha}, \gamma^{(t)}) - P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \boldsymbol{\alpha}, \gamma)| \\
&\quad \left. + |P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \boldsymbol{\alpha}, \gamma)| |f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| \right\}.
\end{aligned}
$$

By (A.1) and Scheffe's theorem (Billingsley, 1995; Section 16), the result follows.

We note that Theorem 4.2 follows by considering $f_{\boldsymbol{\alpha}}^{(t)}$ and $f_{\boldsymbol{\alpha}}$ as point mass distributions at $\boldsymbol{\alpha}^{(t)}$ and $\boldsymbol{\alpha}$, respectively. Theorem 3.1 follows by assuming fixed proposal distribution parameter γ.

Theorem 4.1 follows analogously though with focus on the distributions $g_{\gamma}^{(t)}$ and g_{γ} as opposed to $\gamma^{(t)}$ and γ.

In particular, let $\eta_1,\ \eta_2,\ \varepsilon_1, \varepsilon_2 > 0$. By Egoroff's theorem and condition (a) there exists a T_1 such that

$$
|f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| < \varepsilon_1
$$

for all $t \geqslant T$ except on a set of measure less than η_1. Similarly, by Egoroff's theorem and condition (b) there exists a T_2 such that

$$
|\gamma^{(t)}(\gamma \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) - \gamma| < \varepsilon_2
$$

for all $t \geqslant T_2$ except on a set of measure less than η_2.

By the triangle inequality we have that

$$
\|\mu_t - \pi\|_{TV} \leqslant \|\mu_t - \mu_{t;\boldsymbol{\alpha},\gamma}\|_{TV} + \|\mu_{t;\boldsymbol{\alpha},\gamma} - \pi\|_{TV}
$$

for fixed selection probability distribution $f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ and fixed proposal parameter γ. Here $\mu_{t;\boldsymbol{\alpha},\gamma}$ denotes the distribution of $\mathbf{X}^{(t)}$ at iteration t of the componentwise Hastings sampler with fixed selection probability distribution $f_{\boldsymbol{\alpha}}$ and random proposal distribution g_{γ}.

By condition (c), there exists T such that for all $t \geqslant T$, $\|\mu_{t;\boldsymbol{\alpha},\gamma} - \pi\|_{TV} < \varepsilon/2$. Furthermore, we note that

$$
\begin{aligned}
&\|\mu_t - \mu_{t;\boldsymbol{\alpha}.\gamma}\|_{TV} \\
&= \sup_A \left| \int_A \int \int_{(0,1)^d} \int \sum_i \alpha_i P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \boldsymbol{\alpha}, \gamma) f_{\boldsymbol{\alpha}}^{(t)}(\boldsymbol{\alpha} \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \right. \\
&\quad \times g_{\gamma}^{(t)}(\gamma \mid \mathbf{X}^{(0)}, \ldots, \mathbf{X}^{(t-1)}) \mu_0(\mathrm{d}\mathbf{X}^{(0)}) \cdots \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)}) \, \mathrm{d}\boldsymbol{\alpha} \, \mathrm{d}\gamma \, \mathrm{d}\mathbf{X}^{(t)} \\
&\quad - \int_A \int \int_{(0,1)^d} \int \sum_i \alpha_i P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}; \boldsymbol{\alpha}, \gamma) f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) g_{\gamma}(\gamma) \\
&\quad \left. \times \mu_0(\mathrm{d}\mathbf{X}^{(0)}) \cdots \mu_{t-1}(\mathrm{d}\mathbf{X}^{(t-1)}) \, \mathrm{d}\boldsymbol{\alpha} \, \mathrm{d}\gamma \, \mathrm{d}\mathbf{X}^{(t)} \right|.
\end{aligned}
$$

By condition (d) and the definition of the transition kernel for the componentwise Hastings sampler, a similar set of inequalities to (A.1) follows. Consequently, an application of Scheffe's theorem implies that for all $t \geqslant T$, $\|\mu_t - \mu_{t;\boldsymbol{\alpha},\gamma}\|_{TV} < \varepsilon/2$.

References

Andrieu, C., Robert, C.P., 2003. Controlled MCMC for optimal sampling. Technical Report, Universitè Paris-Dauphine. http://www.ceremade.dauphine.fr/xian/publications.html.

Besag, J., 2000. Markov chain Monte Carlo for statistical inference. University of Washington, Center of Statistics and the Social Sciences Working paper #9.

Besag, J., Green, P.J., 1993. Spatial statistics and Bayesian computation. J. Roy. Statist. Soc. B 55, 25–37.

Besag, J., Green, P., Higdon, D., Mengersen, K., 1995. Bayesian computation and stochastic systems. Statist. Sci. 10, 3–41.

Billingsley, P., 1995. Probability and Measure, 3rd Edition. Wiley, NY.

Clayton, D.G., 1991. A Monte Carlo method for Bayesian inference in frailty models. Biometrics 47, 467–485.

Gelman, A.G., Roberts, G.O., Gilks, W.R., 1996. Efficient metropolis jumping rules. In: Bernardo, J.M., Berger, J.O., David, A.F., Smith, A.F.M. (Eds.), Oxford University Press, Bayesian Statistics V, pp. 599–608.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, 721–741.

Gilks, W.R., Best, N.G., Tan, K.K.C., 1995. Adaptive rejection metropolis sampling within Gibbs sampling. Appl. Statist. 44, 455–472.

Haario, H., Saksman, E., Tamminen, J., 1999. Adaptive proposal distribution for random walk metropolis algorithm. Comput. Statist. 14, 375–395.

Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive metropolis algorithm. Bernoulli 7, 223–242.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Holden, L., 1998. Adaptive chains. Norwegian Computing Center Publication 904009, Nr. SAND/11/98. http://publications.nr.no/ad8.pdf.

Ibrahim, J.G., Chen, M-H., Sinha, D., 2001. Bayesian Survival Analysis. Springer, NY.

Levine, R.A., Casella, G., 2001. Implementations of the Monte Carlo EM algorithm. J. Comput. Graph. Statist. 10, 422–439.

Levine, R.A., Casella, G., 2004. Optimizing random scan Gibbs samplers. Technical Report, Department of Mathematics and Statistics, San Diego State University.

Mantel, N., Bohidar, N.R., Ciminera, J.L., 1977. Mantel–Haenszel analysis of litter-matched time-to-response data with modifications for recovery of interlitter information. Can. Res. 37, 3863–3868.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092.

Mira, A., 2001. Ordering and improving the performance of Monte Carlo Markov chains. Statist. Sci. 16, 340–350.

Robert, C.P., Casella, G., 1999. Monte Carlo Statistical Methods. Springer, NY.

Roberts, G.O., Sahu, S.K., 1997. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. J. Roy. Statist. Soc. B 59, 291–317.

Roberts, G.O., Sahu, S.K., 2001. Approximate predetermined convergence properties of the Gibbs sampler. J. Comput. Graph. Statist. 10, 216–229.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Statist. Assoc. 12, 1151–1172.