J. R. Statist. Soc. B (2011)
73, *Part* 2, *pp.* 1–37

Riemann manifold Langevin and Hamiltonian Monte Carlo methods

Mark Girolami and Ben Calderhead

University College London, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, October 13th, 2010*, Professor D. M. Titterington *in the Chair*]

Summary. The paper proposes Metropolis adjusted Langevin and Hamiltonian Monte Carlo sampling methods defined on the Riemann manifold to resolve the shortcomings of existing Monte Carlo algorithms when sampling from target densities that may be high dimensional and exhibit strong correlations. The methods provide fully automated adaptation mechanisms that circumvent the costly pilot runs that are required to tune proposal densities for Metropolis–Hastings or indeed Hamiltonian Monte Carlo and Metropolis adjusted Langevin algorithms. This allows for highly efficient sampling even in very high dimensions where different scalings may be required for the transient and stationary phases of the Markov chain. The methodology proposed exploits the Riemann geometry of the parameter space of statistical models and thus automatically adapts to the local structure when simulating paths across this manifold, providing highly efficient convergence and exploration of the target density. The performance of these Riemann manifold Monte Carlo methods is rigorously assessed by performing inference on logistic regression models, log-Gaussian Cox point processes, stochastic volatility models and Bayesian estimation of dynamic systems described by non-linear differential equations. Substantial improvements in the time-normalized effective sample size are reported when compared with alternative sampling approaches. MATLAB code that is available from the authors allows replication of all the results reported.

Keywords:

1. Introduction

For an unnormalized probability density function $\tilde{p}(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^D$, the normalized density follows as $p(\boldsymbol{\theta}) = \tilde{p}(\boldsymbol{\theta}) / \int \tilde{p}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$, which for many statistical models is analytically intractable. Monte Carlo estimates of integrals with respect to $p(\boldsymbol{\theta})$, which commonly appear in Bayesian statistics, are therefore required. The predominant methodology for sampling from such a probability density is Markov chain Monte Carlo (MCMC) sampling; see for example Robert (2004), Gelman *et al.* (2004) and Liu (2001). The most general algorithm defining a Markov process with invariant density $p(\boldsymbol{\theta})$ is the *Metropolis–Hastings* algorithm (Metropolis *et al.*, 1953; Hastings, 1970), which is arguably one of the *most successful and influential* Monte Carlo algorithms (Beichl and Sullivan, 2000).

The Metropolis–Hastings algorithm proposes transitions $\boldsymbol{\theta} \mapsto \boldsymbol{\theta}^*$ with density $q(\boldsymbol{\theta}^* | \boldsymbol{\theta})$, which are then accepted with probability

Address for correspondence: Mark Girolami, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK.
E-mail: girolami@stats.ucl.ac.uk

 1369–7412/11/73000

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\{1, \tilde{p}(\boldsymbol{\theta}^*)\, q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) / \tilde{p}(\boldsymbol{\theta})\, q(\boldsymbol{\theta}^*|\boldsymbol{\theta})\}.$$

This acceptance probability ensures that the Markov chain is reversible with respect to the stationary target density $p(\boldsymbol{\theta})$ and satisfies detailed balance; see for example Robert (2004), Neal (1993a, 1996) and Liu (2001). Typically, the proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ which drives the Markov chain takes the form of a random walk; for example $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \boldsymbol{\Lambda})$ is a D-dimensional normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Lambda}$.

High acceptance rates can be achieved by proposing smaller transitions; however, larger amounts of time will then be required to make long traversals of parameter space. In high dimensions, when D is large, the random walk becomes inefficient, resulting in low rates of acceptance, poor mixing of the chain and highly correlated samples. A consequence of this is a small effective sample size ESS from the chain; see Robert (2004), Neal (1996) and Liu (2001). Although there have been various suggestions to overcome this inefficiency, guaranteeing detailed balance and ergodicity of the chain places constraints on what can be achieved in alleviating this problem (Andrieu and Thoms, 2008; Robert, 2004; Neal, 1993a). Design of a good general purpose proposal mechanism providing large proposal transitions that are accepted with high probability remains something of an engineering art form.

Major steps forward in this regard were made when a proposal process derived from a discretized Langevin diffusion with a drift term based on the gradient information of the target density was suggested in the Metropolis adjusted Langevin algorithm (MALA) (Roberts and Stramer, 2003). Likewise the Hamiltonian Monte Carlo (HMC) method (Duane *et al.*, 1987) was proposed in the statistical physics literature as a means of efficiently simulating states from a physical system which was then applied to problems of statistical inference (Neal, 1993a, b, 1996; Liu, 2001). Duane *et al.* (1987) referred to the method as hybrid Monte Carlo sampling; however, we shall follow others and use the term Hamiltonian to make it explicit that the method is based on Hamiltonian dynamics. In HMC sampling, a deterministic proposal process based on Hamiltonian dynamics is employed along with additional stochastic proposals that together provide an ergodic Markov chain that is capable of making large transitions that are accepted with high probability.

Despite the potential efficiency gains to be obtained in MCMC sampling from such proposal mechanisms that are inherent in the MALA and HMC methods, the tuning of these MCMC methods remains a major issue especially for challenging inference problems. This paper seeks to address these issues in a systematic manner by adopting an overarching geometric framework for the overall development of MCMC methods such as these.

Brief reviews of the MALA and HMC methods within the context of statistical inference are provided in the following two sections. In Section 4 differential geometric concepts that are employed in the study of asymptotic statistics are considered within the context of MCMC methodology. Section 5 proposes a generalization of the MALA that takes into account the natural geometry of the target density, making use of the definition of a Langevin diffusion on a Riemann manifold. Likewise in Section 6 a generalization of HMC sampling, Riemann manifold HMC (RMHMC) sampling, is presented, which takes advantage of the manifold structure of the parameter space and allows for more efficient proposal transitions to be made. Finally, in Sections 7–10, this new methodology is demonstrated and assessed on some interesting statistical problems, namely Bayesian logistic regression, stochastic volatility modelling, log-Gaussian Cox point processes and parameter inference in dynamical systems.

2. Metropolis adjusted Langevin algorithm

Consider the random vector $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$ and denote the log-density by $\mathcal{L}(\boldsymbol{\theta}) \equiv$

$\log\{p(\boldsymbol{\theta})\}$; then the MALA is based on a Langevin diffusion, with stationary distribution $p(\boldsymbol{\theta})$, defined by the stochastic differential equation (SDE)

$$\mathrm{d}\boldsymbol{\theta}(t) = \nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\}\,\mathrm{d}t/2 + \mathrm{d}\mathbf{b}(t)$$

where $\mathbf{b}$ denotes a D-dimensional Brownian motion. A first-order Euler discretization of the SDE gives the proposal mechanism

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^n + \varepsilon^2\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^n)/2 + \varepsilon\mathbf{z}^n$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$ and ε is the integration step size. Convergence to the invariant distribution $p(\boldsymbol{\theta})$ is no longer guaranteed for finite step size ε owing to the first-order integration error that is introduced. This discrepancy can be corrected by employing a Metropolis acceptance probability after each integration step, thus ensuring convergence to the invariant measure. As $\mathbf{z}$ is an isotropic standardized normal variate and with

$$\boldsymbol{\mu}(\boldsymbol{\theta}^n, \varepsilon) = \boldsymbol{\theta}^n + \frac{\varepsilon^2}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^n)$$

then the discrete form of the SDE defines a proposal density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n) = \mathcal{N}\{\boldsymbol{\theta}^*|\boldsymbol{\mu}(\boldsymbol{\theta}^n, \varepsilon), \varepsilon^2\mathbf{I}\}$ with acceptance probability of standard form $\min\{1, p(\boldsymbol{\theta}^*)\,q(\boldsymbol{\theta}^n|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^n)\,q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n)\}$.

The optimal scaling ε for the MALA has been theoretically analysed in the limit as $D \to \infty$ for factorizable $p(\boldsymbol{\theta})$ (Roberts and Rosenthal, 1998). Although the drift term in the proposal mechanism for the MALA defines the direction for the proposal based on the gradient information (albeit the Euclidean form) it is clear that the isotropic diffusion will be inefficient for strongly correlated variables $\boldsymbol{\theta}$ with widely differing variances forcing the step size to accommodate the variate with smallest variance. This issue can be circumvented by employing a preconditioning matrix (Roberts and Stramer, 2003) $\mathbf{M}$ such that

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^n + \varepsilon^2\mathbf{M}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^n)/2 + \varepsilon\sqrt{\mathbf{M}}\mathbf{z}^n$$

where $\sqrt{\mathbf{M}}$ can be obtained by diagonalization of $\mathbf{M}$ or via Cholesky decomposition such that $\mathbf{M} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$ and $\sqrt{\mathbf{M}} = \mathbf{U}$. It is unclear how this matrix should be defined in any systematic and principled manner; indeed a global level of preconditioning may be inappropriate for differing transient and stationary regimes of the Markov process as demonstrated in Christensen *et al.* (2005).

3. Hamiltonian Monte Carlo methods

We now give a brief introduction to the HMC method; for a detailed description and extensive review see Neal (2010). As in the previous section consider the random variable $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$. In HMC sampling an independent auxiliary variable $\mathbf{p} \in \mathbb{R}^D$ with density $p(\mathbf{p}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$ is introduced. The joint density follows in factorized form as $p(\boldsymbol{\theta}, \mathbf{p}) = p(\boldsymbol{\theta})\,p(\mathbf{p}) = p(\boldsymbol{\theta})\,\mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$. If we denote the logarithm of the desired density by $\mathcal{L}(\boldsymbol{\theta}) \equiv \log\{p(\boldsymbol{\theta})\}$, the negative joint log-probability is

$$H(\boldsymbol{\theta}, \mathbf{p}) = -\mathcal{L}(\boldsymbol{\theta}) + \tfrac{1}{2}\log\{(2\pi)^D|\mathbf{M}|\} + \tfrac{1}{2}\mathbf{p}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{p}. \quad (1)$$

The physical analogy of this negative joint log-probability is a Hamiltonian (Duane *et al.*, 1987; Leimkuhler and Reich, 2004), which describes the sum of a potential energy function $-\mathcal{L}(\boldsymbol{\theta})$, defined at the position $\boldsymbol{\theta}$, and a kinetic energy term $\mathbf{p}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{p}/2$ where the auxiliary variable $\mathbf{p}$ is interpreted as a momentum variable and the covariance matrix $\mathbf{M}$ denotes a mass matrix.

The derivatives of H with respect to $\boldsymbol{\theta}$ and $\mathbf{p}$ have a physical interpretation as the time evolution, with respect to a fictitious time τ, of the dynamic system as given by Hamilton's equations

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\tau} &= \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p}, \\ \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}\tau} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}). \end{aligned} \tag{2}$$

The solution flow for the differential equations, $(\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)) = \Phi_\tau(\boldsymbol{\theta}(0), \mathbf{p}(0))$,

(a) preserves the total energy, i.e. $H\{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\} = H\{\boldsymbol{\theta}(0), \mathbf{p}(0)\}$, and hence the joint density $p\{\boldsymbol{\theta}(\tau), \mathbf{p}(\tau)\} = p\{\boldsymbol{\theta}(0), \mathbf{p}(0)\}$,
(b) preserves the volume element $\mathrm{d}\boldsymbol{\theta}(\tau)\,\mathrm{d}\mathbf{p}(\tau) = \mathrm{d}\boldsymbol{\theta}(0)\,\mathrm{d}\mathbf{p}(0)$ and
(c) is time reversible (Leimkuhler and Reich, 2004).

For practical applications of interest the differential equations (2) cannot be solved analytically and numerical methods are required. There is a class of numerical integrators for Hamiltonian systems which will fully satisfy criteria (b) and (c), volume preservation and time reversibility, and approximately satisfy (a), energy conservation, to a given order of error; see Leimkuhler and Reich (2004). The Stormer–Verlet or leapfrog integrator was employed in Duane *et al.* (1987), and in various statistical applications, e.g. and Liu (2001) and Neal (1993b, 2010), as described below:

$$\mathbf{p}(\tau + \varepsilon/2) = \mathbf{p}(\tau) + \varepsilon\,\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(\tau)\}/2, \tag{3}$$

$$\boldsymbol{\theta}(\tau + \varepsilon) = \boldsymbol{\theta}(\tau) + \varepsilon\mathbf{M}^{-1}\mathbf{p}(\tau + \varepsilon/2), \tag{4}$$

$$\mathbf{p}(\tau + \varepsilon) = \mathbf{p}(\tau + \varepsilon/2) + \varepsilon\,\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(\tau + \varepsilon)\}/2. \tag{5}$$

Since the joint probability is factorizable (i.e., in physical terms, the Hamiltonian is separable), it is obvious by inspection that each complete leapfrog step (equations (3), (4) and (5)) is reversible by the negation of the integration step size ε. Likewise as the Jacobians of the transformations $(\boldsymbol{\theta}, \mathbf{p}) \mapsto (\boldsymbol{\theta}, \mathbf{p} + \varepsilon\,\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})/2)$ and $(\boldsymbol{\theta}, \mathbf{p}) \mapsto (\boldsymbol{\theta} + \varepsilon\mathbf{M}^{-1}\mathbf{p}, \mathbf{p})$ have unit determinant then volume is preserved. As total energy is only approximately conserved with the Stormer–Verlet integrator then a corresponding bias is introduced into the joint density which can be corrected by an accept–reject step. Owing to the volume preserving property of the integrator the determinant of the Jacobian matrix for the defined mapping does not need to be taken into account in the Hastings ratio of the acceptance probability. Therefore for a deterministic mapping $(\boldsymbol{\theta}, \mathbf{p}) \mapsto (\boldsymbol{\theta}^*, \mathbf{p}^*)$ obtained from a number of Stormer–Verlet integration steps the corresponding acceptance probability is $\min[1, \exp\{-H(\boldsymbol{\theta}^*, \mathbf{p}^*) + H(\boldsymbol{\theta}, \mathbf{p})\}]$, and owing to the reversibility of the dynamics the joint density and hence the marginals $p(\boldsymbol{\theta})$ and $p(\mathbf{p})$ are left invariant. If the integration error in the total energy is small then the acceptance probability will remain at a high level.

The overall HMC sampling from the invariant density $p(\boldsymbol{\theta})$ can be considered as a Gibbs sampler where the momentum $\mathbf{p}$ acts simply as an auxiliary variable drawn from a symmetric density

$$\mathbf{p}^{n+1}|\boldsymbol{\theta}^n \sim p(\mathbf{p}^{n+1}|\boldsymbol{\theta}^n) = p(\mathbf{p}^{n+1}) = \mathcal{N}(\mathbf{p}^{n+1}|\mathbf{0}, \mathbf{M}), \tag{6}$$

$$\boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1} \sim p(\boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1}) \tag{7}$$

where samples of $\boldsymbol{\theta}^{n+1}$ from $p(\boldsymbol{\theta}^{n+1}|\mathbf{p}^{n+1})$ are obtained by running the Stormer–Verlet integrator from initial values $\mathbf{p}^{n+1}$ and $\boldsymbol{\theta}^{n}$ for a certain number of steps to give proposed moves $\boldsymbol{\theta}^*$ and $\mathbf{p}^*$ and accepting or rejecting with probability $\min[1, \exp\{-H(\boldsymbol{\theta}^*, \mathbf{p}^*) + H(\boldsymbol{\theta}^n, \mathbf{p}^{n+1})\}]$. This Gibbs sampling scheme produces an ergodic, time reversible Markov chain satisfying detailed balance whose stationary marginal density is $p(\boldsymbol{\theta})$ (Duane *et al.*, 1987; Liu, 2001; Neal, 1996, 2010).

The combination of equations (3) and (4) in a single step of the integrator yields an update of the form

$$\boldsymbol{\theta}(\tau + \varepsilon) = \boldsymbol{\theta}(\tau) + \varepsilon^2/2\mathbf{M}^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(\tau)\} + \varepsilon\mathbf{M}^{-1}\mathbf{p}(\tau)$$

which is nothing more than a discrete preconditioned Langevin diffusion as employed in the MALA (Roberts and Stramer, 2003) (see Neal (1993a, 1996, 2010) for further discussion on this point). Viewed in this form it is clear that the choice of the mass matrix $\mathbf{M}$, as in the MALA, will be critical for the performance of HMC sampling, and like the MALA there is no guiding principle on how this should be chosen and tuned.

The demonstrated ability of HMC sampling to overcome random walks in MCMC sampling suggests that it should be a highly successful tool for Bayesian inference. A study suggests in excess of 300 citations of Duane *et al.* (1987) within the literature devoted to molecular modelling and simulation, physics and chemistry. However, there is a much smaller number of citations in the literature devoted to statistical methodology and application, e.g. Liu (2001), Neal (1993b, 1996), Gustafson (1997), Ishwaran (1999), Husmeier *et al.* (1999) and Hanson (2001), indicating that it has not been widely adopted as a practical inference method.

Although the choice of the step size ε and number of integration steps can be tuned on the basis of the overall acceptance rate of the HMC sampler, as already mentioned it is unclear how to select the values of the weight matrix $\mathbf{M}$ in any automated or principled manner that does not require some knowledge of the target density, similar to the situation with the MALA. Although heuristic rules of thumb have been suggested (Liu, 2001; Neal, 1993a, 1996, 2010) these typically rely on knowledge of the marginal variance of the target density, which is of course not known at the time of simulation and thus requires preliminary pilot runs of the HMC algorithm, this is also so for the MALA although asymptotic settings were suggested in Christensen *et al.* (2005). Sections 7–10 of this paper will demonstrate how crucial this tuning is to obtain acceptable performance of HMC methods and the MALA.

The potential of both the MALA and HMC methodology may be more fully realized by employing transitions that take into account the *local structure* of the target density when proposing moves to different probability regions, as this may improve the overall mixing of the chain. Therefore, rather than employing a fixed global covariance matrix in the proposal density $\mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$, a position-specific covariance could be adopted. Furthermore, the *deterministic* proposal mechanism of HMC sampling, when viewed as the deterministic component of the discrete preconditioned Langevin diffusion, relies on the gradient preconditioned by the inverse of a globally constant mass matrix. We turn our attention now to geometric concepts which will be shown to be of fundamental importance in addressing these shortcomings.

4. Exploiting geometric concepts in Markov chain Monte Carlo methods

The relationship between Riemann geometry and statistics has been employed in the development of, primarily asymptotic, statistical theory; see for example Murray and Rice (1993) and Barndorff-Nielsen *et al.* (1986). Geometric concepts of distance, curvature, manifolds, geodesics

and invariants are of natural interest in statistical methodology and in what follows we shall exploit some of these in the development of novel MCMC methods.

4.1. Fisher–Rao metric tensor

The formal definition of distance between two parameterized density functions $p(\mathbf{y};\boldsymbol{\theta})$ and $p(\mathbf{y};\boldsymbol{\theta}+\delta\boldsymbol{\theta})$ first appeared in Rao (1945) and took the quadratic form $\delta\boldsymbol{\theta}^{\mathrm{T}}\,\mathbf{G}(\boldsymbol{\theta})\,\delta\boldsymbol{\theta}$ where $\mathbf{G}(\boldsymbol{\theta})$ was shown to be equal to

$$-E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log\{p(\mathbf{y}|\boldsymbol{\theta})\}\right]=\mathrm{cov}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log\{p(\mathbf{y}|\boldsymbol{\theta})\}\right],$$

the expected Fisher information matrix. Rao noted that as the matrix $\mathbf{G}(\boldsymbol{\theta})$ is by definition positive definite it is a position-specific metric of a Riemann manifold. Therefore the space of parameterized probability density functions is endowed with a natural Riemann geometry. Given this geometry Rao went further and showed that expressions for the curvature of the manifold and geodesics on the manifold between two densities could be derived (Rao, 1945) and these ideas have been extended and formalized in the study of statistical inference, e.g. Amari and Nagaoka (2000), Kass (1989), Murray and Rice (1993), Barndorff–Nielsen *et al.* (1986), Critchley *et al.* (1993), Lauritzen (1987), Dawid (1975) and Efron (1975)). The Fisher metric also emerges from purely geometric arguments (Skilling, 2006) and it is straightforward to show for a probability simplex, $p^i \geqslant 0$, $\Sigma_{i=1}^{D} p^i = 1$, that the metric is $g_{ij} = \delta_{ij}/p^i$ where $\delta_{ij} = 1$ if and only if $i = j$. It then follows that a small displacement δl has squared length $(\delta l)^2 = \Sigma_{i,j}\,\delta p^i \delta p^j g_{ij} = \Sigma_i\,(\delta p^i)^2/p^i$, which is nothing more than the Fisher information matrix for a discrete probability distribution, suggesting this as the fundamental metric for probability spaces.

4.2. General form of metric tensor for Markov chain Monte Carlo methods

There are, however, many possible choices of metric for a specific manifold, each having different properties that may be of benefit in different forms of statistical analysis and specific applications. For example the motivating requirement for asymmetry in statistical inference is captured in the preferred point metric and associated geometry (Critchley *et al.*, 1993), whereas in Efron and Hinkley (1978) an argument is made for the use of the observed Fisher information matrix

$$-\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log\{p(\mathbf{y}|\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\mathrm{ML}}}$$

as an assessment of the conditional variance of a maximum likelihood estimator $\boldsymbol{\theta}^{\mathrm{ML}}$. For developing effective proposal mechanisms for MCMC sampling the potential utility of adopting the observed Fisher information matrix is intuitively apparent given that it is the negative Hessian of the log-probability at a specific point, although not strictly positive definite.

One can motivate the choice of the observed Fisher information matrix or indeed the empirical Fisher information matrix,

$$\widehat{\mathrm{cov}}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log\{p(\mathbf{y}|\boldsymbol{\theta})\}\right]$$

(the finite sample estimate of the covariance of the score function) for applications in MCMC methods for Bayesian inference where the metric is then conditional on the observed data rather than the asymptotic sampling mechanism. Indeed for many statistical models where the expected

Fisher information matrix is non-analytic, e.g. mixture models, the observed or empirical versions may define suitable, pragmatic, conditional manifolds for MCMC purposes.

It should be stressed that the MCMC methods which follow in this paper exploit the Riemann geometry that is induced by the metric defined by any arbitrary positive definite matrix $\mathbf{G}(\boldsymbol{\theta})$ and the practitioner is completely free in this choice. Indeed more general definitions of distance between densities such as Hellinger distance, or integrated squared distance, may be employed in deriving metrics to define a manifold if there is sufficient justification for their use in MCMC applications.

As a Bayesian perspective is adopted in this paper, the examples that are reported employ the joint probability of data and parameters when defining the metric tensor, i.e.

$$-E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log\{p(\mathbf{y},\boldsymbol{\theta})\}\right]$$

which is the expected Fisher information matrix plus the negative Hessian of the log-prior. For further discussion on ways to capture prior informativeness in the metric tensor see for example Tsutakawa (1972) and Ferreira (1981). Of course other choices could have been made but for illustration this suffices. The freedom to choose the metric does, however, open up a new line of investigation regarding the intrinsic geometry that is obtained by the choice and design of metrics and the characteristics which may make them appropriate for specific MCMC applications.

In summary, the parameter space of a statistical model is a Riemann manifold. Therefore the natural geometric structure of the density model $p(\boldsymbol{\theta})$ is defined by the Riemann manifold and associated metric tensor. Given this geometric structure of the parameter space of statistical models, the appropriate selection and adoption of a position-specific metric, $\mathbf{G}(\boldsymbol{\theta})$, within an MCMC scheme may yield more effective transitions that respect and exploit the geometry of the manifold in the overall algorithm. We now show how the Riemann manifold structure may be exploited within a correct MCMC framework for the MALA.

5. Riemann manifold Metropolis adjusted Langevin algorithm

Given the geometric structure for probability models a Langevin diffusion with invariant measure $p(\boldsymbol{\theta})$, $\boldsymbol{\theta}\in\mathbb{R}^D$, can be defined directly on a Riemann manifold with arbitrary metric tensor $\mathbf{G}(\boldsymbol{\theta})$ (Roberts and Stramer, 2003; Chung, 1982; Kent, 1978). The SDE defining the Langevin diffusion on the manifold is

$$\mathrm{d}\boldsymbol{\theta}(t)=\tfrac{1}{2}\tilde{\nabla}_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\}\,\mathrm{d}t+\mathrm{d}\tilde{\mathbf{b}}(t) \tag{8}$$

where the natural gradient (Amari and Nagaoka, 2000) is

$$\tilde{\nabla}_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\}=\mathbf{G}^{-1}\{\boldsymbol{\theta}(t)\}\,\nabla_{\boldsymbol{\theta}}\mathcal{L}\{\boldsymbol{\theta}(t)\}$$

and the Brownian motion on the Riemann manifold (Chung, 1982) is

$$\mathrm{d}\tilde{\mathbf{b}}_i(t)=|\mathbf{G}\{\boldsymbol{\theta}(t)\}|^{-1/2}\sum_{j=1}^{D}\frac{\partial}{\partial\boldsymbol{\theta}_j}[\mathbf{G}^{-1}\{\boldsymbol{\theta}(t)\}_{ij}|\mathbf{G}\{\boldsymbol{\theta}(t)\}|^{1/2}]\mathrm{d}t+[\sqrt{\mathbf{G}^{-1}}\{\boldsymbol{\theta}(t)\}\,\mathrm{d}\mathbf{b}(t)]_i. \tag{9}$$

Clearly in a Euclidean space where the metric tensor is an identity matrix equation (8) reduces to the standard form of SDE. The first term on the right-hand side of equation (9) relates to changes in local curvature of the manifold and reduces to 0 if curvature is everywhere constant.

The second right-hand term provides a position-specific axis alignment of the Brownian motion based on the local metric by transformation of the independent Brownian motion, $\mathbf{b}(t)$.

By expansion of the gradient term in equation (9) the discrete form of the above SDE employing a first-order Euler integrator provides a proposal mechanism which follows as

$$\begin{aligned}\boldsymbol{\theta}_i^* &= \boldsymbol{\theta}_i^n + \frac{\varepsilon^2}{2}\{\mathbf{G}^{-1}(\boldsymbol{\theta}^n)\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^n)\}_i - \varepsilon^2 \sum_{j=1}^{D}\left\{\mathbf{G}^{-1}(\boldsymbol{\theta}^n)\frac{\partial \mathbf{G}(\boldsymbol{\theta}^n)}{\partial \boldsymbol{\theta}_j}\mathbf{G}^{-1}(\boldsymbol{\theta}^n)\right\}_{ij} \\ &\quad + \frac{\varepsilon^2}{2}\sum_{j=1}^{D}\{\mathbf{G}^{-1}(\boldsymbol{\theta}^n)\}_{ij}\,\mathrm{tr}\left\{\mathbf{G}^{-1}(\boldsymbol{\theta}^n)\frac{\partial \mathbf{G}(\boldsymbol{\theta}^n)}{\partial \boldsymbol{\theta}_j}\right\} + \{\varepsilon\sqrt{\mathbf{G}^{-1}(\boldsymbol{\theta}^n)}\mathbf{z}^n\}_i \\ &= \boldsymbol{\mu}(\boldsymbol{\theta}^n, \varepsilon)_i + \{\varepsilon\sqrt{\mathbf{G}^{-1}(\boldsymbol{\theta}^n)}\mathbf{z}^n\}_i \end{aligned} \tag{10}$$

with proposal density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n) = \mathcal{N}\{\boldsymbol{\theta}^*|\boldsymbol{\mu}(\boldsymbol{\theta}^n, \varepsilon), \varepsilon^2\mathbf{G}^{-1}(\boldsymbol{\theta}^n)\}$ and standard acceptance probability $\min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^n|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^n)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^n)\}$ to ensure convergence to the invariant density $p(\boldsymbol{\theta})$. Immediately it is clear that the proposal mechanism makes moves in $\mathbb{R}^D$ according to the Riemann metric rather than according to the standard Euclidean distance. Pseudocode describing the full manifold MALA (MMALA) scheme is given in supplementary material that is available on request. For a manifold with constant curvature this reduces further to a position-specific preconditioned MALA proposal

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^n + \varepsilon^2\mathbf{G}^{-1}(\boldsymbol{\theta}^n)\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^n)/2 + \varepsilon\sqrt{\mathbf{G}^{-1}(\boldsymbol{\theta}^n)}\mathbf{z}^n.$$

Of course even if the curvature of the manifold is not constant the above simplified proposal mechanism, used in conjunction with the acceptance probability, will still define a correct MCMC method which converges to the target measure. However, dependent on the characteristics of the curvature the proposal process may not be so efficient in converging to the stationary distribution and this will be explored further in the experimental evaluation. To illustrate this geometric approach and to gain some insight into the MMALA a simple example is now given.

5.1. Illustrative example: parameters of a normal distribution

For N observations drawn from the normal distribution $\mathcal{N}(x|\mu, \sigma)$ the metric tensor based on the Fisher information matrix is $\mathbf{G}(\mu, \sigma) = \mathrm{diag}(N/\sigma^2, 2N/\sigma^2)$. Employing a flat prior on both parameters this metric defines a Riemann manifold with constant curvature which is a hyperbolic space on the upper half-plane that is defined by the horizontal and vertical co-ordinates μ and σ (Amari and Nagaoka, 2000). The distance between two densities $\mathcal{N}(x|\mu, \sigma)$ and $\mathcal{N}(x|\mu + \delta\mu, \sigma + \delta\sigma)$ as defined on this manifold is $(\delta\mu^2 + 2\delta\sigma^2)/\sigma^2$, indicating that, as the value of σ increases, the distance between the densities decreases. The first-order Euler approximations for the Langevin diffusion with invariant measure proportional to $\Pi_l \mathcal{N}(x_l|\mu, \sigma)$ follows as

$$\begin{aligned}\mu^{n+1} &= \mu^n + \frac{\varepsilon^2 m_1^n}{2(\sigma^n)^2} + \varepsilon z^n, \\ \sigma^{n+1} &= \sigma^n + \frac{\varepsilon^2 m_2^n}{2(\sigma^n)^3} - \frac{N\varepsilon^2}{2\sigma^n} + \varepsilon w^n \end{aligned} \tag{11}$$

where $m_1^n = \Sigma_l (x_l - \mu^n)$ and $m_2^n = \Sigma_l (x_l - \mu^n)^2$, with z^n and w^n standardized normal variates. When the diffusion is defined on the Riemann manifold then the approximate diffusion follows as

$$\mu^{n+1} = \mu^n + \frac{\varepsilon_m^2 m_1^n}{2N} + \frac{\varepsilon_m \sigma^n}{\sqrt{N}} z^n,$$
$$\sigma^{n+1} = \sigma^n + \frac{\varepsilon_m^2 m_2^n}{4N\sigma^n} - \frac{\varepsilon_m^2 \sigma^n}{4} + \frac{\varepsilon_m \sigma^n}{\sqrt{(2N)}} w^n. \tag{12}$$

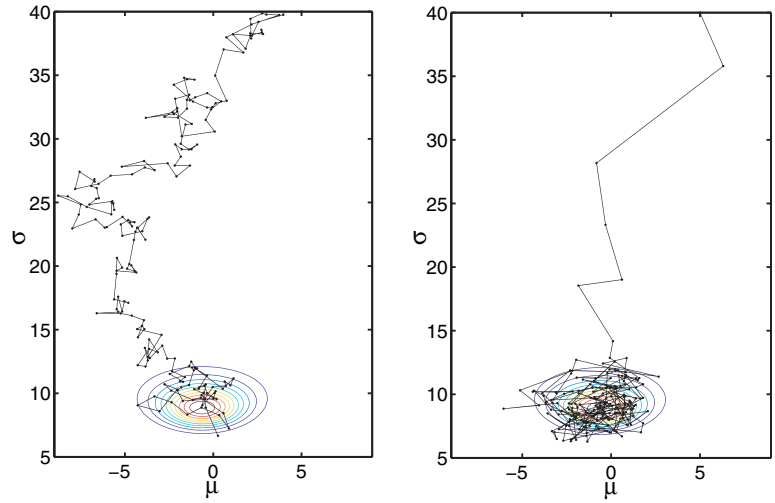

Fig. 1. Contours representing the sample estimate of $p(\mu, \sigma|X)$ where a sample of size $N = 30$ was drawn from $\mathcal{N}(X|\mu = 0, \sigma = 10)$ (both MALA and MMALA discrete diffusions were forward simulated from initial points $\mu_0 = 5$ and $\sigma_0 = 40$ with a step size $\varepsilon = 0.75$ for 200 steps): (a) sample path of the MALA proposal process (as the space is hyperbolic and a Euclidean metric is employed the proposals take inefficient steps of almost equal length thoughout); (b) MMALA proposals (in contrast, MMALA proposals are defined on the basis of the metric for the hyperbolic space with constant negative curvature and as such the distances covered by each step reflect the natural distances on the manifold, resulting in much more efficient traversal of the space)

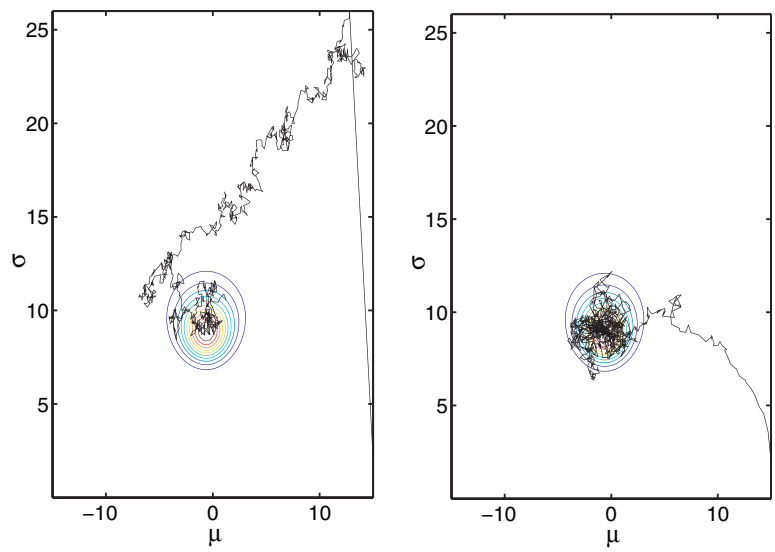

Fig. 2. Same data sample as in Fig. 1, but with $\mu_0 = 15$ and $\sigma_0 = 2$ (the step size is reduced to $\varepsilon = 0.2$ so that the MALA converges and 1000 proposal steps are taken): as in Fig. 1, from (a) it is clear that the Euclidean metric of the MALA does not exploit the hyperbolic geometry and overshoots dramatically at the start, whereas in (b) it is clear that the MMALA converges efficiently owing to the exploitation of the metric

The discrete diffusion based on a Euclidean metric (11) has diffusion terms εz^n and εw^n whose scaling is fixed by the integration step size ε irrespective of position. In contrast the approximate Langevin diffusion that is obtained by employing the Riemann metric tensor (12) produces terms $\varepsilon_m \sigma^n z^n / \sqrt{N}$ for the mean parameter and $\varepsilon_m \sigma^n w^n / \sqrt{(2N)}$ for the variance which are position dependent, thus ensuring appropriate scaling of the diffusion. The integration step size ε_m is effectively dimensionless whereas ε requires dimension proportional to σ^n, thus indicating proposal inefficiency with ε set at a fixed value as demonstrated in Figs 1 and 2. Extensive detailed investigation of the performance of the MMALA will be provided in the experimental sections.

6. Riemann manifold Hamiltonian Monte Carlo methods

Following on from the previous section the Hamiltonian which forms the basis of HMC sampling will now be defined in general form on a Riemann manifold. Zlochin and Baram (2001) originally attempted to exploit this manifold structure in HMC sampling; however, their use of a numerical integration method that did not guarantee reversibility or volume preservation prevented them from developing a correct MCMC procedure.

The definition of the Hamiltonian on a Riemann manifold is straightforward and is a technique that is employed in geometric mechanics to solve partial differential equations (Calin and Chang, 2004). From equation (2), it follows that $\mathbf{p} = \mathbf{M}\dot{\boldsymbol{\theta}}$, so the squared norm of each $\dot{\boldsymbol{\theta}}$ under the metric $\mathbf{M}$ is $\|\dot{\boldsymbol{\theta}}\|^2_{\mathbf{M}} = \dot{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{M}\dot{\boldsymbol{\theta}} = \mathbf{p}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{p}$. In a more general form, as the statistical model is defined on a Riemann manifold, the metric tensor defines the position-specific squared norm such that $\|\dot{\boldsymbol{\theta}}\|^2_{\mathbf{G}(\boldsymbol{\theta})} = \dot{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{G}(\boldsymbol{\theta})\dot{\boldsymbol{\theta}} = \mathbf{p}^{\mathrm{T}}\mathbf{G}^{-1}(\boldsymbol{\theta})\mathbf{p}$ and thus the kinetic energy term can be defined via the inverse metric (Calin and Chang, 2004). To ensure that the Hamiltonian can be interpreted as a log-density and that the desired marginal density for $\boldsymbol{\theta}$ is obtained, the addition of the normalizing constant for the Gaussian distribution is included in the potential energy term. Therefore, the Hamiltonian defined on the Riemann manifold follows as

$$H(\boldsymbol{\theta}, \mathbf{p}) = -\mathcal{L}(\boldsymbol{\theta}) + \tfrac{1}{2}\log\{(2\pi)^D |\mathbf{G}(\boldsymbol{\theta}|\} + \tfrac{1}{2}\mathbf{p}^{\mathrm{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p} \tag{13}$$

so that $\exp\{-H(\boldsymbol{\theta}, \mathbf{p})\} = p(\boldsymbol{\theta}, \mathbf{p}) = p(\boldsymbol{\theta})\, p(\mathbf{p}|\boldsymbol{\theta})$ and the marginal density

$$p(\boldsymbol{\theta}) \propto \int \exp\{-H(\boldsymbol{\theta}, \mathbf{p})\}\mathrm{d}\mathbf{p} = \frac{\exp\{\mathcal{L}(\boldsymbol{\theta})\}}{\sqrt{\{2\pi^D |\mathbf{G}(\boldsymbol{\theta})|\}}} \int \exp\{-\tfrac{1}{2}\mathbf{p}^{\mathrm{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\}\,\mathrm{d}\mathbf{p} = \exp\{\mathcal{L}(\boldsymbol{\theta})\}$$

is the desired target density.

Unlike the previous case for HMC sampling this joint density is no longer factorizable and therefore the log-probability does not correspond to a separable Hamiltonian. The conditional distribution for momentum values given parameter values is a zero-mean Gaussian distribution with the point-specific metric tensor acting as the covariance matrix $p(\mathbf{p}|\boldsymbol{\theta}) = \mathcal{N}\{\mathbf{p}|\mathbf{0}, \mathbf{G}(\boldsymbol{\theta})\}$, which will resolve the scaling issues that are associated with HMC methods, as will be demonstrated in the following sections. The dynamics are defined by Hamilton's equations as

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}\tau} = \frac{\partial H}{\partial p_i} = \{\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}\}_i, \tag{14}$$

$$\frac{\mathrm{d}p_i}{\mathrm{d}\tau} = -\frac{\partial H}{\partial \theta_i} = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} - \frac{1}{2}\mathrm{tr}\left\{\mathbf{G}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{G}(\boldsymbol{\theta})}{\partial \theta_i}\right\} + \frac{1}{2}\mathbf{p}^{\mathrm{T}}\mathbf{G}(\boldsymbol{\theta})^{-1}\frac{\partial \mathbf{G}(\boldsymbol{\theta})}{\partial \theta_i}\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}. \tag{15}$$

The Hamiltonian dynamics on the manifold are simulated by solving the continuous time derivatives and it is straightforward to see that they satisfy Liouville's theorem of volume pres-

ervation (Leimkuhler and Reich, 2004). However, for the discrete integrator it is not so straightforward. Naively employing the discrete Stormer–Verlet leapfrog integrator (equations (3)–(5)) gives transformations of the form $(\boldsymbol{\theta}, \mathbf{p}) \mapsto (\boldsymbol{\theta}, \mathbf{p} - \varepsilon \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{p}))$ and $(\boldsymbol{\theta}, \mathbf{p}) \mapsto (\boldsymbol{\theta} + \varepsilon \nabla_{\mathbf{p}} H(\boldsymbol{\theta}, \mathbf{p}), \mathbf{p})$, neither of which admits a Jacobian with unit determinant. In addition, it is straightforward to see that reversibility for $\boldsymbol{\theta}$ and $\mathbf{p}$ is not satisfied for finite step size ε, as $\mathbf{G}\{\boldsymbol{\theta}(\tau)\} \neq \mathbf{G}\{\boldsymbol{\theta}(\tau + \varepsilon)\}$. Therefore proposals that are generated from this integrator will not satisfy detailed balance in an HMC scheme. What is required is a time reversible volume preserving numerical integrator for solving this non-separable Hamiltonian to ensure a correct MCMC algorithm. A second-order semiexplicit symmetric integrator that is symplectic can be formed by the composition of a first-order implicit Euler integrator with its corresponding adjoint method. This is referred to as the generalized leapfrog algorithm and because it is symmetric and symplectic it has the required properties of volume preservation and reversibility. See for example Hairer *et al.* (1997), pages 187–190, and Leimkuhler and Reich (2004), pages 81–87, for a detailed derivation and proofs.

$$\mathbf{p}\left(\tau + \frac{\varepsilon}{2}\right) = \mathbf{p}(\tau) - \frac{\varepsilon}{2} \nabla_{\boldsymbol{\theta}} H\left\{\boldsymbol{\theta}(\tau), \mathbf{p}\left(\tau + \frac{\varepsilon}{2}\right)\right\}, \tag{16}$$

$$\boldsymbol{\theta}(\tau + \varepsilon) = \boldsymbol{\theta}(\tau) + \frac{\varepsilon}{2}\left[\nabla_{\mathbf{p}} H\left\{\boldsymbol{\theta}(\tau), \mathbf{p}\left(\tau + \frac{\varepsilon}{2}\right)\right\} + \nabla_{\mathbf{p}} H\left\{\boldsymbol{\theta}(\tau + \varepsilon), \mathbf{p}\left(\tau + \frac{\varepsilon}{2}\right)\right\}\right], \tag{17}$$

$$\mathbf{p}(\tau + \varepsilon) = \mathbf{p}\left(\tau + \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2} \nabla_{\boldsymbol{\theta}} H\left\{\boldsymbol{\theta}(\tau + \varepsilon), \mathbf{pu}\left(\tau + \frac{\varepsilon}{2}\right)\right\}. \tag{18}$$

If the Hamiltonian is separable then the generalized leapfrog reduces to the standard Stormer–Verlet leapfrog integrator. For the case of interest where the Hamiltonian is non-separable then equations (16) and (17) are defined implicitly. These require to be solved and we employ simple fixed point iterations run to convergence for this (see Hairer *et al.* (1997), pages 325–334); typically five or six iterations were required in the experiments conducted. The repeated application of the above steps provides the means to obtain a deterministic proposal that is guided not only by the derivative information of the target density, as in HMC sampling or the MALA, but also exploits the local geometric structure of the manifold as determined by the metric tensor. Intuitively, comparing the two Hamiltonians (1) and (13) shows that the constant mass matrix $\mathbf{M}$, defining a globally constant metric, is now replaced with the position-specific metric, thus removing the requirement to tune the values of the elements of $\mathbf{M}$, which so dramatically affects the performance of HMC methods. Since the integration scheme that is detailed above is both time reversible and volume preserving, employing it as a proposal process provides a correct MCMC scheme satisfying detailed balance and convergence to the desired target density. The overall RMHMC scheme can once again be written as a Gibbs sampler

$$\mathbf{p}^{n+1} | \boldsymbol{\theta}^n \sim p(\mathbf{p}^{n+1} | \boldsymbol{\theta}^n) = \mathcal{N}\{\mathbf{p}^{n+1} | \mathbf{0}, \mathbf{G}(\boldsymbol{\theta}^n)\}, \tag{19}$$

$$\boldsymbol{\theta}^{n+1} | \mathbf{p}^{n+1} \sim p(\boldsymbol{\theta}^{n+1} | \mathbf{p}^{n+1}) \tag{20}$$

where samples $\boldsymbol{\theta}^{n+1}$ from $p(\boldsymbol{\theta}^{n+1} | \mathbf{p}^{n+1})$ are obtained by running the generalized leapfrog integrator from initial values $\mathbf{p}^{n+1}$ and $\boldsymbol{\theta}^n$ for a certain number of steps to give proposed moves $\boldsymbol{\theta}^*$ and $\mathbf{p}^*$ and accepting or rejecting with probability $\min[1, \exp\{-H(\boldsymbol{\theta}^*, \mathbf{p}^*) + H(\boldsymbol{\theta}^n, \mathbf{p}^{n+1})\}]$. As for standard HMC sampling this Gibbs sampling scheme produces an ergodic, time reversible Markov chain satisfying detailed balance and whose stationary marginal density is $p(\boldsymbol{\theta})$ (Duane *et al.*, 1987; Liu, 2001; Neal, 1996, 2010). However, in this case there is no need to select and tune the mass matrix manually as it is defined at each step by the underlying geometry. Pseudocode is provided in the supplementary material.

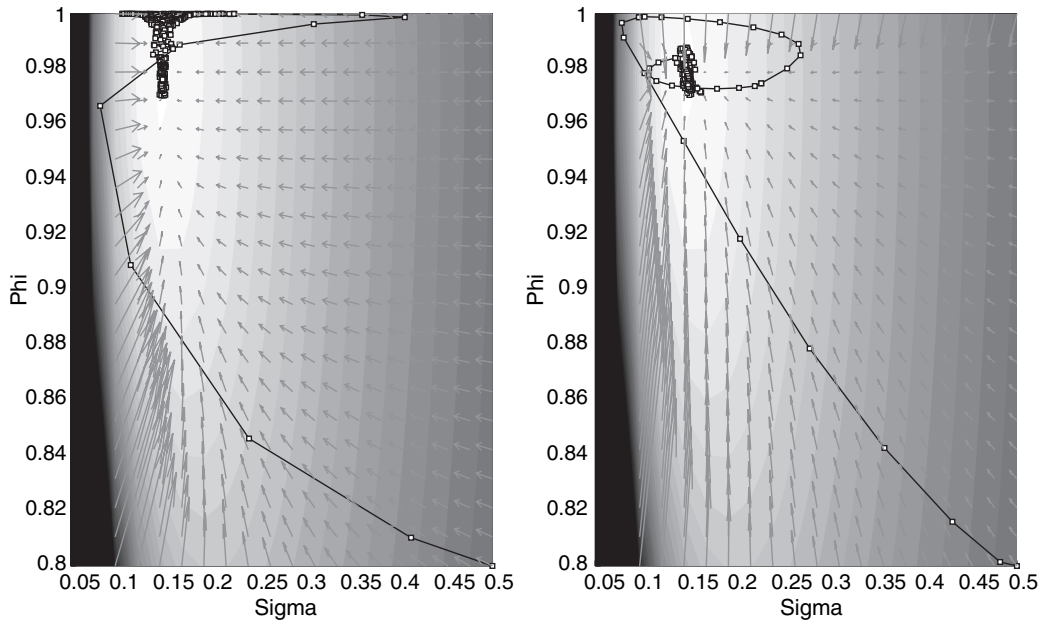

Fig. 3. Contours plotted from the stochastic volatility model investigated later in Section 8 (the latent volatilities and the parameter β are set to their true values, whereas the log-joint-probability given different values of the parameters σ and ϕ is shown by the contour plot): (a) evolution of a Markov chain by using HMC sampling with a unit mass matrix; (b) evolution of a chain from the same starting point by using RMHMC sampling (note how the use of the metric allows the RMHMC algorithm to converge much more quickly to the target density)

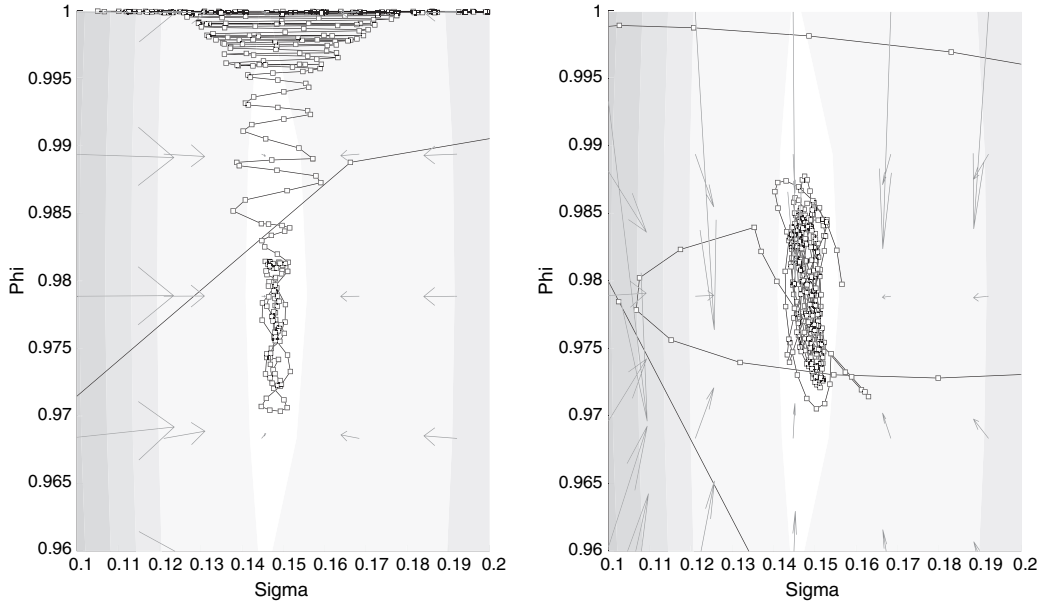

Fig. 4. Close-up of the Markov chain paths shown in Fig. 3: it is clear that RMHMC sampling effectively normalizes the gradients in each direction, whereas HMC sampling, with a unit mass matrix, exhibits stronger gradients along the horizontal direction compared with the vertical direction and therefore takes longer to converge to the target density; a carefully tuned mass matrix may improve HMC sampling, whereas RMHMC sampling deals with this automatically

An interesting point to note is that the Hamiltonian flow (solutions of the differential equations) for a purely kinetic Hamiltonian, i.e. in the absence of a potential energy term, is a geodesic flow (Calin and Chang, 2004). In other words paths that are produced by the solution of Hamilton's equations follow the geodesics (paths of least distance between points) on the manifold. For the case that we consider where there also is a potential term then the flows are locally geodesic (McCord *et al.*, 2002). This observation suggests an optimality, in terms of path length traversed across the manifold, for the RMHMC deterministic proposal mechanism. This presents an interesting area for further theoretical analysis and characterization of the properties of the RMHMC method.

Figs 3 and 4 provide an intuitive visual demonstration of the differences between HMC and RMHMC sampling when converging to and sampling from a target density. To illustrate the RMHMC sampling scheme and to evaluate performance against alternative MCMC methods, some example applications are now presented. We begin with posterior sampling for logistic regression models.

7. Manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo methods for Bayesian logistic regression

Consider an $N \times D$ design matrix $\mathbf{X}$ comprising N samples each with D covariates and a binary response variable $\mathbf{t} \in \{0, 1\}^N$. If we denote the logistic link function by $s(\cdot)$, a Bayesian logistic regression model of the binary response (Gelman *et al.*, 2004; Liu, 2001) is obtained by the introduction of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^D$ with an appropriate prior, which for illustration is given as $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$ where α is given. As already mentioned, throughout the practical examples, we form the metric tensor on the basis of the expected Fisher information plus the negative Hessian of the log-prior to include the effect of the prior on the geometry, although in this particular model the expected and observed Fisher information are the same. The metric tensor therefore follows as $\mathbf{G}(\boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}} \boldsymbol{\Lambda} \mathbf{X} + \alpha^{-1} \mathbf{I}$ where the diagonal $N \times N$ matrix $\boldsymbol{\Lambda}$ has elements $\boldsymbol{\Lambda}_{n,n} = s(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{n,\cdot}^{\mathrm{T}})\{1 - s(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{n,\cdot}^{\mathrm{T}})\}$ where $\mathbf{X}_{n,\cdot}$ denotes the vector that is the nth row of the $N \times D$ matrix $\mathbf{X}$. Finally the derivative matrices of the metric tensor take the form $\partial \mathbf{G}(\boldsymbol{\beta}) / \partial \beta_i = \mathbf{X}^{\mathrm{T}} \boldsymbol{\Lambda} \mathbf{V}^i \mathbf{X}$ where the $N \times N$ diagonal matrix $\mathbf{V}^i$ has elements $\{1 - 2s(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{n,\cdot}^{\mathrm{T}})\} X_{ni}$. The above identities are all that are required to define the RMHMC and MMALA sampling methods, which will be illustrated in the following experimental section.

7.1. Experimental results for Bayesian logistic regression

We present results from the analysis of five data sets (Michie *et al.*, 1994; Ripley, 1996), which are summarized in Table 1. These data sets exhibit a wide range of characteristics which provides a challenging test for any applied sampling method; the number of covariates ranges from 2 to 24, and the number of data points ranges from 250 to 1000. Although the manifold-based methods can easily cope with the raw data, we follow standard practice and normalize the data sets such that each covariate has zero mean and a standard deviation of 1. This allows a fair comparison with other sampling methods which would generally run into numerical problems with unnormalized data. We investigate the use of RMHMC methods and the MMALA applied to this problem and also implement the following sampling methods:

(a) componentwise adaptive Metropolis–Hastings methods (Robert (2004), chapter 7);
(b) joint updating Gibbs sampler (Holmes and Held, 2005);
(c) MALA (Roberts and Stramer, 2003);

Table 1. Summary of data sets for logistic regression

Name	*Covariates (D)*	*Data points (N)*	*Dimension of β (b)*
Pima Indian	7	532	8
Australian credit	14	690	15
German credit	24	1000	25
Heart	13	270	14
Ripley	2	250	7

(d) HMC sampling (Duane *et al.*, 1987; Neal, 1993a; Liu, 2001);
(e) iterated weighted least squares (IWLS) (Gamerman, 1997).

Given each data set we wish to sample from the posterior distribution over the regression coefficients $\boldsymbol{\beta}$, and in each experiment wide Gaussian prior distributions were employed such that $\pi(\beta_i) \sim \mathcal{N}(0, 100)$. A linear logistic regression model with intercept was used for each of the data sets with the exception of the Ripley data set, for which a cubic polynomial regression model was employed.

We reproduce the results of Holmes and Held (2005) by allowing 5000 burn-in iterations so that each sampler reaches the stationary distribution and has time to adapt as necessary. The next 5000 iterations were used to collect posterior samples for each of the methods and the central processor unit time that was required to collect these samples was recorded. Each method was implemented in the interpreted language MATLAB to ensure fair comparison. We compared the relative efficiency of these methods by calculating the effective sample size ESS using the posterior samples for each covariate, $\mathrm{ESS} = N\{1 + 2\Sigma_k \gamma(k)\}^{-1}$, where N is the number of posterior samples and $\Sigma_k \gamma(k)$ is the sum of the K monotone sample auto-correlations as estimated by the initial monotone sequence estimator (see Geyer (1992)). Such an approach was also taken by Holmes and Held (2005), in which they report the *mean* ESS, averaged over each of the covariates. However, we feel that this could give a rather inflated measure of the true ESS, since ideally we want a measure of the number of samples which are uncorrelated over *all* covariates. In this paper we therefore report the *minimum* ESS of the sampled covariates. This minimum ESS is then normalized relative to the central processor unit time by calculating the time that is taken to obtain one sample which is effectively uncorrelated across all covariates. The minimum, median and maximum ESS-values are obtained from each of the 10 runs and the reported values are averages of these results.

7.1.1. Metropolis–Hastings scheme

We employed an adaptive Metropolis–Hastings scheme, such that each covariate was updated individually with its step size being adapted every 100 iterations during burn-in to achieve an acceptance rate of between 20% and 40%. The step size was then fixed at the end of the burn-in period. With the Metropolis–Hastings algorithm subsampling or thinning is often employed in practice to improve ESS-values. Since our current measure of efficiency is time normalized, it automatically takes into account the trade-off between the additional computational cost of drawing more samples and the improved ESS that results. Simple simulations can show that the computational effort that is required to take additional steps through parameter space is generally greater than the benefit of increased ESS that results, such that the time that is taken

to produce one effectively independent sample increases as the number of discarded samples increases by using subsampling. In the main experiments we therefore compare the best case scenario which results from simply using all the available samples.

7.1.2. Auxiliary variable Gibbs sampler

The auxiliary variable Gibbs sampler of Holmes and Held (2005) was implemented with a joint update of $\{\mathbf{z}, \boldsymbol{\beta}\}$, where $\mathbf{z} \in \mathbb{R}^N$ is the auxiliary variable designed to improve mixing of the covariate samples. We implemented the algorithm based on the very detailed pseudocode that is given in the appendix of Holmes and Held (2005), and in contrast with the Metropolis–Hastings algorithm this method has the advantage of requiring no tuning of parameters. The main computational expense, however, is in the repeated sampling from truncated normal distributions, for which we implemented code based on the efficient method that was defined in Johnson *et al.* (1999).

7.1.3. Metropolis adjusted Langevin algorithm

We implemented an MALA sampler with proposed covariates being drawn from the multivariate normal distribution $\mathcal{N}\{\boldsymbol{\beta}^* | \boldsymbol{\mu}(\boldsymbol{\beta}^n, \varepsilon), \varepsilon^2 \mathbf{I}\}$ as defined previously. We follow the advice of Roberts and Rosenthal (1998) by scaling ε like $O(D^{-1/3})$, where D is the number of covariates, such that we achieve an acceptance rate of between 40% and 70%.

7.1.4. Hamiltonian Monte Carlo sampling

HMC sampling has promised to offer more efficient sampling from high dimensional probability distributions by effectively reducing the amount of random walk that is present in the parameter values being proposed. This has indeed been shown to be so for relatively simple, although high dimensional, multivariate normal distributions; however, there has been relatively little application to more complex data models, with the notable exception of Neal (1996). We believe that the reason for this lies in the amount of tuning that is required to obtain reasonable mixing and rates of acceptance, although there are heuristics for certain classes of models used for linear and non-linear regression (Neal, 1996). The two main parameters which require tuning are the number of leapfrog steps, N, and the size of each leapfrog step, ε. Setting different leapfrog step sizes along different directions can be equivalently encoded in the so-called mass matrix (Neal, 1993a, 1996). The use of exploratory runs of a Metropolis sampler to obtain initial estimates of the target distribution, and thus step sizes, has been suggested (Hajian, 2007); however, there is the obvious associated computational cost and the fact that this may not be feasible for very complex distributions.

In our experiments we employ 100 leapfrog steps and vary the step size manually for each data set to achieve an acceptance rate of between 70% and 90%. This requires a number of exploratory runs of the algorithm. The unit mass matrix that we employ works well for distributions in which the standard deviations of the posterior distributions are of similar magnitudes, as is the case here since we have normalized our data sets for fairest comparison. However, for models where no heuristic is available to set the step sizes or mass matrix and where the posterior distribution is highly correlated, the HMC algorithm soon becomes challenging to tune.

7.1.5. Iterated weighted least squares

We consider in addition the second-order method IWLS (Gamerman, 1997), which makes use of second derivatives in its Metropolis proposal steps. It should be noted that the term involving

the second derivatives for IWLS is indeed different from the metric tensor expression that is employed in RMHMC sampling and the MMALA, and we shall see how this impacts on the results shortly. This method is relatively straightforward to implement and has the advantage that it requires no tuning, similarly to the auxiliary variable Gibbs sampler of Holmes and Held (2005).

7.2. Comparison of Markov chain Monte Carlo methods

We begin by investigating the RMHMC method in detail for the more challenging of our five data sets, German credit, which consists of 24 covariates and 1000 data points. We then compare the results for all data sets by employing the alternative sampling methods that were described previously.

Since RMHMC sampling automatically adapts its (non-diagonal) mass matrix via the metric tensor depending on its current position, we consider fixing ε and adjusting the number of leapfrog steps. Table 2 shows the results of the generalized leapfrog integration scheme by using a variety of choices for these parameters. We found that sampling generally became more efficient as εL was increased, i.e. when the chain could traverse a greater distance. The value of 0.5 was found to be a suitable choice for ε, and the choice of six leapfrog steps was implemented for the data sets.

We find that the RMHMC and MMALA sampling methods work very well for a variety of data sets and RMHMC sampling is fairly robust to the choice of algorithm parameters. For comparison with the alternative sampling methods, we chose the settings for RMHMC sampling on the basis of the above analysis. The scaling for the MMALA was chosen to obtain an acceptance rate of around 70%. We repeated the sampling experiments 10 times and averaged the results, which are shown for each of the data sets in Tables 3–7.

All methods converged within 5000 burn-in iterations for all the normalized data sets. The manifold-based methods generally outperform their non-manifold counterparts, HMC sampling and the MALA, particularly for data sets which have stronger correlations between the covariates.

Fig. 5 shows the trace and auto-correlation plots for 1000 posterior samples by using the heart data set. The difference in auto-correlation is quite striking, both from inspection of the traces and from examination of the auto-correlation plots themselves. The auto-correlations of the RMHMC samples drop towards zero far quicker than for any of the other methods.

As the number of covariates in the data set increases, so the overall performance of RMHMC sampling and the MMALA decreases owing to the increased computational burden of calculating partial derivatives of the metric tensor with respect to each of the covariates. It is clear that

Table 2. RMHMC sampling with a generalized leapfrog integration scheme—investigating the effect of parameter settings on sampling efficiency with the German credit data set

εL	*Maximum* ε	*Mean time (s)*	*Minimum ESS*	*s/minimum ESS*
1	0.5	131.7	674	0.195
2	0.5	193.6	2139	0.090
3	0.5	287.9	4791	0.060

Table 3. Australian credit data set ($D = 14$, $N = 690$ and 15 regression coefficients)—comparison of sampling methods

Method	*Time*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	10.8	(314, 709, 979)	0.034	320
Auxiliary variables	407.5	(7.5, 1054, 1405)	10.9	1
MALA	22.3	(22.3, 576.8, 990.6)	0.122	89.3
HMC	87.3	(3197, 3612, 3982)	0.027	403
IWLS	4.7	(3.6, 9.3, 73.7)	1.31	8.3
MMALA	11.7	(702, 867, 1037)	0.0167	652
Simplified MMALA	3.2	(487, 625, 746)	0.006	1817
RMHMC	81.7	(4975, 5000, 5000)	0.016	681
RMHMC (Student t)	87.3	(1083, 1625, 2002)	0.081	134

Table 4. German credit data set ($D = 24$, $N = 1000$ and 25 regression coefficients)—comparison of sampling methods

Method	*Time*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	23.4	(167, 613, 1015)	0.140	13.3
Auxiliary variables	618.8	(1006, 2211, 2640)	0.614	3
MALA	3.5	(95.5, 316, 667)	0.037	50.3
HMC	117.9	(3182, 3632, 3986)	0.037	50.3
IWLS	7.8	(4.2, 9.9, 69)	1.862	1
MMALA	42.3	(604, 766, 902)	0.070	26.6
Simplified MMALA	5.0	(435, 615, 747)	0.012	155
RMHMC	246.6	(4757, 5000, 5000)	0.052	35.8
RMHMC (Student t)	257.4	(3981, 4934, 5000)	0.065	28.6

Table 5. Pima Indian data set ($D = 7$, $N = 532$ and eight regression coefficients)—comparison of sampling methods

Method	*Time*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	3.8	(347, 552, 980)	0.011	35.1
Auxiliary variables	304.3	(1432, 1888, 2295)	0.212	1.8
MALA	1.56	(316, 550, 895)	0.005	77.2
HMC	45.7	(3265, 3605, 3893)	0.014	27.6
IWLS	2.9	(7.5, 14, 171)	0.386	1
MMALA	4.2	(1135, 1286, 1412)	0.0037	104
Simplified MMALA	1.9	(1046, 1160, 1300)	0.0018	214
RMHMC	34.4	(5000, 5000, 5000)	0.0069	55.9
RMHMC (Student t)	38.6	(3928, 4432, 4688)	0.0098	39.4

for logistic regression problems with a very high number of covariates, e.g. in excess of 100, the use of manifold methods will become computationally infeasible if the metric tensor is position dependent and the same number of derivative matrices as covariates must be computed and manipulated. From a practical perspective we can impose a constant curvature manifold on the logistic regression model by employing a constant metric tensor of the form $\mathbf{G} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \alpha^{-1}\mathbf{I}$

Table 6. Heart data set ($D = 13$, $N = 270$ and 14 regression coefficients)—comparison of sampling methods

Method	*Time*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	4.4	(418, 637, 905)	0.010	85
Auxiliary variables	150.9	(711, 1233, 1676)	0.212	4
MALA	1.1	(279, 524, 814)	0.0038	223
HMC	27.6	(3246, 3647, 4003)	0.0085	100
IWLS	2.8	(3.3, 6.2, 27)	0.85	1
MMALA	5.6	(656, 789, 903)	0.0085	100
Simplified MMALA	1.6	(371, 481, 617)	0.0043	197
RMHMC	42.2	(4862, 5000, 5000)	0.0087	97.7
RMHMC (Student t)	48.0	(2603, 2904, 3171)	0.018	47.2

Table 7. Ripley data set ($D = 2$, $N = 250$ and seven regression coefficients)—comparison of sampling methods

Method	*Time*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	2.1	(59, 99, 271)	0.035	201
Auxiliary variables	139.6	(19, 44, 283)	7.06	1
MALA	0.97	(33, 58, 101)	0.029	243
HMC	24.8	(3326, 3719, 4053)	0.0076	928
IWLS	2.3	(6, 11, 26)	0.39	19.6
MMALA	3.3	(447, 579, 685)	0.0075	941
Simplified MMALA	1.3	(291, 403, 473)	0.0045	1569
RMHMC	28.0	(4273, 4677, 4961)	0.0065	1086
RMHMC (Student t)	31.9	(2829, 3088, 3289)	0.011	641

emerging from the linear regression model. This captures the correlation structure in the covariates and as such provides an obvious decorrelating operator for the manifold methods. This will now have the same computational burden as the non-manifold methods.

We consider also an alternative second-order method, IWLS, which makes use of terms involving second derivatives and therefore some measure of the curvature of the parameter space. IWLS performs fairly poorly; indeed in the examples it performs about the same as parameterwise Metropolis sampling.

7.3. Comparison of manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo variants

We now investigate variants of RMHMC sampling and the MMALA to see whether results may be improved on the basis of slight alterations to the standard forms. We first consider a simplified version of the MMALA, which assumes a locally flat metric tensor during each Metropolis step and will still converge to the stationary distribution owing to the Metropolis adjustment. It is clear that this is computationally much less expensive than the full MMALA as it avoids the calculation of metric tensor derivatives. It is interesting that the simplified MMALA has worse ESS than the complete MMALA, which intuitively makes sense since proposed steps across

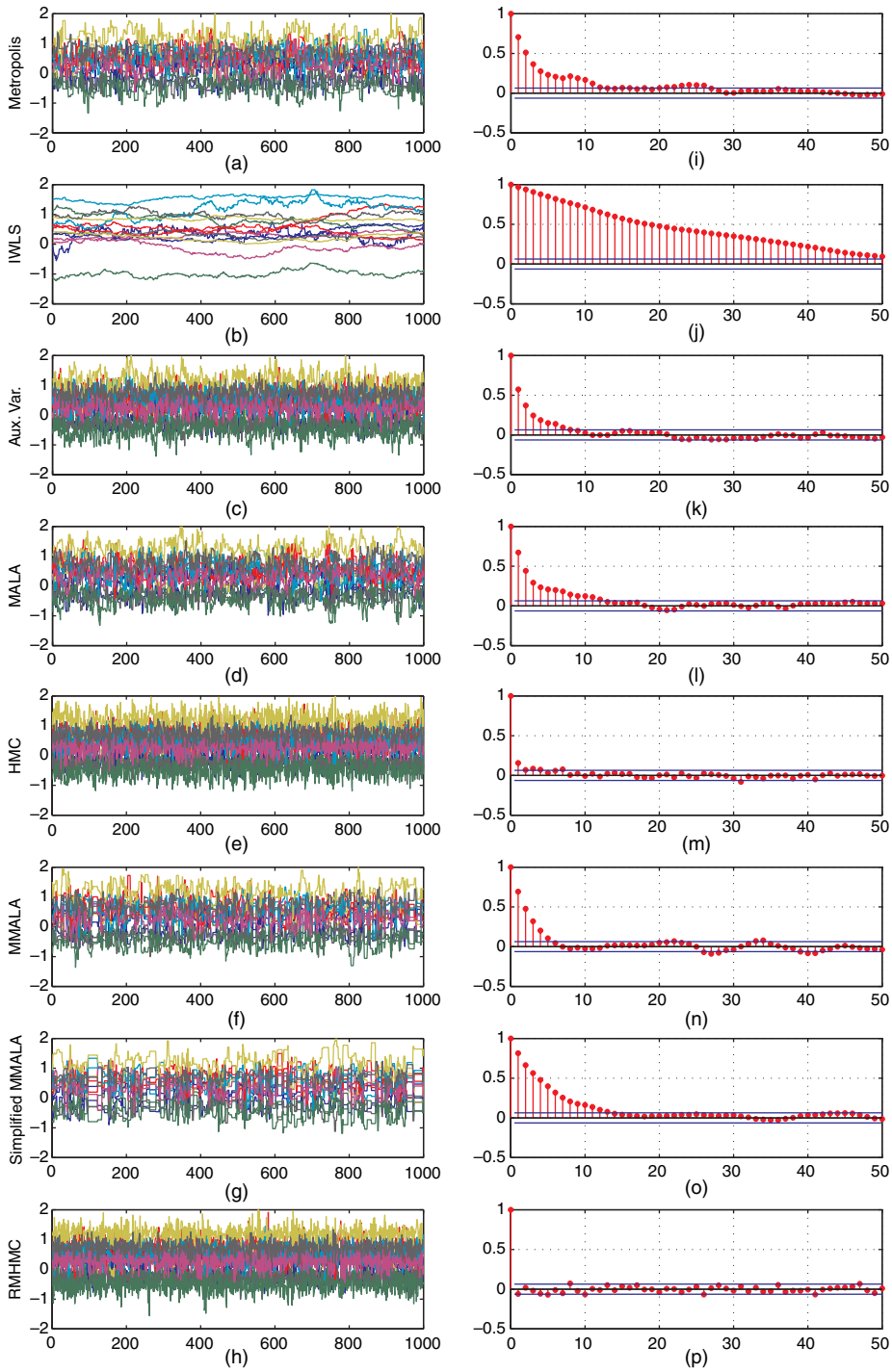

Fig. 5. Trace plots for 1000 posterior samples for the heart data set by using (a) Metropolis sampling, (b) IWLS, (c) the auxiliary variable sampler, (d) the standard MALA, (e) standard HMC sampling, (f) the MMALA, (g) the simplified MMALA and (h) RMHMC sampling: (i)–(p) corresponding auto-correlation plots for the first sampled covariate

the manifold will have greater error by not taking into account any changes in curvature. The time-normalized ESS, however, is much better, as the computational complexity is far less.

It is also interesting to investigate the use of an alternative kinetic energy function in RMHMC sampling; this idea was also briefly mentioned in Liu (2001) although no example was given. We consider therefore the use of a kinetic energy term based on the Student t-density, with the idea that, since the heavy tails might occasionally mean that a larger momentum is sampled, this could plausibly result in less correlated samples of the target distribution. We note that, since the multivariate Student t-distribution is symmetric, then the resulting Hamiltonian is still reversible. The simulations take slightly longer to run than with standard Gaussian-distributed momentum using the same integration time steps. This is due simply to the increased computation that is required to sample from a Student t-distribution, and also to the more involved computation that is required to calculate the dynamics of this new Hamiltonian. The results show that the ESS is actually significantly less than that of a Hamiltonian defined with Gaussian momentum. This is possibly a result of a higher concentration of mass producing momenta with values closer to zero, even though there will be occasional samples of momentum with much larger magnitude.

In our simulations, manifold-based methods perform extremely well compared with the other methods using small to medium-sized data sets. It is interesting to note that, owing to the dense matrix form of the metric tensor and its inverse, the computational cost of the MMALA and RMHMC sampling on Bayesian logistic regression will not scale favourably and it can be seen that their time-normalized efficiency does indeed decrease as the number of regression coefficients in the data set increases. This issue of scaling can, however, be eased somewhat by employing simplified MMALA sampling, which assumes a locally constant metric tensor, thus avoiding expensive computation of the derivatives of the metric tensor and for RMHMC sampling a globally constant metric based on the linear regression metric. A further, more complex, example based on a stochastic latent volatility model is now considered where the metric tensor and its inverse are sparse, permitting scaling of RMHMC sampling to very high dimensions.

8. Manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo sampling for a stochastic volatility model

A stochastic volatility model that was studied in Liu (2001) and Kim *et al.* (1998) is defined with the latent volatilities taking the form of an auto-regressive AR(1) process such that $y_t = \varepsilon_t \beta \exp(x_t/2)$ with $x_{t+1} = \phi x_t + \eta_{t+1}$ where $\varepsilon_t \sim \mathcal{N}(0,1)$, $\eta_t \sim \mathcal{N}(0, \sigma^2)$ and $x_1 \sim \mathcal{N}\{0, \sigma^2/(1-\phi^2)\}$ having joint probability

$$p(\mathbf{y}, \mathbf{x}, \beta, \phi, \sigma) = \prod_{t=1}^{T} p(y_t|x_t, \beta)\, p(x_1) \prod_{t=2}^{T} p(x_t|x_{t-1}, \phi, \sigma)\, \pi(\beta)\, \pi(\phi)\, \pi(\sigma). \tag{21}$$

We split up the sampling procedure into two steps, which as will be seen allow the implementation of both the MMALA and RMHMC sampling in a computationally efficient manner. Firstly we simulate ϕ, σ and β from $p(\beta, \phi, \sigma|\mathbf{y}, \mathbf{x})$, where the priors are chosen to be $p(\beta) \propto 1/\beta$, $\sigma^2 \sim \text{Inv-}\chi^2(10, 0.05)$ and $(\phi+1)/2 \sim \text{beta}(20, 1.5)$. One way to deal with the constraints on the values ϕ and σ is to implement a transformation of these to the real line, which we do by letting $\sigma = \exp(\gamma)$ and $\phi = \tanh(\alpha)$, and noting that this introduces a Jacobian factor into the acceptance ratio in the standard manner. Secondly we sample the latent volatilities by simulating from the conditional $p(\mathbf{x}|\mathbf{y}, \beta, \sigma, \phi)$. We shall consider the use of the MMALA, RMHMC sampling, the MALA and HMC sampling for the purpose of sampling both the parameters and the latent volatilities.

8.1. Manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo sampling for stochastic volatility model parameters

We require the partial derivatives of the joint log-probability with respect to the transformed parameters to implement the MALA and HMC sampling, as well as expressions for the metric tensor and its partial derivatives, to employ the MMALA and RMHMC algorithm. All these quantities may be obtained straightforwardly (see Appendix A for details). We then use these methods to draw samples from the conditional posterior $p(\beta, \alpha, \gamma|\mathbf{y}, \mathbf{x},)$.

8.2. Manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo sampling for stochastic volatility model latent volatilities

Defining $\mathbf{u} = (x_3, \ldots, x_T)^{\mathrm{T}}$, $\mathbf{v} = (x_2, \ldots, x_{T-1})^{\mathrm{T}}$, $\mathbf{w} = (\phi/\sigma^2)(\mathbf{u} - \phi\mathbf{v})$, $\mathbf{s} = (s_1, \ldots, s_T)^{\mathrm{T}}$ such that $s_i = 0.5\{1 - y_i^2\beta^{-2}\exp(-x_i)\}$, $\delta_1 = -\sigma^{-2}(x_1 - \phi x_2)$ and $\delta_T = -\sigma^{-2}(x_T - \phi x_{T-1})$, we define the vector $\mathbf{r} = (\delta_1, \mathbf{w}^{\mathrm{T}}, \delta_T)^{\mathrm{T}}$ and the gradient $\nabla_{\mathbf{x}} \log\{p(\mathbf{y}, \mathbf{x}|\beta, \phi, \sigma)\} = \mathbf{s} - \mathbf{r}$.

To devise an MMALA and RMHMC sampler for the latent volatilities $\mathbf{x}$, we also require an expression for the metric tensor and its partial derivatives with respect to the latent volatilities. For the data probability of the model, $p(\mathbf{y}|\mathbf{x}, \beta)$, the expected Fisher information matrix is the scaled identity matrix $\frac{1}{2}\mathbf{I}$. The latent volatility is an AR(1) process having covariance matrix $\mathbf{C}$ with elements $E(x_{t+n}x_t) = \phi^{|n|}\sigma^2/(1 - \phi^2)$ and as in the previous examples the metric tensor is defined as the sum of the expected Fisher information matrix and the negative Hessian of the log-prior, $\mathbf{G} = \frac{1}{2}\mathbf{I} + \mathbf{C}^{-1}$, conditional on current values of σ, ϕ and β. Now the expression for the covariance matrix is completely dense and is therefore computationally expensive to manipulate. Fortunately, this AR(1) process admits a simple analytical expression for the precision matrix in the form of a sparse tridiagonal matrix, such that the diagonal elements are equal to $(1 + \phi^2)/\sigma^2$, with the exception of the first and last diagonal elements which are equal to $1/\sigma^2$, and the superdiagonal and subdiagonal elements are equal to $-\phi/\sigma^2$. Thus the metric tensor also has a tridiagonal form. For large numbers of observations this sparse structure allows great gains in computational efficiency, since the inverse of this tridiagonal metric tensor may be computed in $\mathcal{O}(T)$ as opposed to the usual $\mathcal{O}(T^3)$. We note that computationally efficient methods for manipulating tridiagonal matrices are automatically implemented by the standard routines in MATLAB.

We notice that the metric tensor in this case is not a function of the latent volatilities $\mathbf{x}$ and so the associated partial derivatives with respect to the latent volatilities are zero. In this case as the manifold has constant curvature the RMHMC scheme is effectively an HMC scheme with mass matrix $\mathbf{M}$ now defined, based on the Riemann geometric principles, by the globally constant metric tensor $\mathbf{G}$. Likewise the MMALA collapses to an MALA scheme preconditioned by the constant matrix $\mathbf{G}^{-1}$. It is clear that this preconditioning will improve both the mixing and overall ESS; see Lambert and Eilers (2009) for a recent application of this type of preconditioning in the MALA. We point out that, as in the case of RMHMC sampling, the preconditioning matrix emerges naturally from the underlying geometric principles.

8.3. Experimental results for stochastic volatility model

We now compare the computational efficiency of RMHMC sampling, the MMALA, HMC sampling and the MALA for sampling both the parameters and the latent variables of the stochastic volatility model as previously defined: Tables 8 and 9. 2000 observations were simulated from the model with the parameter values $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$ as given in Liu (2001). Using these data, 20 000 posterior samples were collected after a burn-in period of 10 000 samples. This sampling procedure was repeated 10 times (Fig. 6). The efficiency was compared in

Table 8. 2000 simulated observations with $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$—comparison of sampling the parameters β, σ and ϕ after 20000 posterior samples averaged over 10 runs

Method	*Mean time*	*ESS* (β, σ, ϕ)	*Standard error* (β, σ, ϕ)	*s/minimum ESS*	*Relative speed*
MALA	44.0	(19.1, 11.3, 30.1)	(1.9,0.8,2.1)	3.89	36.7
HMC	424.8	(117, 81, 198)	(9.3, 3.1, 10.3)	5.19	27.5
MMALA	2455.9	(17.2, 17.4, 44.5)	(2.8, 2.4, 9.2)	142.8	1
RMHMC	329.4	(325, 139, 344)	(19.0, 7.3, 25.2)	2.37	60.3

Table 9. 2000 simulated observations with $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$—comparison of sampling the latent volatilities after 20000 posterior samples averaged over 10 runs

Method	*Mean time*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
MALA	44.0	(9.7, 16.7, 28.4)	4.53	7.5
HMC	424.8	(409, 624, 1239)	1.04	32.9
MMALA	2455.9	(71.8, 131.0, 329.8)	34.2	1
RMHMC	329.4	(977, 1689, 3376)	0.34	100.6

terms of time-normalized ESS, as in the previous section, for the parameters and the latent volatilities. The MALA was tuned such that the acceptance ratio was between 40% and 70%, and it was necessary to use a tuning for the transient phase that was different from that for the stationary phase. HMC sampling was implemented again by using 100 leapfrog steps and tuning the step size to obtain an acceptance rate of between 70% and 90%, which resulted in a step size of 0.015 for hyperparameters and a step size of 0.03 for the latent volatilities. RMHMC sampling was implemented by using a step size of 0.5 and six integration steps per parameter proposal, and a step size of 0.1 and 50 integration steps per volatility proposal.

In terms of sampling the hyperparameters, manifold methods offer little advantage over standard sampling approaches owing to the small dimensionality of the problem. RMHMC sampling and the MALA give the best performance in terms of time-normalized ESS. The MALA exhibits a very poor ESS; however, the computation time is also extremely small compared with the other two methods. RMHMC sampling has the highest raw ESS but has much more computational overhead compared with the MALA. When we consider sampling the latent variable, RMHMC sampling offers greater advantages. In particular, it runs faster than HMC sampling, partly because of the computationally efficient tridiagonal structure of the metric tensor and partly because RMHMC sampling follows the natural tensor gradient through the parameter space and requires significantly fewer leapfrog iterations to explore the target density. See Figs 3 and 4 for an illustration of the contrast between HMC and RMHMC sampling of the parameters of this model. In this example, the MMALA performs very badly owing to the need to take a Cholesky decomposition of the inverse metric tensor of the latent variables, which is a dense matrix, compared with RMHMC sampling, which only requires use of the tridiagonal metric tensor. It should be noted that RMHMC sampling again requires very little tuning compared with the other methods; unlike the MALA it does not require different tuning in different parts

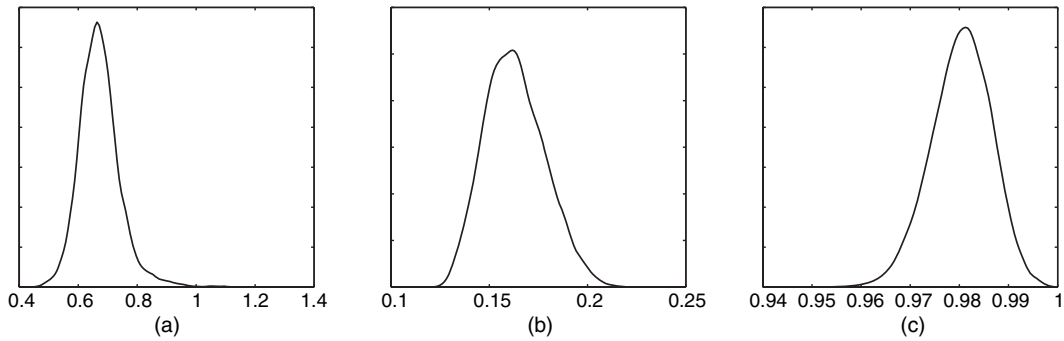

Fig. 6. Posterior marginal densities for (a) β, (b) σ and (c) ϕ, employing RMHMC sampling to draw 20 000 samples of the parameters and latent volatilities by using a simulated data set consisting of 2000 observations: the true values are $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$

of the parameter space, and unlike HMC sampling it requires no manual setting of a mass matrix. It would be interesting to consider the use of the MMALA and RMHMC sampling as a part of the particle MCMC methodology (Andrieu *et al.*, 2010) for this particular model.

We now consider an example where the target density is extremely high dimensional, which is encountered when performing inference using spatial data modelled by a log-Gaussian Cox process.

9. Manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo sampling for log-Gaussian Cox point processes

RMHMC sampling and the MMALA are further studied by using the example of inference in a log-Gaussian Cox point process as detailed in Christensen *et al.* (2005). This is a particularly useful example in that the target density is of high dimension with strong correlations and provides a severe test of MCMC capability. The data, model and experimental protocol as described in Christensen *et al.* (2005) are adopted here. A 64×64 grid is overlaid on the area $[0, 1]^2$ with the number of points in each grid cell denoted by the random variables $\mathbf{Y} = \{Y_{i,j}\}$ which are assumed conditionally independent, given a latent intensity process $\Lambda(\cdot) = \{\Lambda(i, j)\}$, and are Poisson distributed with means $m\,\Lambda(i, j) = m \exp(X_{i,j})$, where $m = 1/4096$, $\mathbf{X} = \{X_{i,j}\}$, $\mathbf{x} = \mathrm{vec}(\mathbf{X})$, and $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, with $\mathbf{X}$ a Gaussian process having mean $E(\mathbf{x}) = \mu\mathbf{1}$, and covariance function $\boldsymbol{\Sigma}_{(i,j),(i',j')} = \sigma^2 \exp\{-\delta(i, i', j, j')/64\beta\}$, where $\delta(i, i', j, j') = \sqrt{\{(i - i')^2 + (j - j')^2\}}$. The joint density is

$$p(\mathbf{y}, \mathbf{x}|\mu, \sigma, \beta) \propto \prod_{i,j} \exp\{y_{i,j} x_{i,j} - m \exp(x_{i,j})\} \exp\{-(\mathbf{x} - \mu\mathbf{1})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mu\mathbf{1})/2\}. \tag{22}$$

As in the previous example an overall Gibbs scheme in which we alternately sample from $p(\mathbf{x}|\mathbf{y}, \sigma, \beta, \mu)$ and $p(\sigma, \beta|\mathbf{y}, \mathbf{x}, \mu)$ is considered. If we let $\mathcal{L} \equiv \log\{p(\mathbf{y}, \mathbf{x}|\mu, \sigma, \beta)\}$ and $\mathbf{e} = m \exp(x_{i,j})$, then the derivative with respect to the latent variables follows straightforwardly as $\nabla_{\mathbf{x}}\mathcal{L} = \mathbf{y} - \mathbf{e} - \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu\mathbf{1})$, and $-E_{\mathbf{y},\mathbf{x}|\boldsymbol{\theta}}(\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}\mathcal{L}) = \boldsymbol{\Lambda} + \boldsymbol{\Sigma}^{-1}$, where the diagonal matrix $\boldsymbol{\Lambda}$, whose ith diagonal element is defined as $m \exp\{\mu + (\boldsymbol{\Sigma})_{ii}\}$, follows from the expectation of the exponential of normal random variables. The metric tensor describing the manifold for the random field is constant, $\mathbf{G} = \boldsymbol{\Lambda} + \boldsymbol{\Sigma}^{-1}$, and the MMALA and RMHMC schemes for the conditional, $p(\mathbf{x}|\mathbf{y}, \sigma, \beta, \mu)$, are basically the MALA, HMC sampling with mass and preconditioning matrices $\mathbf{M} = \boldsymbol{\Lambda} + \boldsymbol{\Sigma}^{-1}$ and $\mathbf{M}^{-1}$. The computational cost of calculating the required inverse of the metric tensor scales as $\mathcal{O}(N^3)$; however, once this quantity has been calculated, for

HMC sampling a large number of leapfrog steps may be made with little additional overhead, which as we shall see results in very efficient sampling of the latent variables.

To sample from the conditional $p(\sigma, \beta|\mathbf{y}, \mathbf{x}, \mu)$ we employ a metric tensor based on the expected Fisher information for the parameters $\boldsymbol{\theta} = [\sigma, \beta]$ which follows as $\mathbf{D}_{\boldsymbol{\theta}}$ whose (l, m)th element is

$$\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_l}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_m}\right).$$

See Appendix B for details.

Since the metric tensor for the latent variables has dimension $N \times N$, where $N = 4096$, the $\mathcal{O}(N^3)$ operations that are required in the MMALA and RMHMC schemes will clearly be computationally costly. However, it should also be noted that, in previous studies of this log-Gaussian Cox process (Christensen *et al.*, 2005), a transformation of the latent Gaussian field is necessary based on the Cholesky decomposition of $\boldsymbol{\Sigma}^{-1} + \mathrm{diag}(\mathbf{x})$, which will therefore also scale as $\mathcal{O}(N^3)$.

It is possible to consider jointly sampling the hyperparameters and the latent variables. Now with $\mathcal{L} \equiv \log\{p(\mathbf{y}, \mathbf{x}, \sigma, \beta|\mu)\}$, we see that the expected Fisher information matrix is block diagonal with blocks $\boldsymbol{\Lambda} + \boldsymbol{\Sigma}^{-1}$ and $\mathbf{D}_{\boldsymbol{\theta}}^{-1}$. Unfortunately, jointly sampling the latent variables and the hyperparameters proves to be computationally too costly to implement, as the metric tensor is now no longer fixed and so the generalized leapfrog integration scheme must be implemented in RMHMC sampling with fixed point iterations, during each of which the metric tensor and its inverse must be recalculated.

9.1. Experimental results for Log-Gaussian Cox processes

Following the example given by Christensen *et al.* (2005), we fix the parameters $\beta = 1/33$, $\sigma^2 = 1.91$ and $\mu = \log(126) - \sigma^2/2$. We generate a latent Gaussian field $\mathbf{x}$ from the Gaussian process and use these values to generate count data $\mathbf{y}$ from the latent intensity process $\boldsymbol{\Lambda}$. Given the generated data and the fixed hyperparameters, we infer $\mathbf{x}$ by using the MMALA, RMHMC and MALA method as in Christensen *et al.* (2005). The algorithms were run on a single AMD Opteron processor with 8 Gbytes of memory and were coded in MATLAB for consistency.

In many settings the MALA, like HMC sampling, is particularly sensitive to the choice of scaling and very often a reparameterization of the target density is required for these methods to be effective. Indeed this is seen to be so with this particular example, where the MALA cannot sample $\mathbf{x}$ directly. We therefore follow Christensen *et al.* (2005) and employ the transformation $\mathbf{X} = \mu\mathbf{1} + \mathbf{L}\boldsymbol{\Gamma}$, where $\mathbf{L}$ is obtained by Cholesky factorization such that $\{\boldsymbol{\Sigma} + \mathrm{diag}(\mathbf{x})\}^{-1} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. Even after this reparameterization, it is still necessary to tune the scaling factor carefully for this method to work at all. This challenging aspect of employing the MALA has been investigated in detail by Christensen *et al.* (2005) who characterized the problem very well, advising great care in its implementation, but could not ultimately offer any panacea. In contrast with the necessary transformation and fine-tuning that are required by the MALA, both the MMALA and RMHMC sampling allow us to sample the latent variables $\mathbf{x}$ directly *without* reparameterizing the target density.

Fig. 7 shows the traces of the log-joint-probability for both methods by using the starting position $x_{i,j} = \mu$ for $i, j = 1, \ldots, 64$. For the MALA these starting positions must be transformed into corresponding values for $\boldsymbol{\Gamma}$. The RMHMC sampler quickly converges to the true mode after very minimal tuning of the integration step size based on the integration error, which corresponds directly to the acceptance rate. The MMALA also converges very quickly to the true posterior mode. The MALA converges in a similar number of iterations, but only for a

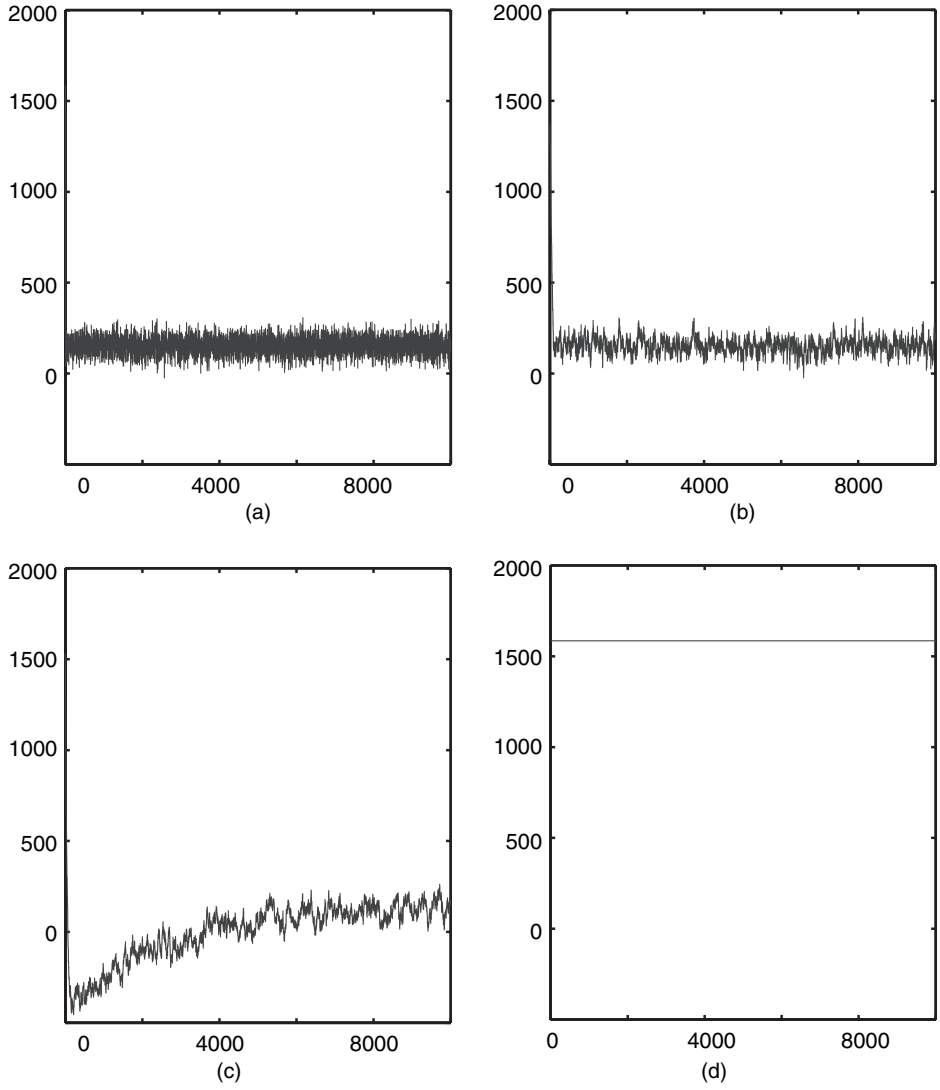

Fig. 7. Trace plots of the log-joint probability for the first 10 000 samples of the latent variables of a log-Gaussian Cox process: (a) convergence of the RMHMC scheme which can directly sample the latent variables **x** without the need for *ad hoc* reparameterizations and pilot runs for fine-tuning; (b) convergence of the MMALA scheme which, since it also uses information about the manifold in the form of the metric tensor, can directly sample without any reparameterizations; (c) log-joint probability for samples drawn by the MALA by using a reparameterization of the latent variables (the scaling was carefully tuned to allow traversal of the parameter space to the posterior mode); (d) trace of the MALA sampler tuned for optimally sampling from the posterior mode (we note that the algorithm cannot now traverse the parameter space when initialized away from this mode; such fine-tuning and reparameterization are frequently necessary when employing the MALA)

suitable choice of scaling factor. Fig. 7(c) shows convergence when the scaling factor is carefully tuned for the transient phase of the Markov chain; however, Fig. 7(d) demonstrates how it fails to converge at all given a scaling factor which is tuned for stationarity. Detailed results of the sampling efficiency of each method are given in Table 10. In this example the RMHMC method required just 1.5 s per effectively independent sample compared with more than 2 h needed by

Table 10. Sampling the latent variables of a log-Gaussian Cox process—comparison of sampling methods

Method	*Time*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
MALA with transformation (transient)	31577	(3, 8, 50)	10605	1
MALA with transformation (stationary)	31118	(4, 16, 80)	7836	1.35
MMALA	634	(26, 84, 174)	24.1	440
RMHMC	2936	(1951, 4545, 5000)	1.5	7070

the MALA. In addition to taking far longer to sample, the MALA also generates much more highly correlated samples and as a result has a far worse effective sample size. This can also be seen in Fig. 8 which shows the inferred posterior latent field, the posterior latent process and the variance that is associated with the Monte Carlo estimate. For RMHMC sampling, the variance in the estimates increases where the data sample is small, i.e. in the top right-hand corner of the field. The MMALA has slightly more variability, whereas the low ESS of the MALA methods manifests itself in patchy regions of high variability across the entire field. We note that the MALA tuned for stationarity has slightly lower variance than the MALA tuned for the transient phase, as we would expect.

Conditionally sampling the hyperparameters by using RMHMC sampling proves more costly, with 5000 posterior samples taking around 90 h of computation time. However, the posterior estimates for the hyperparameters correspond extremely well to their true values; Fig. 9.

Inferring the latent field of a log-Gaussian Cox process with a finely grained discretization is clearly a very challenging problem due to the high dimensionality and strong spatial correlations between the latent variables. The major challenges that we associated with employing the MALA are firstly finding a suitable reparameterization of the target density, and secondly making a suitable choice for the scaling factor according to whether the Markov chain is in a transient or stationary regime. In contrast, the MMALA and RMHMC sampling do not exhibit such extreme technical difficulties. We have demonstrated that RMHMC sampling can sample the latent variables directly with minimal tuning and effort and without the need for reparameterization. By employing a Gibbs style sampling scheme we could additionally sample the hyperparameters of the covariance function in a relatively computationally efficient manner. An investigation into the sparse approaches that were presented in Vanhatalo and Vehtari (2007) and Rue *et al.* (2009) may provide further computational efficiencies. We shall now turn our attention to the very topical application of statistical inference to non-linear differential equations.

10. Manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo sampling for inference in non-linear differential equation models

An important class of problems recently gaining attention is the statistical analysis of uncertainty in dynamical systems defined by a system of non-linear differential equations (Ramsay *et al.*, 2007; Calderhead and Girolami, 2009; Vyshemirsky and Girolami, 2008). For example a dynamical system may be described by a collection of N non-linear ordinary differential equations (ODEs) and model parameters $\boldsymbol{\theta}$ which define a functional relationship between

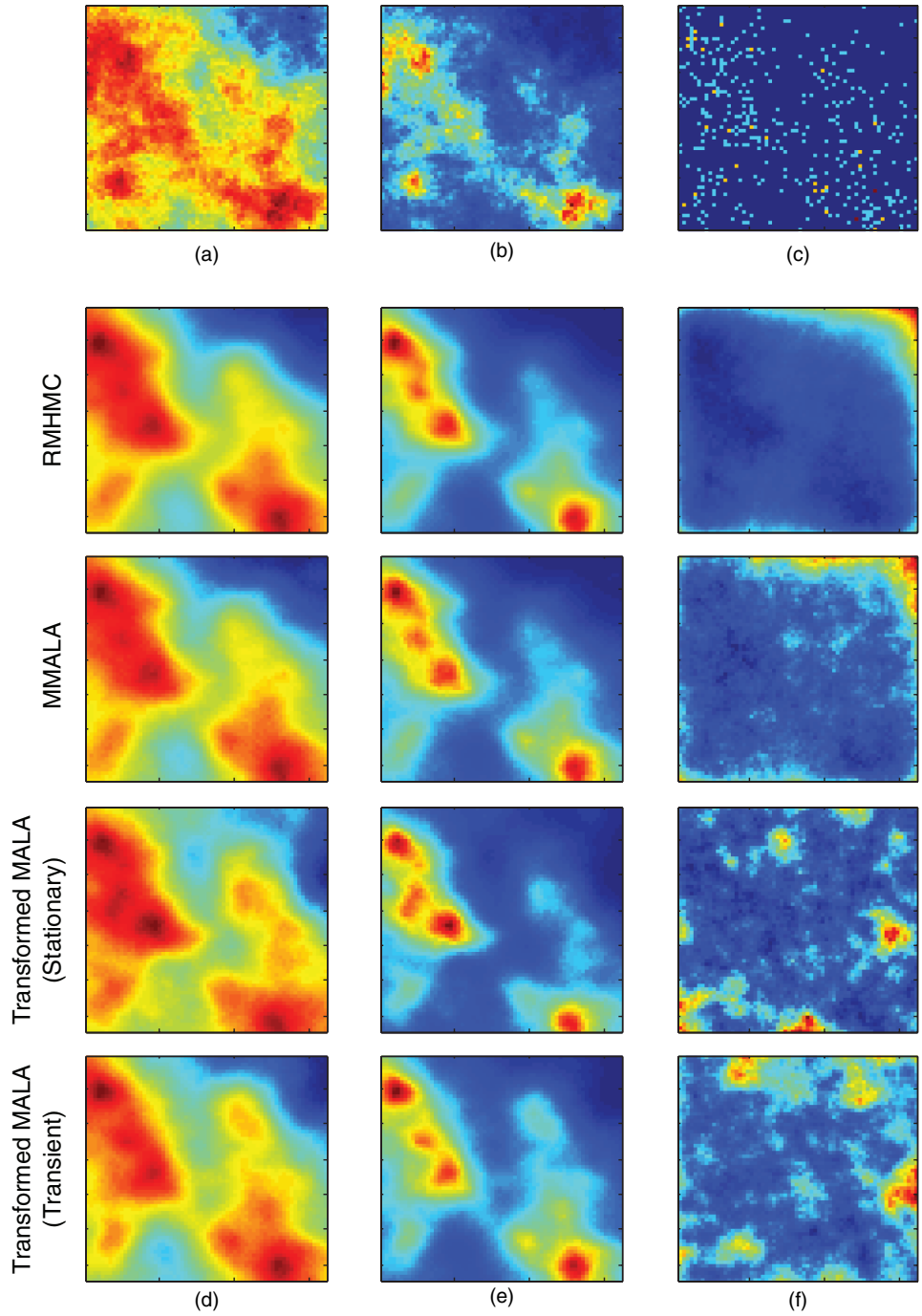

Fig. 8. Posterior latent fields and processes and associated variance, using each of the sampling methods, compared with the true latent field and process (the data employed to infer the latent field are shown in (c); RMHMC sampling produces the lowest variance estimates, which corresponds with its having the highest ESS; for RMHMC sampling there is higher variance where there is less data; however, for the other methods there are patchy areas of high variance due to correlations in the samples collected): (a) true latent field; (b) true latent process; (c) data; (d) posterior latent field; (e) posterior latent process; (f) posterior variance

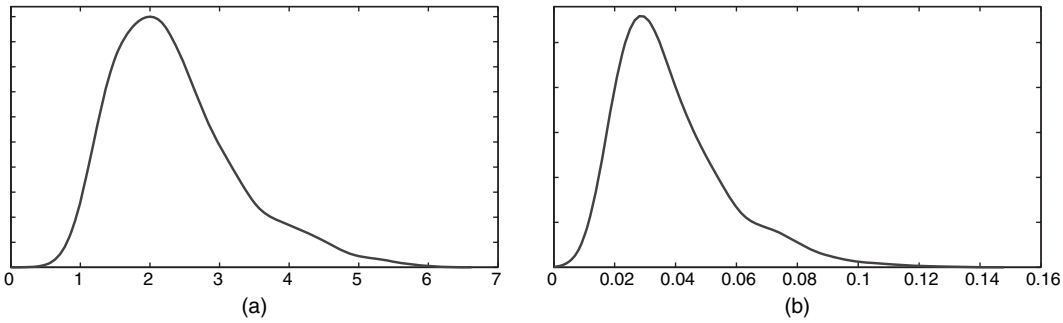

Fig. 9. Kernel density estimates of the hyperparameter samples obtained from Gibbs style sampling from the log-Gaussian Cox model: (a) true value $\sigma = 1.9$; (b) true value $\beta = 0.03$

the process state $\mathbf{x}(t)$ and its time derivative such that $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}, t)$. A sequence of process observations, $\mathbf{y}(t)$, is usually contaminated with some measurement error, which is modelled as $\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\varepsilon}(t)$, where $\boldsymbol{\varepsilon}(t)$ defines an appropriate multivariate noise process, e.g. a zero-mean Gaussian distribution with variance σ_n^2 for each of the N states. If observations are made at T distinct time points, the $T \times N$ matrices summarize the overall observed system as $\mathbf{Y} = \mathbf{X} + \mathbf{E}$. To obtain values for $\mathbf{X}$, the system of ODEs must be solved, so in the case of an initial value problem $\mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)$ denotes the solution of the system of equations at the specified time points for the parameters $\boldsymbol{\theta}$ and initial conditions $\mathbf{x}_0$. The posterior density follows by employing appropriate priors such that

$$p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{x}_0, \boldsymbol{\sigma}) \propto \pi(\boldsymbol{\theta}) \prod_n \mathcal{N}\{\mathbf{Y}_{n,\cdot}|\mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)_{n,\cdot}, \boldsymbol{\Sigma}_n\}.$$

By considering the Gaussian noise model that was described above, where $\boldsymbol{\Sigma}_n = \mathbf{I}_T \sigma_n^2$, using the expected Fisher information matrix we straightforwardly obtain the following analytical expressions for the metric tensor and its derivatives in terms of the first- and second-order sensitivities of the states of the differential equations. The T-dimensional vectors of first-order sensitivities for the nth component of state relative to the ith parameter are denoted by $\mathbf{S}^i_{\cdot,n} = \partial \mathbf{X}_{\cdot,n} / \partial \theta_i$. The metric tensor and its derivatives follow as

$$\mathbf{G}(\boldsymbol{\theta})_{ij} = \sum_{n=1}^{N} \mathbf{S}^{i^{\mathrm{T}}}_{\cdot,n} \boldsymbol{\Sigma}_n^{-1} \mathbf{S}^j_{\cdot,n},$$

$$\frac{\partial \mathbf{G}(\boldsymbol{\theta})_{ij}}{\partial \theta_k} = \sum_{n=1}^{N} \left(\frac{\partial \mathbf{S}^{i^{\mathrm{T}}}_{\cdot,n}}{\partial \theta_k} \boldsymbol{\Sigma}_n^{-1} \mathbf{S}^j_{\cdot,n} + \mathbf{S}^{i^{\mathrm{T}}}_{\cdot,n} \boldsymbol{\Sigma}_n^{-1} \frac{\partial \mathbf{S}^j_{\cdot,n}}{\partial \theta_k} \right).$$

This expression for the metric tensor has an appealing interpretation in that the actual sensitivity equations of the underlying dynamic system model explicitly enter the proposal process for the MCMC scheme. One method for obtaining the required sensitivities at all time points is to approximate them by using finite differences; however, for our purposes this may be inaccurate. For this example we differentiate the system of equations with respect to each of the parameters and directly solve the first-order sensitivity equations defined as

$$\dot{\mathbf{S}}^i_{t,n} = \frac{\partial \mathbf{f}_n(\mathbf{x}, \boldsymbol{\theta}, t)}{\partial \theta_i} = \sum_{l=1}^{N} \frac{\partial \mathbf{f}_{t,n}}{\partial x_l} \mathbf{S}^i_{t,l} + \frac{\partial \mathbf{f}_{t,n}}{\partial \theta_i}.$$

We must take the total derivative with respect to $\boldsymbol{\theta}$, since the states $\mathbf{x}$ also depend on the parameter values. We may augment the original system with these new differential equations, such

that we may solve to obtain both the states and the sensitivities of the states. This will incur an increase in the computational time as it is required to solve both the equations for state and state sensitivity. Similarly we may augment the system with additional equations to solve for the second-order sensitivities, which are required for calculating the partial derivatives of the metric tensor with respect to the model parameters. These equations follow as

$$\frac{\partial \dot{\mathbf{S}}^i_{t,n}}{\partial \theta_k} = \sum_{l=1}^{N} \left\{ \left(\sum_{m=1}^{N} \frac{\partial^2 \mathbf{f}_{t,n}}{\partial x_l \, \partial x_m} \mathbf{S}^k_{t,m} + \frac{\partial^2 \mathbf{f}_{t,n}}{\partial x_l \, \partial \theta_k} \right) \mathbf{S}^i_{t,l} + \frac{\partial \mathbf{f}_{t,n}}{\partial x_l} \frac{\partial \mathbf{S}^i_{t,l}}{\partial \theta_k} \right\} + \sum_{l=1}^{N} \frac{\partial^2 \mathbf{f}_{t,n}}{\partial \theta_i \, \partial x_l} \mathbf{S}^k_{t,l} + \frac{\partial^2 \mathbf{f}_{t,n}}{\partial \theta_i \, \partial \theta_k}.$$

We now have everything that is required to implement RMHMC and MMALA sampling schemes for system models defined by systems of non-linear differential equations.

Interestingly the structure of the equations that are required for the metric tensor and its derivatives are such that RMHMC sampling can be used to form a parallel tempering or population Monte Carlo scheme where the numerical solution of the sensitivity equations and their derivatives can be used at all tempered posterior distributions defined as

$$p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{x}_0, \boldsymbol{\sigma}, \beta) \propto \pi(\boldsymbol{\theta}) \prod_n \mathcal{N}\{\mathbf{Y}_{n,\cdot}|\mathbf{X}(\boldsymbol{\theta}, \mathbf{x}_0)_{n,\cdot}, \boldsymbol{\Sigma}_n\}^{\beta}$$

where $0 \leqslant \beta \leqslant 1$ and the metric is a simple scaling, i.e.

$$\mathbf{G}(\boldsymbol{\theta}, \beta)_{ij} = \beta^2 \sum_{n=1}^{N} \mathbf{S}^{i^{\mathrm{T}}}_{\cdot,n} \boldsymbol{\Sigma}_n^{-1} \mathbf{S}^j_{\cdot,n}.$$

10.1. Experimental results for non-linear differential equations

We present results comparing the sampling efficiency for the parameters of the Fitzhugh–Nagumo differential equations (Ramsay *et al.*, 2007),

$$\begin{aligned} \dot{V} &= c\left(V - \frac{V^3}{3} + R\right), \\ \dot{R} &= -\left(\frac{V - a + bR}{c}\right). \end{aligned} \tag{23}$$

We obtain samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{x}_0, \boldsymbol{\sigma})$, and so in this example $\mathbf{X}_{1,\cdot} = \mathbf{V}$ and $\mathbf{X}_{2,\cdot} = \mathbf{R}$. The sampling schemes that we employ are Metropolis–Hastings, the MALA, HMC sampling, the MMALA, simplified MMALA and RMHMC sampling, as first described in Section 7.1. We again compare the simulations by calculating the effective sample size ESS normalized by the computational time that is required to produce the samples.

Before proceeding we require the first and second partial derivatives of the Fitzhugh–Nagumo equations to calculate the metric tensor for employing manifold sampling approaches to explore the posterior distribution; these are detailed in Appendix C. In practice, all these expressions may be obtained automatically by using symbolic differentiation and we supply MATLAB code for this purpose.

10.1.1. Comparison of sampling schemes

We used 200 data points generated from the Fitzhugh–Nagumo ODE model between $t = 0$ and $t = 20$ with the model parameters $a = 0.2$, $b = 0.2$ and $c = 3$ and initial conditions $V(0) = -1$ and $R(0) = 1$. Gaussian-distributed noise with standard deviation equal to 0.5 was then added to the data; Fig. 10.

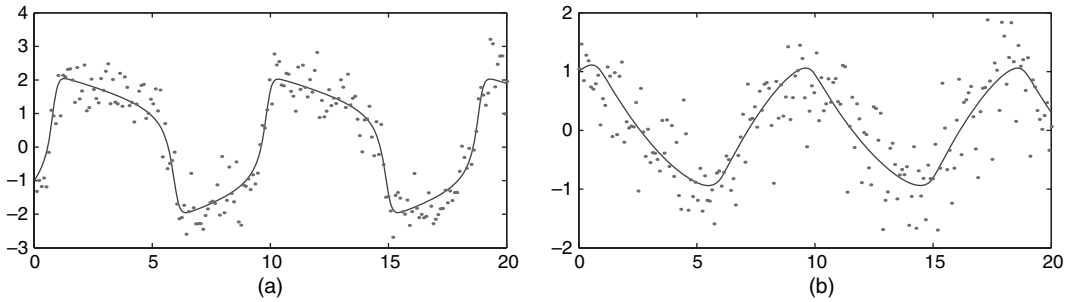

Fig. 10. Output for (a) species V and (b) species R of the Fitzhugh–Nagumo model with parameters $a = 0.2$, $b = 0.2$ and $c = 3$: •, noisy data set

Non-linear ODEs generally induce corresponding non-linearities in the target distribution, which may result in many local maxima (Calderhead and Girolami, 2009). Careful attention must therefore be paid so that the Markov chains do not converge to the wrong mode, but rather sample from the correct distribution. All the sampling methods that are employed in this section may be embedded within a population MCMC framework to allow full exploration of and convergence to the target density (Calderhead *et al.*, 2009); however, for comparing sampling efficiency we employ a single Markov chain initialized on the true mode. We collected 5000 posterior samples and calculated ESS for each parameter, using the minimum value to calculate the time per effectively independent sample. 10 simulations were run for each method, using the same data set, and all methods were implemented in the interpreted language MATLAB for consistency of comparison. All sampling methods were implemented in the same manner as previously described in Section 7.

The results of our simulations are shown in Table 11. Standard HMC sampling takes the longest time for this problem owing to the large number of leapfrog steps that it needs to traverse the parameter space. RMHMC sampling in contrast requires relatively few leapfrog steps, as it takes into account the local geometry to make better moves. We note, however, the additional computational cost of the leapfrog steps, during each of which it is necessary to solve the system of ODEs to evaluate the gradients and metric tensor. The first momentum update of RMHMC sampling is relatively quick since only a vector–matrix multiplication is necessary; however, updating the parameter values requires the metric tensor to be evaluated for each fixed point iteration in the generalized leapfrog algorithm as the parameter values converge, thus adding a considerable amount of computation to the overall algorithm. The MMALA methods offer

Table 11. Fitzhugh–Nagumo model: summary of results for 10 runs of the model parameter sampling scheme with 5000 posterior samples

Sampling method	*Time (s)*	*Mean ESS (a, b, c)*	*Total time/minimum mean ESS*	*Relative speed*
Metropolis	18.5	132, 130, 108	0.17	3.9
MALA	14.4	125, 21, 46	0.67	1
HMC	815	4668, 3483, 3811	0.23	2.9
MMALA	34.9	1057, 925, 956	0.037	18.1
Simplified MMALA	14.9	1007, 479, 762	0.031	21.6
RMHMC	266	4302, 4202, 3199	0.083	8

the best performance for this particular example, as they have the benefit of using manifold information to guide the direction of the chain, but without the required fixed point iterations, thus only requiring the ODEs to be numerically solved once per iteration. This suggests that the MMALA is perhaps particularly suited to settings in which there is a non-constant metric tensor which is expensive to compute, as in this case.

The Fitzhugh–Nagumo model has only three parameters and we see that the MALA and HMC method perform adequately in this low dimensional setting; indeed the largest marginal parameter variance is only four times larger than the smallest marginal variance. We would expect the MALA and HMC sampling to perform worse in cases where there is a greater difference in the marginal variances, since the step size of each is restricted by the smallest marginal variance. Similarly, although componentwise Metropolis sampling performs adequately in this setting, we would expect its performance to deteriorate in higher dimensions where there are greater correlations in the parameters.

11. Conclusions and discussion

In this paper Riemann manifold Metropolis adjusted Langevin and HMC sampling methods have been proposed and evaluated, on a representative range of inference problems. The development of these methods is an attempt to improve on existing MCMC methodology when sampling from target densities that may be of high dimension and exhibit strong correlations. It is argued that the methods are fully automated in terms of tuning the overall proposal mechanism to accommodate target densities which may exhibit strong correlations, widely varying scales in each dimension and significant changes in the geometry of the manifold between the transitional and stationary phases of the Markov chain. By exploiting the natural Riemann structure of the parameter space of statistical models the methods proposed can be viewed as generalizations of both HMC and MALA methods and as such have the potential to overcome the oftentimes complex manual tuning that is required of both methods.

Clearly there are two main overheads when employing the MMALA or RMHMC sampling, the first being the ability to develop analytical expressions and stable numerical or finite sample estimates for the metric tensor (once it has been chosen) along with its associated derivatives. The second is the worst case $\mathcal{O}(N^3)$ scaling of solving the linear systems when updating the parameter vectors, i.e. inverting the metric tensor, especially for high dimensional problems. The issue of the $\mathcal{O}(N^3)$ scaling is something which deserves further consideration. In some statistical models there is a natural sparsity in the metric tensor; the stochastic volatility model example is a case in point where owing to this structure RMHMC sampling was computationally more efficient than the MMALA and HMC sampling. In other models this is not so, e.g. the logistic regression model and the log-Gaussian Cox model. It should be noted that adaptive MCMC methods (see for example Andrieu and Thoms (2008)) also incur the same level of cubic scaling. At the very high dimensional end of the scale a decorrelating transformation is required for the MALA and HMC sampling and this will also incur an $\mathcal{O}(N^3)$ scaling; however, further work to characterize the incurred computational costs at the intermediate dimensionality regime will be of value. The use of *guiding Hamiltonians*, as described in Duane *et al.* (1987), may be a way of reducing the computational cost of proposals in RMHMC sampling; however, at the moment it is unclear how this could make any dramatic reduction in this respect. As far as the computational issues are concerned automatic or adjoint differentiation methods may prove to be of use, and Hanson (2002) has proposed adjoint methods for HMC sampling. There are clearly many numerical and computational avenues of investigation that may be followed in this regard.

Interestingly the simplified MMALA method can be seen to employ a drift term which is based on the natural gradient as defined in Amari and Nagaoka (2000) and this form of natural gradient, which is the contravariant gradient as defined in differential geometry, has been exploited in approximate Bayesian inference by Honkela *et al.* (2008). On page 319 of Robert (2004) the MALA is derived from a second-order approximation of the target density, which can also be seen to be a simplified MMALA where the metric is the negative of the Hessian matrix; a similar approach was taken in Qi and Minka (2002) when seeking to exploit the Hessian in the design of MCMC methods. The geometric perspective that is adopted in this paper provides the overarching framework that generalizes these specific approaches.

In this paper all the examples that have been considered have had analytic expressions for the expected Fisher information matrix. However, there are whole families of statistical models for which the Fisher information matrix is not available in closed analytic form, mixture models being an obvious example. In these cases it may be possible to employ the empirical Fisher information matrix (Spall, 2005) in the form of an estimate of the covariance of the score, which has the advantage that the overall methods require only second-order derivatives. The other option is to employ the observed Fisher information, although numerical issues such as the loss of guaranteed positive definiteness would require consideration. It is unclear what type of manifold structure this would induce, so the theoretical and practical implications of the difference between the expected, empirical and observed information matrices would be worthy of further investigation.

This leads onto the discussion about the particular choice of metric to be employed if one takes the view that the Fisher information is only one possible metric that could be adopted. Alternatives have already been considered in the literature, e.g. the preferred point metric (Critchley *et al.*, 1993); although not within the context of MCMC sampling and this presents a new area of analysis and study to characterize the principles of optimality in appropriate metric design for MCMC sampling.

A note of caution regarding the exploitation of the geometry that is induced by the Fisher information metric in inference problems is spelled out in Skilling (2006). Two distributions may be a short distance apart on the probability simplex; however, if the parameter submanifold (which we are interested in) is locally *rough* they may be distantly separated and hence following small scale detailed paths on the submanifold will be highly inefficient. This is not an observation that is made in this paper; however, there may be examples where this will be a real problem; for example inference over dynamic systems that exhibit complex limit cycles is challenging owing to the small scale structure that is induced in the probability density (Calderhead *et al.*, 2009). Further theoretical and applied investigation will help to understand this issue more fully.

The work of Christensen *et al.* (2005), Roberts and Rosenthal (1998) and Roberts and Stramer (2003) has provided theoretical analysis of limiting rates of convergence, ergodicity, optimal step sizes and acceptance rates for the MALA, and more recently HMC methods (Beskos *et al.*, 2010). This type of theoretical study will be required for the MMALA and RMHMC class of MCMC methods to characterize their theoretical properties in a rigorous manner. The highly promising performance that was reported in the experimental evaluation of the MMALA and RMHMC methods on challenging inference problems gives further motivation for this theoretical analysis.

From the experimental evaluation the raw ESS-values for RMHMC sampling far exceeds that of the MMALA despite both methods being based on geometric principles. There are several reasons for this; firstly the MMALA proposal is based on a single forward step of the Euler integrator whereas the proposal mechanism for RMHMC sampling can take multiple integration

steps, thus travelling further on the manifold (parameter space) for each proposal. Secondly the discrete version of the Langevin diffusion is being driven by a diffusion term that is defined by the metric tensor at the current point rather than the new point. Depending on the step size this will introduce further inefficiency based on deviation from the manifold of the effective path. Thirdly, as has already been commented on, Hamiltonian flows of the form that are employed in RMHMC sampling are locally geodesic flows (Calin and Chang 2004; McCord *et al.*, 2002), suggesting a possible optimality, in terms of distance, in the paths that are simulated across the manifold by HMC and RMHMC sampling. This is an interesting point which requires further theoretical analysis to characterize the nature of these local geodesics and how they may be exploited further in this regard.

In summary the MMALA and RMHMC methods provide novel MCMC algorithms whose performance has been assessed on a range of statistical models and in all cases has been shown to be superior to similar MCMC methods. The adoption of this geometric viewpoint when designing MCMC algorithms provides a framework in which to develop further the theory, methodology and application of this promising avenue of statistical inference.

Acknowledgements

M. Girolami is supported by Engineering and Physical Sciences Research Council Advanced Research Fellowship EP/E052029/1, Engineering and Physical Sciences Research Council project grant EP/F009429/1 and Biotechnology and Biological Sciences Research Council project grant BB/G006997/1. B. Calderhead is supported by a Microsoft Research European doctoral scholarship. M. Girolami is indebted to Sui Chin, Nial Friel, Andrew Gelman, Dirk Husmeier, Tom Minka, Iain Murray, Radford Neal, Gareth Roberts, John Skilling, Andrew Stuart, Aki Vehtari and the reviewers, for valuable comment, helpful suggestions and constructive criticism regarding the ideas that are developed in this paper.

Appendix A: Expressions required for stochastic volatility model

We employ the transformations $\sigma = \exp(\gamma)$ and $\phi = \tanh(\alpha)$ to deal with constrained parameters. The derivatives of the transformations follow as $\mathrm{d}\sigma/\mathrm{d}\gamma = \exp(\gamma) = \sigma$ and $\mathrm{d}\phi/\mathrm{d}\alpha = 1 - \tanh^2(\alpha) = 1 - \phi^2$. The partial derivatives of joint-log-probability $L = \log\{p(\mathbf{y}, \mathbf{x}|\beta, \sigma, \phi)\}$ with respect to the transformed parameters are

$$\frac{\partial L}{\partial \beta} = -\frac{T}{\beta} + \sum_{t=1}^{T} \frac{y_t^2}{\beta^3 \exp(x_t)}, \tag{24}$$

$$\frac{\partial L}{\partial \gamma} = \frac{\partial L}{\partial \sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}\gamma} = -T + \frac{x_1^2(1-\phi^2)}{\sigma^2} + \sum_{t=2}^{T} \frac{(x_t - \phi x_{t-1})^2}{\sigma^2}, \tag{25}$$

$$\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial \phi}\frac{\mathrm{d}\phi}{\mathrm{d}\alpha} = -\phi + \frac{\phi x_1^2(1-\phi^2)}{\sigma^2} + \sum_{t=2}^{T} \frac{x_{t-1}(x_t - \phi x_{t-1})(1-\phi^2)}{\sigma^2}. \tag{26}$$

If we want to sample the parameters by using the MMALA or RMHMC sampling, then we also need expressions for the metric tensor and its partial derivatives with respect to β, σ and ϕ. We can obtain the following expressions for the individual components of the metric tensor for the log-probability-density:

$$E\left(\frac{\partial^2 L}{\partial \beta^2}\right) = -\frac{2T}{\beta^2}, \qquad E\left(\frac{\partial^2 L}{\partial \gamma^2}\right) = -2T, \qquad E\left(\frac{\partial^2 L}{\partial \beta\, \partial \gamma}\right) = E\left(\frac{\partial^2 L}{\partial \beta\, \partial \alpha}\right) = 0, \tag{27}$$

$$E\left(\frac{\partial^2 L}{\partial \gamma \partial \alpha}\right) = -2\phi, \qquad E\left(\frac{\partial^2 L}{\partial \alpha^2}\right) = -2\phi^2 - (T-1)(1-\phi^2). \tag{28}$$

Thus the expected Fisher information matrix and its partial derivatives follow as

$$\mathbf{G}(\alpha,\gamma,\beta)=\begin{pmatrix}2T/\beta^2 & 0 & 0\\ 0 & 2T & 2\phi\\ 0 & 2\phi & 2\phi^2+(T-1)(1-\phi^2)\end{pmatrix},$$

$$\frac{\partial\mathbf{G}}{\partial\beta}=\begin{pmatrix}-4T/\beta^3 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix},$$

$$\frac{\partial\mathbf{G}}{\partial\gamma}=\begin{pmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix},$$

$$\frac{\partial\mathbf{G}}{\partial\alpha}=\begin{pmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 2\phi(3-T)(1-\phi^2)\end{pmatrix}.$$

We therefore require expressions for the second-order derivatives of the log-priors to obtain the overall metric tensor, and also the third-order derivatives of the log-priors to calculate the partial derivatives of the metric tensor, which follow straightforwardly.

Appendix B: Expressions required for log-Gaussian Cox process model

We employ a change of variables $\sigma^2=\exp(\varphi_1)$ and $\beta=\exp(\varphi_2)$ to allow constrained sampling such that σ^2 and β are both strictly positive. The log-probability and gradients that are required for sampling the hyperparameters of the Gaussian process follow in standard form, where $i=1,2$, as

$$\frac{\partial\mathcal{L}}{\partial\varphi_i}=-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_i}\right)+\frac{1}{2}(\mathbf{x}-\mu\mathbf{1})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_i}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu\mathbf{1}) \tag{29}$$

and the Fisher information matrix also follows in standard form as

$$\mathbf{G}(\boldsymbol{\varphi})_{ij}=\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_j}\right). \tag{30}$$

Application of standard derivatives of trace operators provides an analytical expression for the derivative of the metric tensor with respect to the transformed parameters:

$$\begin{aligned}\frac{\partial\mathbf{G}(\boldsymbol{\varphi})_{ij}}{\partial\varphi_k}&=\frac{\partial}{\partial\varphi_k}\left\{\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_j}\right)\right\}\\ &=-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_k}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_j}\right)+\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\varphi_i\,\partial\varphi_k}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_j}\right)\\ &\quad-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_k}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_j}\right)+\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\varphi_i}\boldsymbol{\Sigma}^{-1}\frac{\partial^2\boldsymbol{\Sigma}}{\partial\varphi_j\,\partial\varphi_k}\right).\end{aligned}$$

In our experiments we employ an infinitely differentiable stationary covariance function to calculate the (i,j)th entry of the covariance matrix,

$$\boldsymbol{\Sigma}_{(i,j),(i',j')}=\sigma^2\exp\left\{-\frac{1}{64\beta}\delta(i,i',j,j')\right\}, \tag{31}$$

where $\delta(i,i',j,j')=\sqrt{\{(i-i')^2+(j-j')^2\}}$. The gradients and the Fisher information matrix above may therefore be obtained by using the first and second partial derivatives of the covariance function. The first partial derivatives follow as

$$\frac{\partial\boldsymbol{\Sigma}_{i,j}}{\partial\varphi_1}=\sigma^2\exp\left\{-\frac{1}{64\beta}\delta(i,i',j,j')\right\},$$

$$\frac{\partial\boldsymbol{\Sigma}_{i,j}}{\partial\varphi_2}=\frac{\sigma^2}{64\beta}\exp\left\{-\frac{1}{64\beta}\delta(i,i',j,j')\right\}\delta(i,i',j,j').$$

The second partial derivatives may also be easily calculated as follows:

$$\frac{\partial^2 \boldsymbol{\Sigma}_{i,j}}{\partial \varphi_1^2} = \sigma^2 \exp\left\{-\frac{1}{64\beta}\delta(i,i',j,j')\right\},$$
$$\frac{\partial^2 \boldsymbol{\Sigma}_{i,j}}{\partial \varphi_1 \partial \varphi_2} = \frac{\sigma^2}{64\beta} \exp\left\{-\frac{1}{64\beta}\delta(i,i',j,j')\right\}\delta(i,i',j,j'),$$
$$\frac{\partial^2 \boldsymbol{\Sigma}_{i,j}}{\partial \varphi_2^2} = \frac{\sigma^2}{(64\beta)^2} \exp\left\{-\frac{1}{64\beta}\delta(i,i',j,j')\right\}\delta(i,i',j,j')^2 - \frac{\sigma^2}{64\beta} \exp\left\{-\frac{1}{64\beta}\delta(i,i',j,j')\right\}\delta(i,i',j,j').$$

Once again we require expressions for the second-order derivatives of the log-priors to obtain the metric tensor over the full target distribution, and also the third-order derivatives of the log-priors to calculate the partial derivatives of the metric tensor. These follow straightforwardly from the Ga(2,0.5) priors that were employed over the hyperparameters σ^2 and β.

Appendix C: Partial derivatives for ordinary differential equation example

$$\frac{\partial \dot{V}}{\partial a} = \frac{\partial \dot{V}}{\partial b} = 0,$$
$$\frac{\partial \dot{V}}{\partial c} = V - \frac{V^3}{3} + R,$$
$$\frac{\partial \dot{R}}{\partial a} = \frac{1}{c},$$
$$\frac{\partial \dot{R}}{\partial b} = \frac{-R}{c},$$
$$\frac{\partial \dot{R}}{\partial c} = \frac{V - a + bR}{c^2}.$$

All the second derivatives of $\dot{V}$ with respect to the model parameters are equal to 0, and the five non-zero second partial derivatives of $\dot{R}$ are

$$\frac{\partial^2 \dot{R}}{\partial a \partial c} = -\frac{1}{c^2},$$
$$\frac{\partial^2 \dot{R}}{\partial b \partial c} = \frac{R}{c^2},$$
$$\frac{\partial^2 \dot{R}}{\partial c \partial a} = -\frac{1}{c^2},$$
$$\frac{\partial^2 \dot{R}}{\partial c \partial b} = \frac{R}{c^2},$$
$$\frac{\partial^2 \dot{R}}{\partial c^2} = 2\left(\frac{-V + a - bR}{c^3}\right).$$

In addition, the second partial derivatives with respect to all states and parameters are required for writing the differential equation describing the second-order sensitivities. There are again five non-zero second partial derivatives with respect to the states and parameters:

$$\frac{\partial^2 \dot{V}}{\partial V \partial c} = 1 - V^2,$$
$$\frac{\partial^2 \dot{V}}{\partial R \partial c} = 1,$$

$$\frac{\partial^2 \dot{R}}{\partial V \, \partial c} = \frac{1}{c^2},$$
$$\frac{\partial^2 \dot{R}}{\partial R \, \partial b} = -\frac{1}{c},$$
$$\frac{\partial^2 \dot{R}}{\partial R \, \partial c} = \frac{b}{c^2}.$$

References

Amari, S. and Nagaoka, H. (2000) *Methods of Information Geometry*. Oxford: Oxford University Press.

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc.* B, **72**, 269–342.

Andrieu, C. and Thoms, J. (2008) A tutorial on adaptive MCMC. *Statist. Comput.*, **18**, 343–373.

Barndorff-Nielsen, O. E., Cox, D. R. and Reid, N. (1986) The role of differential geometry in statistical theory. *Int. Statist. Rev.*, **54**, 83–96.

Beichl, I. and Sullivan, F. (2000) The Metropolis Algorithm. *Comput. Sci. Engng*, **2**, 65–69.

Beskos, A., Pillai, N., Roberts, G., Serna, S. and Stuart, A. (2010) Optimal tuning of the Hybrid Monte-Carlo algorithm. *Technical Report*. Department of Statistical Science, University of College London, London.

Calderhead, B. and Girolami, M. (2009) Estimating Bayes factors via thermodynamic integration and population MCMC. *Computnl Statist. Data Anal.*, **53**, 4028–4045.

Calderhead, B., Girolami, M. and Lawrence, N. D. (2009) Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *Adv. Neur. Inform. Process.*, **21**, 217–224.

Calin, O. and Chang, D. C. (2004) *Geometric Mechanics on Riemannian Manifolds*. Basel: Birkhäuser.

Christensen, O. F., Roberts, G. O. and Rosenthal, J. S. (2005) Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *J. R. Statist. Soc.* B, **67**, 253–268.

Chung, K. L. (1982) *Lectures from Markov Processes to Brownian Motion*. New York: Springer.

Critchley, F., Marriot, P. K. and Salmon, M. (1993) Preferred point geometry and statistical manifolds. *Ann. Statist.*, **21**, 1197–1224.

Dawid, A. P. (1975) Discussion on 'Defining the curvature of a statistical problem' (with applications to second-order efficiency (by B. Efron). *Ann. Statist.*, **3**, 1231–1234.

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987) Hybrid Monte Carlo. *Phys. Lett.* B, **55**, 2774–2777.

Efron, B. (1975) Defining the curvature of a statistical problem (with applications to second-order efficiency). *Ann. Statist.*, **3**, 1189–1242.

Efron, B. and Hinkley, D. V. (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, **65**, 457–487.

Ferreira, P. E. (1981) Extending Fisher's measure of information. *Biometrika*, **68**, 695–698.

Gamerman, D. (1997) Sampling from the posterior distribution in generalized linear mixed models. *Statist. Comput.*, **7**, 57–68.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*. New York: Chapman and Hall.

Geyer, C. J. (1992) Practical Markov Chain Monte Carlo. *Statist. Sci.*, **7**, 473–483.

Gustafson, P. (1997) Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics*, **53**, 230–242.

Hairer, E., Lubich, C. and Wanner, G. (2006) *Geometric Numerical Integration, Structure Preserving Algorithms for Ordinary Differential Equations*, 2nd edn. Berlin: Springer.

Hajian, A. (2007) Efficient cosmological parameter estimation with Hamiltonian Monte Carlo technique. *Phys. Rev.* D, **75**, 083525–1–11.

Hanson, K. M. (2001) Markov Chain Monte Carlo posterior sampling with the Hamiltonian method. *Proc. SPIE*, **4322**, 456–467.

Hanson, K. M. (2002) Use of probability gradients in hybrid MCMC and a new convergence test. *Report LA-UR-02-4105*. Los Alamos National Laboratory, Los Alamos.

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

Holmes, C. C. and Held, L. (2005) Bayesian auxiliary variable models for binary and multinomial regression. *Baysn Anal.*, **1**, 145–168.

Honkela, A., Tornio, M., Raiko, T. and Karhunen, J. (2008) Natural conjugate gradient in variational inference. *Lect. Notes Comput. Sci.*, **4985**, 305–314.

Husmeier, D., Penny, W. and Roberts, S. J. (1999) An empirical evaluation of Bayesian sampling with hybrid Monte Carlo for training neural network classifiers. *Neur. Netwrks*, **12**, 677–705.

Ishwaran, H. (1999) Applications of hybrid Monte Carlo to Bayesian generalised linear models: quasicomplete separation and neural networks. *J. Computnl Graph. Statist.*, **8**, 779–799.
Jeffreys, H. (1948) *Theory of Probability*, 2nd edn. Oxford: Clarendon.
Johnson, V. E., Krantz, S. G. and Albert, J. H. (1999) *Ordinal Data Modeling*. New York: Springer.
Kass, R. E. (1989) The geometry of asymptotic inference. *Statist. Sci.*, **4**, 188–234.
Kent, J. (1978) Time reversible diffusions. *Adv. Appl. Probab.*, **10**, 819–835.
Kim, S., Shephard, N. and Chib, S. (1998) Stochastic volatility: likelihood inference and comparison with ARCH models. *Rev. Econ. Stud.*, **65**, 361–393.
Lambert, P. and Eilers, P. H. C. (2009) Bayesian density estimation from grouped continuous data. *Computnl Statist. Data Anal.*, **53**, 1388–1399.
Lauritzen, S. L. (1987) Statistical manifolds. In *Differential Geometry in Statistical Inference*, pp. 165–216. Hayward: Institute of Mathematical Statistics.
Leimkuhler, B. and Reich, S. (2004) *Simulating Hamiltonian Dynamics*. Cambridge: Cambridge University Press.
Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing*. New York: Springer.
McCord, C., Meyer, K. R. and Offin, D. (2002) Are Hamiltonian flows geodesic flows? *Trans. Am. Math. Soc.*, **355**, 1237–1250.
Metropolis, M., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (1994) *Machine Learning, Neural and Statistical Classification*. Englewood Cliffs: Prentice Hall.
Murray, M. K. and Rice, J. W. (1993) *Differential Geometry and Statistics*. New York: Chapman and Hall.
Neal, R. M. (1993a) Probabilistic inference using Markov Chain Monte Carlo methods. *Technical Report*. University of Toronto, Toronto.
Neal, R. M. (1993b) Bayesian learning via stochastic dynamics. *Adv. Neur. Inform. Process. Syst.*, **5**, 475–482.
Neal, R. M. (1996) Bayesian learning for neural networks. *Lect. Notes Statist*.
Neal, R. M. (2010) MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (eds S. Brooks, A. Gelman, G. Jones and X.-L. Meng). Boca Raton: CRC Press.
Qi, Y. and Minka, T. (2002) Hessian-based Markov Chain Monte-Carlo algorithms. *1st Cape Cod Wrkshp Monte Carlo Methods*. (Available from `http://www.cs.purdue.edu/homes/alanqi/papers/qi-minka-HMH-AMIT-02.pdf`.)
Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007) Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Statist. Soc.* B, **69**, 741–796.
Rao, C. R. (1945) Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calc. Math. Soc.*, **37**, 81–91.
Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
Robert, C. (2004) *Monte Carlo Statistical Methods*. New York: Springer.
Roberts, G. O. and Rosenthal, J. S. (1998) Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc.* B, **60**, 255–268.
Roberts, G. and Stramer, O. (2003) Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. App. Probab.*, **4**, 337–358.
Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *J. R. Statist. Soc.* B, **71**, 319–392.
Skilling, J. (2006) Probability and geometry. In *ESA-EUSC: Image Information Mining for Security and Intelligence*. (Available from `http://earth.eo.esa.int/rtd/Events/ESA-EUSC_2006/Oral/Ar19_Skilling.pdf`.)
Spall, J. C. (2005) Monte Carlo computation of the Fisher information matrix in nonstandard settings. *J. Computnl. Graph. Statist.*, **14**, 889–909.
Tsutakawa, R. K. (1972) Design of experiment for bioassay. *J. Am. Statist. Ass.*, **67**, 584–590.
Vanhatalo, J. and Vehtari, A. (2007) Sparse log Gaussian processes via MCMC for spatial epidemiology. In *JMLR Wrkshp. Conf. Proc. Gaussian Processes in Practice*, vol. 1, pp. 73–89.
Vyshemirsky, V. and Girolami, M. (2008) Bayesian ranking of biochemical system models. *Bioinformatics*, **24**, 833–839.
Zlochin, M. and Baram, Y. (2001) Manifold stochastic dynamics for Bayesian learning. *Neur. Computn*, **13**, 2549–2572.