Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2011)
73, *Part* 2, *pp.* 123–214

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 www.ucl.ac.uk/statistics/research/rmhmc allows replication of all the results reported.

Keywords: Bayesian inference; Geometry in statistics; Hamiltonian Monte Carlo methods; Langevin diffusion; Markov chain Monte Carlo methods; Riemann manifolds

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 and Casella (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

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

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

© 2011 Royal Statistical Society 1369–7412/11/73123

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 and Casella (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 and Casella (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 and Casella 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

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

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. 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 from `www.ucl.ac.uk/statistics/research/rmhmc`. 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 that converges to the target measure. However, dependent on the characteristics of the curvature this 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

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

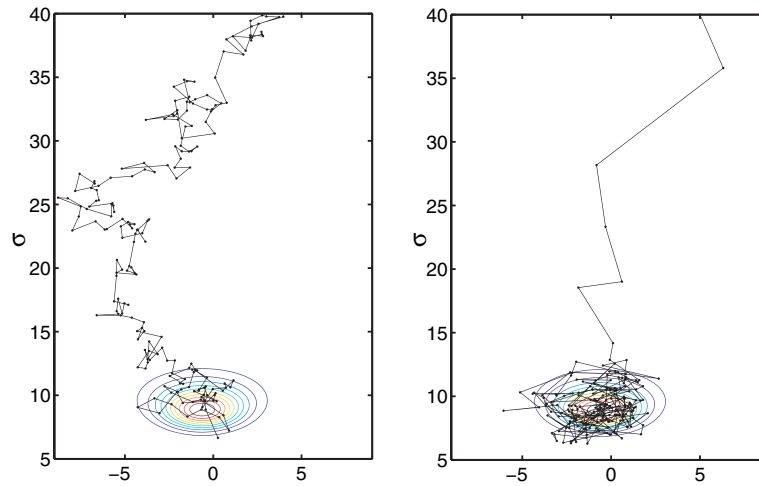

Fig. 1. Contours representing the sample estimate of $p(\mu, \sigma|X)$ where a sample of size $N = 30$ was drawn from $\mathcal{N}(X|\mu = 0, \sigma = 10)$ (both MALA and MMALA discrete diffusions were forward simulated from initial points $\mu_0 = 5$ and $\sigma_0 = 40$ with a step size $\varepsilon = 0.75$ 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)

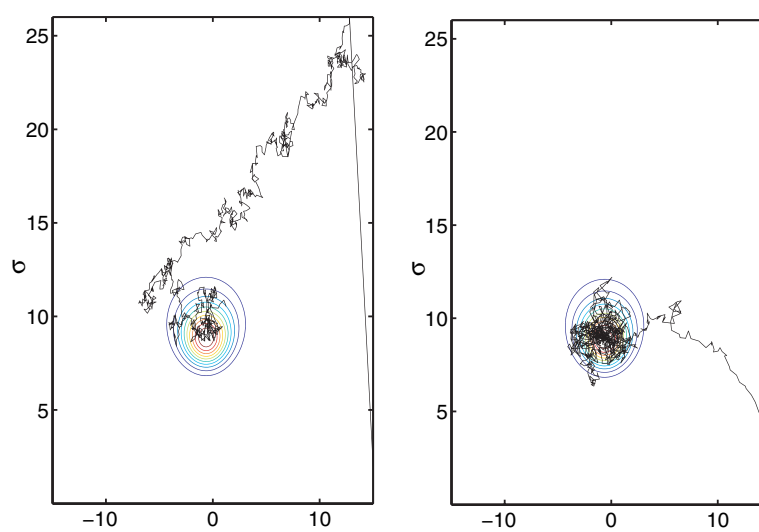

Fig. 2. Same data sample as in Fig. 1, but with $\mu_0 = 15$ and $\sigma_0 = 2$ (the step size is reduced to $\varepsilon = 0.2$ so that the MALA converges and 1000 proposal steps are taken): as in Fig. 1, from (a) it is clear that the Euclidean metric of the MALA does not exploit the hyperbolic geometry 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.* (2006), pages 187–190, and Leimkuhler and Reich (2004), pages 81–87, for a detailed derivation and proofs.

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

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

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

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.* (2006), 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.

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 $\mathbf{s} = (s_1, \ldots, s_T)^{\mathrm{T}}$ where each $s_i = \{y_i^2 \beta^{-2} \exp(-x_i) - 1\}/2$, $\delta_1 = \sigma^{-2}(x_1 - \phi x_2)$, $\delta_T = \sigma^{-2}(x_T - \phi x_{T-1})$, the $(T-2)$-dimensional vector $\mathbf{w}$ with elements $\sigma^{-2}(x_t - \phi x_{t-1}) - \phi\sigma^{-2}(x_{t+1} - \phi x_t)$, where $t = 2, \ldots, T-1$, and $\mathbf{r} = (\delta_1, \mathbf{w}^{\mathrm{T}}, \delta_T)$, then 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 note 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 the 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, 20000 posterior samples were collected after a burn-in period of 10000 samples. This sampling procedure was repeated 10 times. The efficiency was compared in terms of

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 (s)*	*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)	5.24	27.3
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 (s)*	*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

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 (Fig. 6), 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 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

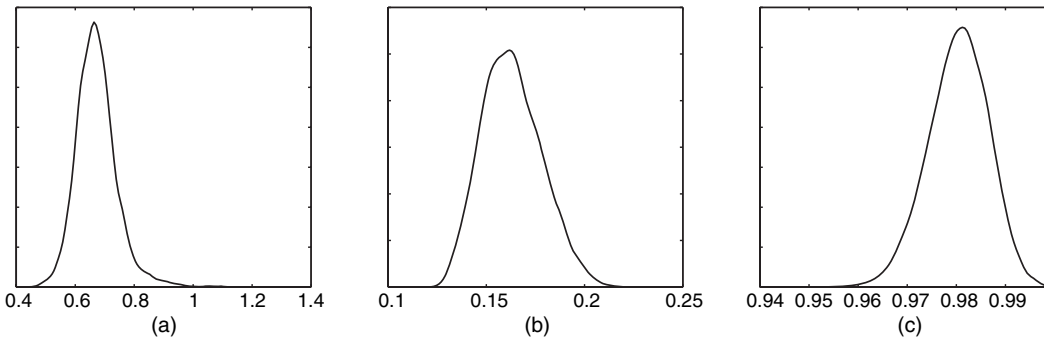

Fig. 6. Posterior marginal densities for (a) β, (b) σ and (c) ϕ, employing RMHMC sampling to draw 20 000 samples of the parameters and latent volatilities by using a simulated data set consisting of 2000 observations: the true values are $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$

parts 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 $\delta(i, i', j, j') = \surd\{(i - i')^2 + (j - j')^2\}$. 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 RMHMC 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

given a more explicit recommendation on how to address this problem. Instead, the setting of simulation lengths is given a fairly sketchy treatment, although suitable evaluation metrics are being used in the fairly extensive experiments.

My final point is the related question of how this methodology can be made available and put into use by the large community of MC practitioners who could benefit. HMC methodology is a little complicated to explain, although the resulting algorithm is surprisingly simple. Nevertheless its use has been restricted to small pockets of the community. The Riemann manifold HMC method is considerably more complex to understand and implement, although it should be easier to run effectively once implemented. Good recommendations on how to run RMHMC algorithms, not requiring intricate understanding of every aspect of the algorithm, may be necessary for the algorithm to realize its potential.

The vote of thanks was passed by acclamation.

Arnaud Doucet (*University of British Columbia, Vancouver*), **Pierre Jacob** (*Université Paris Dauphine and Centre de Recherche en Economie et Statistique, Paris*) **and Adam M. Johansen** (*University of Warwick, Coventry*)

We congratulate the authors for their elegant contribution.

Consider those situations in which we do not have direct access to an appropriate metric but can obtain pointwise, simulation-based estimates of its values. For example, we might be interested in performing Bayesian inference in general state space hidden Markov models by using particle Markov chain Monte Carlo methods (Andrieu *et al.*, 2010). In this context, we integrate out numerically the latent variables of the model by using a sequential Monte Carlo (SMC) scheme. A sensible metric to use is the observed information matrix which can also be estimated in this way (Poyiadjis *et al.*, 2010). We discuss here the use of such estimates in a Markov chain Monte Carlo context.

Assume that we want to sample from a target $\pi(x)$ on $\mathcal{X}$ using the Metropolis–Hastings algorithm. Denote the proposal's parameters (e.g. scale) $r \in \mathcal{R}$. Defining an extended target over $\mathcal{X} \times \mathcal{R}$ as $\bar{\pi}(x, r) = \pi(x)\, q(r|x)$ an algorithm may be defined on $\mathcal{X} \times \mathcal{R}$ in which both R and X are sampled.

At iteration $n+1$ draw $X^* \sim s(\cdot|x_n, r_n)$ and $R^* \sim q(\cdot|x^*)$. Accept this proposal with the standard Metropolis–Hastings acceptance probability on the extended space

$$\begin{aligned}\alpha(x_n, r_n; x^*, r^*) &= \mathbf{1} \wedge \frac{\bar{\pi}(x^*, r^*)}{\bar{\pi}(x_n, r_n)} \frac{s(x_n|x^*, r^*)\, q(r_n|x_n)}{s(x^*|x_n, r_n)\, q(r^*|x^*)} \\ &= \mathbf{1} \wedge \frac{\pi(x^*)}{\pi(x_n)} \frac{s(x_n|x^*, r^*)}{s(x^*|x_n, r_n)}.\end{aligned}$$

Hence it is not necessary to be able to evaluate q, even pointwise, provided that it can be sampled from. The resulting transition is reversible on the extended space and admits π as a marginal of its invariant distribution. This simple result is well known: see Besag *et al.* (1995), appendix 1.

The manifold Metropolis adjusted Langevin algorithm, with metric tensor obtained by sampling, may be justified by using precisely the same argument: a proposal of the form of equation (10) may be implemented with a sampled estimate of the metric tensor and such gradients as are required (objects which can be obtained readily in settings of interest, such as hidden Markov models); the extended space construction above holds with $x = 0$ and $r = (G, \nabla G)$ and the acceptance probability remains of the same form; the constant curvature proposal may be implemented without the need for estimates of ∇G with $x = 0$ and $r = G$.

The Hamiltonian Monte Carlo variant of the same is not trivial. As each step of the implicit integrator requires access to the value of the metric at several (implicitly defined) points, direct application of the above principles does not appear possible. However, more subtle approaches can be employed. In particular one could consider trying to approximate the metric by using the expectation of a function with respect to a probability measure independent of x and using common random variates from this measure during a Hamiltonian Monte Carlo update.

Antti Honkela (*Institute for Information Technology, Helsinki*)

The application of Riemannian geometry of probability distributions to Markov chain Monte Carlo methods proposed in the paper appears a very promising new tool for developing efficient highly automated computational techniques for Bayesian inference. The authors have done an especially beautiful job in successfully combining efficient Hamiltonian numerical solvers with the theoretical framework.

One aspect limiting the practical applicability of the method is the difficulty of deriving the necessary geometric structure of different models as well as potentially the computational complexity of working in that geometry. The authors already suggest a few alternatives to the expected Fisher information matrix. The evaluation of which situations these are optimal both in theory and in practical computational implementations is an important topic for further study. Fortunately, because of the Metropolis adjustment, the metric employed can be changed without sacrificing the correctness of the samplers.

In my own work I have applied Riemannian geometric ideas to speeding up variational inference methods. These are based on fitting a parametric approximation to the posterior by optimizing some objective. Posing this optimization problem on the Riemannian manifold that is defined by the approximations and applying a Riemannian conjugate gradient algorithm can yield orders of magnitude speed-up over Euclidean gradient algorithms (Honkela *et al.*, 2010), even when using some simplifications to the algorithm to avoid the most complex Riemannian vector operations.

The potential connection between these methods suggests interesting opportunities for Riemann manifold samplers. In our method we use the expected Fisher information geometry of the approximations. By selecting a suitable approximation, this can be chosen to be tractable and efficient to evaluate. In Honkela *et al.* (2010) we present a Riemannian geometric method for Gaussian mixtures, which are intractable in the standard expected Fisher information framework. Perhaps these approximation geometries could be used to work around this, to extend further the applicability of the beautiful methods that are presented here.

Serge Guillas (*University College London*)

The authors should be congratulated for this important paper. Dramatic improvements in efficiency of Markov chain Monte Carlo techniques are expected from the algorithms proposed. One interesting area of application of such schemes is the Bayesian calibration of computer models (Kennedy and O'Hagan, 2001). Full Bayesian analysis relies on Gaussian processes and is computationally challenging. The number of parameters can be high, and the outputs can be high dimensional, as in complex models for the environment. For instance, to calibrate the National Center for Atmospheric Research thermosphere–ionosphere electrodynamics general circulation model, Guillas *et al.* (2009) used multiple chains in parallel to reduce wall clock time and to check Markov chain Monte Carlo convergence. The burn-in period was around 200–500 iterations. The combination of samples from 10 chains, after the burn-in period, supplied draws from the posterior. A reduction in the length of this burn-in period, by an order of magnitude as the authors show in some of their examples—where the geometry of the problem can be exploited—will allow further parallelization on large clusters, saving more time and enabling more advanced studies. For this, we need to make use of the Fisher information matrix of a Gaussian process, with the possible addition of hyperparameters. The derivation of the Fisher information matrix in this context has been investigated for covariance structures of Gaussian, triangular and exponential types (Abt and Welch, 1998), and more recently of Matérn type (Loh, 2005).

When the Fisher information matrix has no analytical form, the authors rightly suggest the use of sampling methods (Spall, 2005). The issue is that it is extremely time consuming to include these additional steps in the Markov chain Monte Carlo algorithm. However, more recently, Das *et al.* (2010) have developed an algorithm that makes use of known parts of the Fisher information matrix to improve efficiency. One hopes that a class of problems may become tractable as a result of the combination of such efficient techniques with the manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo approaches.

Simon Barthelmé (*Technische Universität Berlin*) **and Nicolas Chopin** (*Centre de Recherche en Economie et Statistique, Paris*)

One of the many things that we like about this paper is that it forces us to change our perspective on Metropolis–Hastings sampling. We may not be the only ones with the toy example of a bivariate, strongly correlated, Gaussian distribution imprinted in our brain. This example explains well why taking correlations into account is important. However, one often forgets that, contrary to the Gaussian example, the curvature of the log-target-density may be far from constant, which justifies a *local* calibration of hidden Markov strategies. The authors give compelling evidence that local calibration may lead to strong improvements in large dimensional problems.

There are two ways to understand these results. One of them, which was put forward in this paper, stems from the information geometry perspective: the parameter space is endowed with a metric defined by $G(\theta)$, which turns the posterior distribution into a density over a manifold. The general manifold Metropolis adjusted Langevin algorithm (MMALA) based on a diffusion over that manifold is a beautiful

mathematical device, but it is not immediately apparent how this leads to improved (relative) Markov chain Monte Carlo performance. A different viewpoint proceeds from optimization: the MMALA performs better because it uses a better local model of the posterior density.

As often pointed out, the Langevin proposal is a noisy version of a gradient ascent step. Similarly, the simplified MMALA step is a noisy version of a (quasi-)Newton step, in which the Hessian is replaced with the Fisher information matrix, which is an idea known as iteratively reweighted least squares in the literature on generalized linear models. It is worth emphasizing the fact that the simplified versions, which just rely on these local curvature ideas, but do not require third derivatives, do better in terms of relative efficiency (not to mention in terms of human computation time!).

This suggests two avenues for further research. First, many optimization methods have been developed that require only evaluating the gradient. This may be more convenient from the practitioner's point of view, and it also proves more effective whenever computing Hessian matrices is expensive. Methods, such as the Broyden–Fletcher–Goldfarb–Shanno or Barzilai–Borwein method, approximate the Hessian locally from the previous k iterations. Our preliminary experiments indicate that these methods may reduce the correlation in Markov chain Monte Carlo chains.

The second point is that the auxiliary Gaussian distribution is merely a choice that is imposed by the physical interpretation of the Hamiltonian. Do the authors have any intuition on what would be the optimal auxiliary distribution?

Maurizio Filippone (*University of Glasgow*)

Consider non-parametric logistic regression with Gaussian process priors (Rasmussen and Williams, 2006), where a set of n covariates $\mathbf{x}_i \in \mathbb{R}^d$ are associated with response $y_i \in \{0, 1\}$:

$$p(\mathbf{f}|\mathbf{0}) \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, K),$$
$$p(y_i|f_i) = \sigma(f_i)^{y_i}\{1-\sigma(f_i)\}^{1-y_i}.$$

Let K be the covariance matrix parameterized by a vector of (hyper)parameters $\boldsymbol{\theta} = (\psi_\sigma, \psi_{r_1}, \ldots, \psi_{r_d})$:

$$k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) = \exp(\psi_\sigma)\exp\{-\tfrac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} A(\mathbf{x}_i - \mathbf{x}_j)\},$$
$$A^{-1} = \mathrm{diag}\{\exp(\psi_{r_1}), \ldots, \exp(\psi_{r_d})\}.$$

We consider the manifold methods that are presented in this paper in comparison with a set of alternative algorithms to sample from the joint posterior distribution of $\mathbf{f}$ and $\boldsymbol{\theta}$.

Efficiently sampling of $\mathbf{f}$ and $\boldsymbol{\theta}$ is complex because of their strong coupling (Murray and Adams, 2010; Neal, 1999). Gibbs style samplers, as used by the authors in Section 9, based on sampling of $\mathbf{f}|\boldsymbol{\theta}, \mathbf{y}$ and $\boldsymbol{\theta}|\mathbf{f}, \mathbf{y}$ are convenient from an implementation standpoint, but extremely inefficient. This is because fixing $\mathbf{f}$ induces a sharply peaked posterior for $\boldsymbol{\theta}$, resulting in a poor effective sample size (ESS) for the length scale parameters (Murray and Adams, 2010).

The metric tensor comprises the Fisher information matrix and the negative of the Hessian of the prior:

$$G_{\mathbf{f}} = -E_{\mathbf{y}|\mathbf{f}}[\nabla_{\mathbf{f}}\nabla_{\mathbf{f}}\mathcal{L}] = \sigma(\mathbf{f})\{1-\sigma(\mathbf{f})\} + K^{-1} = \Lambda + K^{-1} \qquad \Lambda = \mathrm{diag}[\sigma(\mathbf{f})\{1-\sigma(\mathbf{f})\}],$$
$$G_{\mathbf{f},\theta_i} = -E_{\mathbf{y},\mathbf{f}|\boldsymbol{\theta}}\left[\frac{\partial \nabla_{\mathbf{f}}\mathcal{L}}{\partial \theta_i}\right] = -E_{\mathbf{f}|\boldsymbol{\theta}}\left[K^{-1}\frac{\partial K}{\partial \theta_i}K^{-1}\mathbf{f}\right] = 0,$$
$$G_{\theta_j,\theta_i} = -E_{\mathbf{y},\mathbf{f}|\boldsymbol{\theta}}\left[\frac{\partial^2 \mathcal{L}}{\partial \theta_i\,\partial \theta_j}\right] = \frac{1}{2}\,\mathrm{tr}\left(K^{-1}\frac{\partial K}{\partial \theta_i}K^{-1}\frac{\partial K}{\partial \theta_j}\right) - \frac{\partial^2 \log\{p(\boldsymbol{\theta})\}}{\partial \theta_i\,\partial \theta_j}$$
$$G = \begin{pmatrix} G_{\mathbf{f}} & \mathbf{0} \\ \mathbf{0} & G_{\boldsymbol{\theta}} \end{pmatrix}.$$

The derivatives of G follow from standard properties of matrix derivatives. Taking the expectations with respect to $\mathbf{y}$ alone does not lead to a positive definite matrix G and it is therefore necessary to take them with respect to $\mathbf{y}$ and $\mathbf{f}$ jointly (for $G_{\mathbf{f}}$ we compute the expectation with respect to $\mathbf{y}$ to leave the dependence from $\mathbf{f}$). G is block diagonal, so the geometry-based argument in favour of the decoupling of $\mathbf{f}$ and $\boldsymbol{\theta}$, when sampled jointly by using manifold methods, does not hold.

Results and experimental settings for a bivariate logistic regression problem with $n = 100$ are reported in Table 12. The results confirm that Gibbs style samplers are very inefficient in sampling the length scale parameters.

Table 12. ESS for Gibbs Metropolis–Hastings (Gibbs MH), Gibbs simplified manifold Metropolis adjusted Langevin algorithm (Gibbs SMMALA), Gibbs Riemann manifold Hamiltonian Monte Carlo (Gibbs RMHMC) Gibbs whitening (Gibbs Wht), HMC, SMMALA and RMHMC algorithms all averaged over 10 runs (the standard deviation is in parentheses)†

Parameter	*Gibbs MH*	*Gibbs SMMALA*	*Gibbs RMHMC*	*Gibbs Wht*	*HMC*	*SMMALA*	*RMHMC*
Minimum ESS $\mathbf{f}$	3 (0)	27 (17)	78 (69)	26 (24)	3 (0)	17 (6)	182 (50)
Average ESS $\mathbf{f}$	6 (0)	102 (27)	404 (92)	112 (24)	6 (0)	51 (4)	531 (80)
Maximum ESS $\mathbf{f}$	18 (3)	205 (35)	888 (60)	300 (55)	21 (4)	94 (12)	100 (61)
ESS ψ_σ	30 (11)	54 (38)	30 (13)	56 (20)	5 (2)	18 (10)	530 (250)
ESS ψ_{τ_1}	30 (13)	6 (2)	6 (4)	203 (112)	12 (11)	6 (2)	86 (33)
ESS ψ_{τ_2}	36 (23)	8 (3)	7 (3)	136 (60)	10 (6)	7 (4)	111 (40)
$10^3 \times G$	—	—	—	—	—	3 (0)	257 (18)
$10^3 \times G_{\boldsymbol{\theta}}$	—	3 (0)	80 (6)	—	—	—	—
$10^3 \times \mathrm{chol}(K)$	3 (0)	3 (0)	80 (6)	3 (0)	47 (1)	3 (0)	257 (18)

†In Hamiltonian-based methods, the maximum number of leapfrog steps was set to 30. The Gibbs MH and HMC algorithms were tuned on the basis of posterior covariances estimated from pilot runs of the Gibbs Wht algorithm. We also report the number of calls (in thousands) to the functions computing G, $G_{\boldsymbol{\theta}}$ and the Cholesky decomposition of K that are the main computational bottlenecks (along with the derivatives of $G_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$, although we are not reporting these statistics). All the methods were initialized from the true values used to generate the data; the ESS is computed over 2000 samples collected after 1000 burn-in samples. In Gibbs style samplers, the length scale parameters have a poor ESS, whereas the latent functions are sampled quite efficiently by manifold methods, confirming that the geometric argument is effective in improving the sampling of $\mathbf{f}$.

Gibbs sampling with Riemann manifold Hamiltonian Monte Carlo proposals seems suboptimal in this problem, as it may be for the log-Gaussian Cox model presented by the authors in Section 9. A natural decoupling of $\mathbf{f}$ and $\boldsymbol{\theta}$ is offered by whitening the prior over $\mathbf{f}$. Given the decomposition $K = LL^{\mathrm{T}}$, define $\boldsymbol{\nu} = L^{-1}\mathbf{f}$; sampling $\boldsymbol{\theta}|\mathbf{f}, \mathbf{y}$ is replaced by $\boldsymbol{\theta}|\boldsymbol{\nu}, \mathbf{y}$.

Even if G is block diagonal, the results for computationally demanding runs of Riemann manifold Hamiltonian Monte Carlo algorithms show some potential in achieving an ESS that is comparable with the whitening method. This motivates further investigation on less expensive (guiding) Hamiltonians for the joint update of $\mathbf{f}$ and $\boldsymbol{\theta}$ trading off some efficiency. Also, it would be particularly interesting to start off from the whitened model and to study whether manifold methods can improve sampling efficiency.

Frank Critchley (*The Open University, Milton Keynes*)

It is a truth universally acknowledged that, when a man has been waiting a *very* long time for a bus, two will come at once. I am such a man. I first encountered differential geometries for statistics in Barndorff-Nielsen *et al.* (1986) and have never doubted their *potential* major influence on mainstream applied statistics. However, *realizing* that potential has been a very long wait indeed (almost a quarter of a century!), undoubtedly important results being locked away from everyday use behind conceptual and notational barriers.

To my knowledge, the first bus arrived in the work of Copas and Eguchi (see, for example, Copas and Eguchi (2010)). In particular, their geometrically inspired 'double-the-variance' formula gives an eminently practical way to allow for model uncertainty, at least asymptotically. Close behind, this paper and its accompanying software represent a second bus, whose arrival I wholeheartedly welcome.

Parenthetically, I would like to signal the (hopefully, not-too-distant) arrival of a third: computational information geometry. Being a global approximate analogue of the first, local asymptotic bus, this treatment of the 'uncertainty of uncertainty' is joint work with Anaya-Izquierdo, Marriott and Vos, some of whom will be contributing in writing after the meeting.

Like all good 'read' papers, this one makes important contributions—among which, I find the geometric Hamiltonian dynamic approach particularly appealing—whose very originality opens up new questions of potential further interest. These include the following.

(a) To what extent are the procedures presented equivariant to reparameterization? If not fully, can they be adapted to be so? If not, can an argument be made for a particular choice of parameterization?

(b) Is there a potential role for alternative choices of metric? Possibilities here include one of the *preferred point* metrics of Critchley *et al.* (1993)—perhaps, based on (estimates of) the target distribution, up to proportionality—the preferred point expected score vector also carrying important information in this case.
(c) Again, might non-Riemannian geodesics be of value, notions of 'straightness' based on exponential or mixture connections being natural candidates?

Overall, as will be clear, I warmly welcome this paper, both in itself and as an exemplar of geometric ideas informing applied statistics, and hope that many will want to step on board the bus which it represents.

N. Friel and J. Wyse (*University College Dublin*)
Our discussion explores an extension of the Riemann manifold Hamiltonian Monte Carlo (RMHMC) methodology to Bayesian model selection, where we entertain a collection of plausible models $M_1, \ldots, M_m$ each with parameters $\theta_1, \ldots, \theta_m$ and data y. We consider the augmented target distribution

$$\exp\{-H(\theta_i, p_i, M_i)\} = \pi(\theta_i, M_i|y)\,\pi(p_i|\theta_i)$$

where $\pi(\theta_i, M_i|y)$ is the posterior distribution over model parameters and model index. Further the auxiliary $p_i \sim N\{0, G(\theta_i)\}$, where $G(\theta_i)$ is the metric tensor defined at parameter θ_i. We consider a reversible jump (RJ) Markov chain Monte Carlo (MCMC) extension of the RMHMC algorithm, which we term RJRMHMC. The essential part of our algorithm is the stochastic mechanism whereby we propose a move from (θ_l, p_l, M_l) to (θ_k, p_k, M_k). The scheme is described in Fig. 12. Briefly, this scheme involves leapfrog steps from the current state to an intermediate state within model M_l. A jump move is then proposed to a state in model M_k followed by leapfrog steps in model M_k. Negating the leapfrog integration steps yields a reversible move from the current state to the proposed state. The acceptance probability appears as

$$\alpha\{(\theta_l, p_l), (\theta_l^*, p_l^*), (\theta_k^*, p_k^*), (\theta_k, p_k)\} = \min\left[1, \frac{\exp\{-H(\theta_k, p_k)\}}{\exp\{-H(\theta_l, p_l)\}} \frac{\phi\{p_l^*; 0, G(\theta_l^*)\}\Pr(k \to l)}{q(u)\,\phi\{p_k^*; 0, G(\theta_k^*)\}\Pr(l \to k)}\right],$$

where ϕ is a multivariate normal density. This depends not just on the current and proposed states, but also on the intermediate states (θ_l^*, p_l^*) and (θ_k^*, p_k^*). The leapfrog steps typically result in movement of the chain towards a high density region of the posterior distribution—effectively yielding an adaptive proposal mechanism for the jump move in the RJ algorithm.

We have applied our methodology to the *Pima Indians* logistic regression data set presented in the paper. Our interest is to carry out a Bayesian variable selection of the seven covariates in the data set. We compared an RJMCMC algorithm with the RJRMHMC algorithm where, for both methods, moves to models differing by one variable were proposed. We ran both algorithms for the same central processor unit time (500000 RJMCMC iterations and 5 million RJRMHMC iterations). The posterior model probability estimates were similar and are displayed in Fig. 13. The acceptance rates within each algorithm were quite different. Between- and within-model moves for the RJRMHMC algorithm had acceptance rates of 4% and 96% respectively, whereas, for the RJMCMC algorithm, between- and within-model acceptance rates were 2.5% and 11% respectively. This suggests that the RJRMHMC algorithm provides an improvement

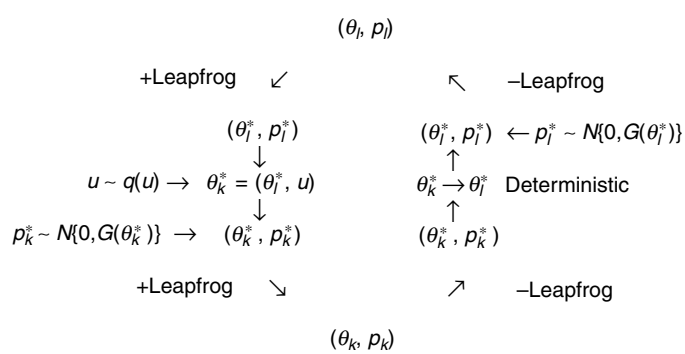

Fig. 12. Schematic diagram of the RJRMHMC 'jump' move from current state (θ_l, p_l) in model M_l to a proposed state (θ_k, p_k) in model M_k

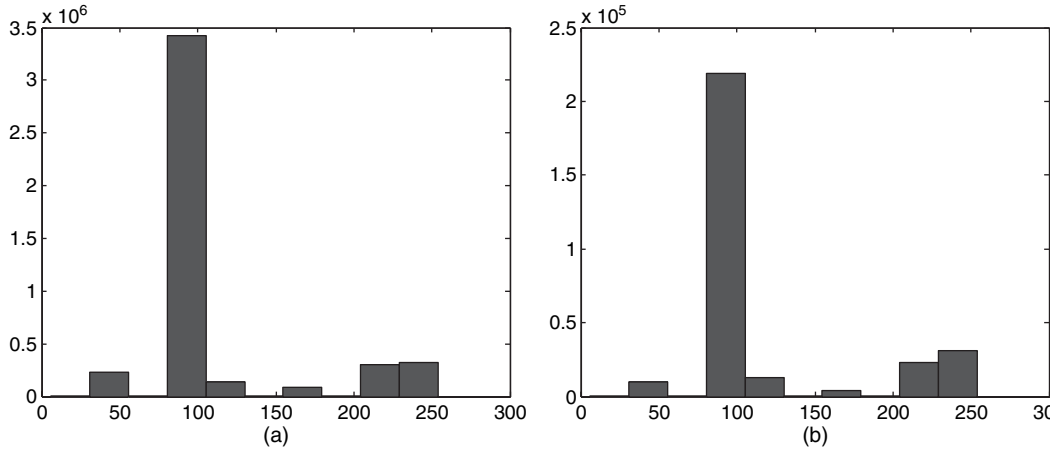

Fig. 13. Posterior model probabilities from (a) the RJRMHMC and (b) the RJMCMC algorithms

in efficiency. The implementation of the RJRMHMC algorithm could be improved and extended in many directions and this will form the basis for future research.

Vassilios Stathopoulos and Maurizio Filippone (*University of Glasgow*)
We consider a univariate binomial probit model where we use X, W and $Y \in \{0, 1\}$ to denote the observed covariates, the latent variables and binomial responses respectively. The latent variables W are modelled as

$$W = X\beta + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The binomial variable is $Y(W) = 0$ if $W < 0$ and $Y(W) = 1$ if $W > 0$. The model in this form is known to be non-identifiable as the likelihood is constant along straight lines out of the origin of the (β, σ)-plane and therefore is only informative about their ratio (Nobile, 1998, 2000; McCulloch *et al.*, 2000; Imai and van Dyk, 2005). This poses significant challenges to the Markov chain Monte Carlo methods discussed in this paper since the Fisher information matrix is not positive definite. The problem can be resolved by considering an informative prior (Nobile, 1998) and by adding the negative of its Hessian to the Fisher information matrix as suggested by the authors in Section 4.2. The resulting posterior, however, is strongly skewed and, as we discuss here, this can lead to very poor mixing of the chains.

For the experiments presented here, we generated a synthetic data set for the binomial model as described in Nobile (1998) and used the priors $p(\beta) = \mathcal{N}(0, 100)$ and $p(1/\sigma^2) = \mathcal{G}(\frac{3}{2}, \frac{1}{6})$ which ensure weak identifiability. Furthermore, we reparameterize σ, such that $\psi = \log(\sigma^2)$, and sample ψ. The log-likelihood is given by

$$\mathcal{L} = \sum_i y_i \log\left\{\Phi\left(\frac{\beta x_i}{\sigma}\right)\right\} + \sum_i (1 - y_i) \log\left\{\Phi\left(-\frac{\beta x_i}{\sigma}\right)\right\}$$

where Φ is the cumulative function of $\mathcal{N}(0, 1)$. The gradient of the log-likelihood and the Fisher information matrix follow as

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_i a_i y_i + b_i,$$
$$\frac{\partial \mathcal{L}}{\partial \psi} = \sum_i c_i y_i + d_i,$$
$$\mathcal{I}_i = \begin{pmatrix} b_i^2 & b_i d_i \\ b_i d_i & d_i^2 \end{pmatrix} + \Phi\left(\frac{\beta x_i}{\sigma}\right) \begin{pmatrix} a_i^2 + 2a_i b_i & a_i c_i + a_i d_i + b_i c_i \\ a_i c_i + a_i d_i + b_i c_i & c_i^2 + 2c_i d_i \end{pmatrix}$$

with

$$a_i = \frac{x_i}{\sigma}\{\xi(\rho) + \xi(-\rho)\},$$
$$b_i = -\frac{x_i}{\sigma}\xi(-\rho),$$

Consider observations $y_1, \ldots, y_n \sim \mathcal{N}(\theta_1 + \theta_2^2, \sigma_y^2)$. The parameters θ_1 and θ_2 are non-identifiable without any additional information beyond the observations: any values such that $\theta_1 + \theta_2^2 = c$ for some constant c explain the data equally well. By imposing a prior distribution $\theta_1 + \theta_2 \sim \mathcal{N}(0, \sigma_\theta^2)$ we create weak identifiability, namely decreased posterior probability for c far from zero. Fig. 16 shows the prior, likelihood and ridge-like posterior for the model. For this problem, we have

$$\mathbf{G}(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{n}{\sigma_y^2} + \dfrac{1}{\sigma_\theta^2} & \dfrac{2n\theta_2}{\sigma_y^2} \\ \dfrac{2n\theta_2}{\sigma_y^2} & \dfrac{4n\theta_2^2}{\sigma_y^2} + \dfrac{1}{\sigma_\theta^2} \end{pmatrix}.$$

Fig. 17 compares typical trajectories of both HMC and RMHMC algorithms, demonstrating the ability of RMHMC sampling to follow the full length of the ridge.

HMC and RMHMC algorithms also differ in sensitivity to the step size. As described by Neal (2010), HMC algorithms suffer from the presence of a critical step size above which the error explodes, accumulating at each leapfrog step. In contrast, RMHMC algorithms occasionally exhibit a sudden jump in the Hamiltonian at one specific leapfrog step, followed by well behaving steps (as seen in Fig. 16(a)). This is due to the possible divergence of the fixed point iterations in the generalized leapfrog equations

$$\mathbf{p}\left(\tau + \frac{\varepsilon}{2}\right) = \mathbf{p}(\tau) - \frac{\varepsilon}{2} \nabla_{\boldsymbol{\theta}} H \left\{ \boldsymbol{\theta}(\tau), \mathbf{p}\left(\tau + \frac{\varepsilon}{2}\right) \right\} \tag{35}$$

for given momentum $\mathbf{p}(\tau)$, parameter $\boldsymbol{\theta}(\tau)$ and step size ε. Fig. 18 shows the probability that equation (35) has a solution $\mathbf{p}(\varepsilon/2)$ as a function of $\boldsymbol{\theta}(0)$, and of the derivative at the fixed point being 'sufficiently small' for the fixed point iterations to converge; the well-known sufficient theoretical threshold on the derivative (see for example Fletcher (1987)) is 1, but we conservatively chose 1.2 on the basis of typical successful runs. When the finite number of fixed point iterations diverges, the Hamiltonian explodes; however, subsequent steps may still admit a fixed point and hence behave normally. Unsurprisingly, this behaviour is much more likely to occur for larger step sizes.

Although the regions of low probability can strongly decrease the mixing of the algorithm, they do not affect the theoretical convergence ensured by the rejection step. Understanding this behaviour can bring

(a) (b)

Fig. 17. Three typical consecutive trajectories of 20 leapfrog steps each, with step size 0.1, for the RMHMC and HMC algorithm, chosen to highlight two acceptances (black) and one rejection (grey), representative of the approximately 65% acceptance ratio for both the HMC and RMHMC algorithms: we see that the RMHMC algorithm can track the contours of the density and reach the furthest tails of the ridge, adapting to the local geometry, whereas the spherical moves of the HMC algorithm oscillate back and forth across the ridge

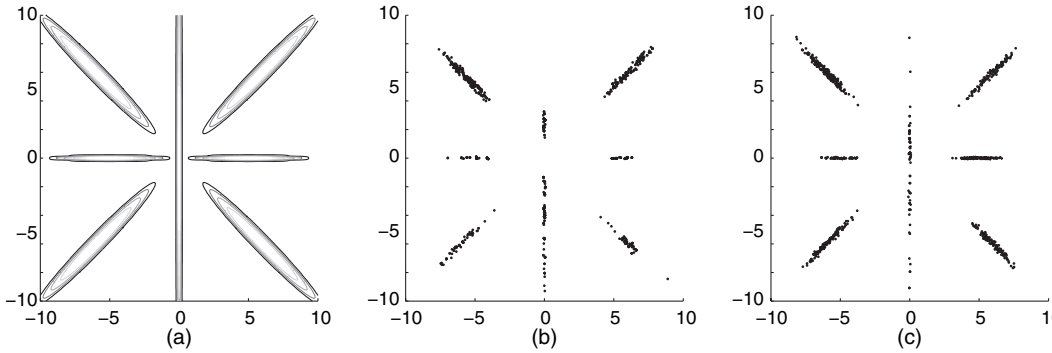

Fig. 26. (a) Contour plot of the target density and 1000 posterior samples from a mixture of seven two-dimensional Gaussian distributions with varying means and covariances by using (b) random-walk Metropolis sampling and (c) the simplified MMALA

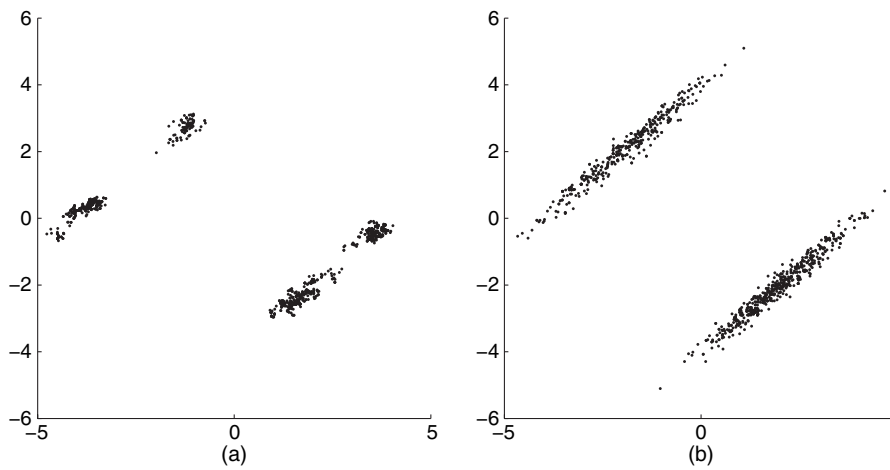

Fig. 27. 1000 posterior samples from the first two dimensions of a mixture of two strongly correlated 20-dimensional Gaussian distributions by using (a) random-walk Metropolis sampling and (b) the simplified MMALA

random-walk Metropolis scheme and the simplified MMALA; however, there is better coverage of each mode when employing the manifold method.

A more challenging example is a mixture of two strongly correlated and unevenly scaled 20-dimensional Gaussian distributions. This time we employ a schedule with 20 temperatures, again evenly distributed between 0 and 1, and then raised to the power 5. Fig. 27 shows the samples collected after the burn-in period for the first two of the 20 dimensions. It is clear that the manifold-based method does a much better job of covering the posterior modes. Fig. 28 shows the trace plots of these samples.

Multiple metric tensors

Following this theme Jasra and Singh suggest the use of a sequence of tempered densities, and hence tempered metrics, in sequential Monte Carlo sampling—this may have a benefit

(a) in terms of reducing correlations and
(b) in that some of the metric tensors may be cheaper to work with and so the sequential Monte Carlo sampler can use the computationally inexpensive metrics.

This is similar to the generalized Hamiltonian that is suggested by Beffy and Robert, employing multiple metric tensors, possibly tempered, or capturing different small-scale and large-scale structures. Another perspective, motivated by balancing computational load with sampling efficiency, is provided by Campbell

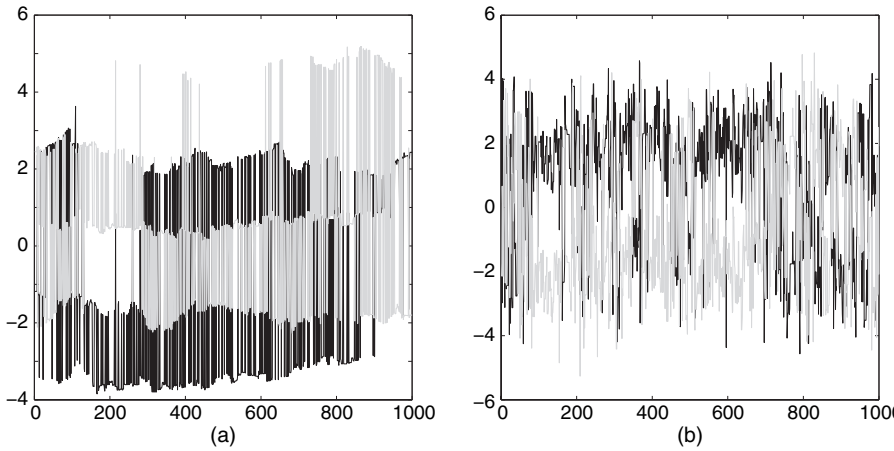

Fig. 28. Trace plots of 1000 posterior samples from the first two dimensions of a mixture of two strongly correlated 20-dimensional Gaussian distributions by using (a) random-walk Metropolis sampling and (b) the simplified MMALA

who suggests careful block updating, with between-block independence assumed, e.g. Chib and Ramamurthy (2010). These are without doubt valuable areas for further methodological developments.

Hierarchic Bayesian and latent variable models

The class of latent Gaussian models (e.g. the log-Gaussian Cox model in the paper) is given detailed consideration by Filippone, Vehtari and Vanhatalos. Filippone highlights the huge computational load that is required of a full RMHMC scheme, as used in the paper, for such a model and notes the decoupled metric for latent functions and model parameters. It is clear that further effort is needed in defining an appropriate metric(s) for such hierarchic models, as noted by Murray and Adams. We consider it is possible that a different geometric view for hierarchic Bayesian models and models with latent variables may be advantageous with possible relationships to the *e–m*-connections of Amari. We also note that in, for example, Chib and Ramamurthy (2010) it is clear that devising successful sampling schemes for complex hierarchic models may require several different strategies to be harnessed. It is clear that manifold MCMC sampling is one sampling strategy which will prove to be important in sampling over such complex models (as the log-Gaussian Cox model in the paper) but, as in Chib and Ramamurthy (2010), cleverly deploying a variety of strategies may prove more successful than naively adopting manifold methods throughout. The, yet-to-appear, *user manual* that Draper speaks of will surely contain examples of when and where manifold methods should be used and combined.

Griffin asks how the manifold methods proposed compare with adaptive methods in general; this is a good question requiring further empirical investigation. However, we highlight that the manifold methods although being *adaptive* have the guarantee of convergence to the desired invariant measure, something which adaptive methods can only approximate.

Higher order integrators and approximations

Chin asks the question about the use of more accurate higher order solvers for the Langevin diffusion suggesting, as does Sermaidis though in an exact setting, bypassing the accept–reject step. We highlight that the Hamiltonian proposal mechanism is based on a deterministic geodesic flow across the manifold whereas the Langevin mechanism is a random diffusion which in most cases will be less efficient even, we argue, when more efficient stochastic differential equation integrators are employed. The opposite strategy of employing approximations is advocated by Honkela and the combination of employing such approximation schemes to define a geometry that is suitable for manifold MCMC sampling which will converge to the correct target density and be computationally inexpensive, appears to be a most promising avenue for methodological development. On this theme Archambeau and Bouchard suggest study of the bias–variance of manifold MCMC *versus* variational approximations, expectation propagation, and indeed the integrated nested Laplace approximation scheme.

The many suggestions for development by Cornebise and Peters are exciting and of potential methodological importance, as are those of Mira and Haario, in particular the incorporation of manifold sampling

as part of the zero-variance Monte Carlo scheme. Robert provides a provocative comment about the necessary transition from continuous Hamiltonian dynamics to the discrete computational setting, and in a similar vein Mansinghka questions the continuum–discrete divide. We conjecture that these questions may suggest relaxations as described by Mansinhgka, as well as to for example volume preservation, reversibility and adherence to the view of the MALA being a discretization of a diffusion process, requirements which Mira and Haario also challenge.

Computation

The computational cost of the manifold methods is a recurring theme in many of the discussion contributions. Roy, Honkela, Filippone, Barthelmé and Chopin, Bhadra, Campbell, Coolen, Draper, Gelman, Guerrera, Rue and Simpson, Jasra and Singh, Mansinghka, Salimans and Welling, all raise the issue and suggest possible approaches to reducing the computational cost.

Computational perspective

Firstly the computational cost of the manifold methods should be put into some perspective alongside other numerically intensive techniques that are commonly employed in day-to-day statistical practice. Consider N covariates and a data sample size $M \gg N$; using leave-one-out cross-validation to assess the value of a single regularization parameter in penalized logistic regression will incur a scaling of $\mathcal{O}(N^4)$, in the worst case, for each value of the parameter.

In the simplified MMALA the dominating $\mathcal{O}(N^3)$ scaling is the worst case scenario. The Newton method, Fisher scoring and iterated weighted least squares methods for optimization, which are routinely employed in maximum likelihood estimation, all have the same order of scaling as the simplified MMALA. For the full MMALA and RMHMC methods where the connection is defined by full matrices of metric tensor derivatives, then the additional cost of N matrix multiplications will be incurred, resulting in the worst case $\mathcal{O}(N^4)$ scaling as seen in the logistic regression example, and is of the same order as the commonly employed leave-one-out estimator.

Approximate models and iterative updating

We expect that algorithmic research work will, over time, reduce this scaling considerably. For example Honkela suggests geometries based on approximate models that may retain the main effects yet have metrics and connections with much simpler computational structures. Barthelmé and Chopin, Roy and Salimans suggest sequential updates of the metric tensor (square root) reducing the existing cubic scaling to quadratic. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) style of rank 1 updating can be employed straightforwardly in the simplified MMALA for example. However, with the iterative BFGS style updating, the MMALA proposals no longer retain their Markov structure and so fall into the category of adaptive MCMC methods where convergence to the invariant measure is approximate under conditions of *diminishing adaptation*. We have assessed the simplified MMALA using BFGS optimization on a variety of problems with mixed results. Application of BFGS style updating to RMHMC sampling requires some further investigation in terms of the effect that this will have on the symplectic properties of the integrator. An elegant approach is taken by Salimans in the stochastic volatility model where the Kalman smoother is employed, resulting in a significant speed-up in performance.

Another approach to reducing computational complexity is suggested by Bhadra where an annealing scheme is employed with the curvature components being removed once reaching the mode of the density as they will have little effect on sampling efficiency. Similarly Jasra and Singh suggest the use of a sequence of tempered densities and associated metrics in sequential Monte Carlo sampling and this may have a benefit in reducing computational costs. Looking to the extensive optimization literature may yield ideas that are useful to MCMC development; indeed Welling draws attention to the importance of the burn-in phase as *optimization with detailed balance*.

Block updating is a further way to reduce computational complexity as offered by Campbell, where interblock independence is assumed in the metric tensor. Coolen goes further by considering the possibility of finding a problem-dependent change of co-ordinates that makes the algorithm less compute intensive. This is reminiscent of the centred and non-centred parameterizations described in Papaspiliopoulos *et al.* (2007), an example of which is the *whitening* transformation of Murray and Adams.

Automatic differentiation, parallel computation and the future

Mansinghka makes a compelling case for the investigation of automatic differentiation methods for the MMALA and RMHMC algorithm. Certainly the work of Siskind and Pearlmutter (2008) makes this argument all the more persuasive and we are eager to see these tools developed further. It remains to

be seen how amenable manifold-based sampling methods will be to leveraging computational benefits in massively parallel environments, as suggested by Guerrera, Rue and Simpson. Both Welling and Gelman ask about very large data sets and large-scale models, and we agree with Welling that those of us working in this area still *have some distance to travel*.

Applications

The discussion highlights several application areas for manifold MCMC sampling, some of which may yet demand further theoretical and methodological development. Sanz-Serna mentions promising results in molecular dynamics, whereas Gripton and Christie describe an application in estimating permeability fields, and the full results of these studies will indeed be of great interest. Guillas highlights the need for efficient sampling methods for Bayesian calibration and emulation of computer models. This is a fascinating area of application which is ripe with many challenges for the statistician; indeed Cao and Wang, and Penny using the Fitzhugh–Nagumo model as an example, indicate some of these issues when calibrating non-linear differential equation models. They correctly point out that starting an MCMC run at points in the parameter space that do not include the *true* parameter value is more realistic and we have addressed this issue by the use of population MCMC sampling. Cao and Wang comment on identifiability for partially observed data. This has implications in the form of the Fisher information matrix as it will be rank deficient and points to issues such as *sloppiness*, and the pathological example that was presented by Stathopoulos and Filippone. We note that Mira and Haario obtain favourable results on the Fitzhugh–Nagumo example where a standard Metropolis–Hastings scheme is employed, with proposal covariance related to the system Jacobian at the maximum likelihood estimate. For more complex models with a larger number of parameters and partial observations the issue of *sloppiness* arises and the matter of finding the maximum likelihood estimate is a challenge in itself; see the contribution by Transtrum, Gutenkunst, Chen and Sethna. We envisage that it is in such situations that the manifold methods will be found to be particularly powerful.

Cox, Vehtari and Vanhatalo, Filippone, and Murray and Adams all raise relevant questions about the application of models with latent Gaussian processes, in particular the log-Gaussian Cox model of the paper. The main issue which arises is the joint sampling of latent functions and covariance function parameters for which Filippone illustrates the inefficiency of the scheme that is adopted in the paper. The potential of a non-centred parameterization (Papaspiliopoulos *et al.*, 2007), such as the whitening transformation, is highlighted by Murray and Adams, as well as Filippone, and is worthy of further generalization in geometric terms.

Further work will be required to address the question by Griffin of the suitability of manifold MCMC sampling on stochastic volatility model extensions. At present it remains unclear how to design manifold-based MCMC algorithms for the epidemic models with discontinuous density functions suggested by Kypraios although a combination of particle MCMC and manifold techniques may be one way forward.

User adoption

Both Roy and Honkela highlight the potential barrier to adoption of manifold Monte Carlo methods as being the need to obtain expressions for the metric and the manifold connections. This amounts to obtaining expressions for first-, second- and potentially third-order derivatives. Newton style optimization schemes require the Hessian matrix, as is the case if confidence intervals are required for maximum likelihood estimators. Indeed third-order derivatives are employed in integrated nested Laplace approximations, so we would argue that the analytical effort in obtaining the metric tensor and the components of the connection is no more than that required in setting up Newton style optimizers, variational, integrated nested Laplace approximation or expectation propagation types of approximate inference methods. It certainly is the case that a rudimentary appreciation of differential geometric concepts is required in employing this methodology and there is no doubt that this would require additional effort by potential users. As Rasmussen points out, recommendations for effective usage, a *user manual* as Draper suggests, that do not require a mastery of the technical details may be necessary to ensure wide adoption of the methodology, and this forms part of our on-going efforts.

References in the discussion

Absil, P. A., Mahony, R. and Sepulchre, R. (2008) *Optimization Algorithms on Matrix Manifolds*. Princeton: Princeton University Press.

Abt, M. and Welch, W. (1998) Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processess. *Can. J. Statist.*, **26**, 127–137.

Amari, S. (1982) Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.*, **10**, 357–385.

Amari, S.-I. (1985) Differential-geometrical methods in statistics. *Lect. Notes Statist.*, **28**.

Amari, S., Chen, T. P. and Chichocki, A. (2000) Nonholonomic orthogonal learning algorithms for blind source separation. *Neur. Computn*, **12**, 1463–1484.

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

Anaya-Izquierdo, K., Critchley, F., Marriott, P. and Vos, P. (2010) On the space of probability distributions. Submitted to *Ann. Inst. Statist. Math.*

Anaya-Izquierdo, K. A. and Marriott, P. (2007) Local mixture models of exponential families. *Bernoulli*, **13**, 623–640.

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 Moulines, E. (2006) On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, **16**, 1462–1505.

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

Assaraf, R. and Caffarel, M. (2003) Zero-variance zero-bias principle for observables in quantum monte carlo: application to forces. *J. Chem. Phys.*, **119**, 10536–10552.

Atchadé, Y. F. (2006) An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.*, **8**, 235–254.

Atchadé, Y., Fort, G., Moulines, E. and Priouret, P. (2009) Adaptive Markov chain Monte Carlo: theory and methods. *Technical Report*.

Attias, H. (1999) A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems* (eds S. A. Solla, T. K. Leen and K.-R. Müller), pp. 209–215. Cambridge: MIT Press.

Ball, F., Dryden, I. and Golalizadeh, M. (2006) Discussion on 'Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes' (by A. Beskos, O. Papaspiliopoulos, G. O. Roberts and P. Fearnhead). *J. R. Statist. Soc.* B, **68**, 367–368.

Ball, F. G., Dryden, I. L. and Golalizadeh, M. (2008) Brownian motion and Ornstein-Uhlenbeck processes in planar shape space. *Methodol. Comput. Appl. Probab.*, **10**, 1–22.

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.

Barndorff-Nielsen, O. E. and Jupp, P. E. (1997a) Yokes and symplectic structures. *J. Statist. Planng Inf.*, **63**, 133–146.

Barndorff-Nielsen, O. E. and Jupp, P. E. (1997b) Statistics, yokes and symplectic geometry. *Ann. Fac. Sci. Toul.*, **6**, 389–427.

Bates, D., Hamilton, D. and Watts, D. (1983) Calculation of intrinsic and parameter-effects curvatures for nonlinear regression models. *Communs Statist. Simuln Computn*, **12**, 469–477.

Bates, D. and Watts, D. (1980) Relative curvature measures of nonlinearity. *J. R. Statist. Soc.* B, **42**, 1–25.

Bates, D. and Watts, D. (1988) *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Beaumont, M. A., Cornuet, J. M., Marin, J. M. and Robert, C. P. (2009) Adaptive approximate Bayesian computation. *Biometrika*, **96**, 983–990.

Bertsekas, D. (1999) *Nonlinear Programming*, 2nd edn. Belmont: Athena Scientific.

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems. *Statist. Sci.*, **10**, 3–66.

Beskos, A., Papaspiliopoulos, O. and Roberts, G. O. (2008) A factorisation of diffusion measure and finite sample path constructions. *Methodol. Comput. Appl. Probab.*, **10**, 85–104.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. R. Statist. Soc.* B, **68**, 333–382.

Beskos, A., Pillai, N., Roberts, G. O., Sanz-Serna, J. M. and Stuart, A. M. (2010a) Optimal tuning of the Hybrid Monte-Carlos algorithm. *Technical Report*. Department of Statistical Science, University College London, London. (Available from `http://arxiv.org/abs/1001.4460`.)

Beskos, A., Pinski, F., Sanz-Serna, J. M. and Stuart, A. M. (2010b) Hybrid Monte-Carlo on hilbert spaces. *Technical Report*.

Beskos, H., Roberts, G., Stuart, A. and Voss, J. (2008) MCMC methods for diffusion bridges. *Stochast. Dyn.*, **8**, 319–350.

Bhadra, A. (2010) Discussion on 'Particle Markov chain Monte Carlo methods' (by C. Andrieu, A. Doucet and R. Holenstein). *J. R. Statist. Soc.* B, **72**, 314–315.

Bretó, C., He, D., Ionides, E. L. and King, A. A. (2009) Time series analysis via mechanistic models. *Ann. Appl. Statist.*, **3**, 319–348.

Brody, D. C. and Hughston, L. P. (1999) Thermalization of quantum states. *J. Math. Phys.*, **40**, 12–18.

Brown, K. S. (2003) Signal transduction, sloppy models, and statistical mechanics. *PhD Thesis*. Cornell University, Ithaca.

Brown, K., Hill, C., Calero, G., Myers, C., Lee, K., Sethna, J. and Cerione, R. (2004) The statistical mechanics of complex signaling networks: nerve growth factor signaling. *Phys. Biol.*, **1**, 184–195.

Brown, K. and Sethna, J. (2003) Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev.* E, **68**, 21904.

Burbea, J. and Rao, C. R. (1982) Entropy differential metric, distance and divergence measures in probability spaces. *J. Multiv. Anal.*, **12**, 575–596.

Burbea, J. and Rao, C. R. (1984) Differential metrics in probability spaces. *Probab. Math. Statist.*, **3**, 241–258.

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

Cardoso, J. and Laheld, B. H. (1996) Equivariant adaptive source separation. *IEEE Trans. Signal Process.*, **44**, 3017–3030.

Chib, S. and Ramamurthy, S. (2010) Tailored randomized block MCMC methods with application to DSGE models. *J. Econmetr.*, **155**, 19–38.

Choo, K. (2000) Learning hyperparameters for neural network models using Hamiltonian dynamics. *Masters Thesis*. Department of Computer Science, University of Toronto, Toronto. (Available from `http://www.cs.toronto.edu/~radford/ftp/kiam-thesis.ps`.)

Chopin, N. (2002) A sequential particle filter for static models. *Biometrika*, **89**, 539–552.

Chopin, N. (2007) Discussion of Del Moral *et al*. In *Bayesian Statistics 8* (eds S. Bayarri, J. O. Berger, J. M. Bernardo, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West). Oxford: Oxford University Press.

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.

Christensen, O. F., Roberts, G. O. and Sköld, M. (2006) Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *J. Computnl Graph. Statist.*, **15**, 1–17.

Copas, J. and Eguchi, S. (2005) Local model uncertainty and incomplete data bias (with discussion). *J. R. Statist. Soc.* B, **67**, 459–512.

Copas, J. and Eguchi, S. (2010) Likelihood for statistically equivalent models. *J. R. Statist. Soc.* B, **72**, 193–217.

Cornuet, J. M., Marin, J. M., Mira, A. and Robert, C. P. (2009) Adaptive multiple importance sampling. *Preprint*. (Available from `http://arxiv.org/abs/0907.1254`.)

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

Daniels, B., Chen, Y., Sethna, J., Gutenkunst, R. and Myers, C. (2008) Sloppiness, robustness, and evolvability in systems biology. *Curr. Opin. Biotechnol.*, **19**, 389–395.

Das, S., Spall, J. and Ghanem, R. (2010) Efficient Monte Carlo computation of Fisher information matrix using prior information. *Computnl Statist. Data Anal.*, **54**, 272–289.

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.

Dellaportas, P. and Kontoyiannis, I. (2010) Control variates for reversible MCMC samplers. Submitted to *J. R. Statistic. Soc.* B.

Del Moral, P., Doucet, A. and Jasra, A. (2006) Sequential Monte Carlo samplers. *J. R. Statist. Soc.* B, **68**, 411–436.

Diaconis, P., Holmes, S. and Neal, R. (2000) Analysis of a nonreversible Markov chain sampler. *Ann. Appl. Probab.*, **10**, 726–752.

Draper, D. and Liu, S. (2006) MCMC acceleration: methods and results. *Technical Report*. Department of Applied Mathematics and Statistics, University of California, Santa Cruz.

Dryden, I. L., Pennec, X. and Peyrat, J.-M. (2010) Power Euclidean metrics for covariance matrices with application to diffusion tensor imaging. *Technical Report*. University of Nottingham, Nottingham. (Available from `http://arxiv.org/abs/1009.3045`.)

Durbin, J. and Koopman, S. (2001) *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

van Dyk, D. A. and Meng, X.-L. (2010) Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: a graphical guide book. *Statist. Sci.*, to be published.

Edelman, A., Arias, T. and Smith, S. (1999) The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Applic.*, **20**, 303–353.

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

Eguchi, S. (1983) Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Statist.*, **11**, 793–803.

Fitzhugh, R. (1961) Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.*, **1**, 445–466.

Fletcher, R. (1987) *Practical Methods of Optimization*, 2nd edn. New York: Wiley.

Forbert, H. A. and Chin, S. A. (2000) Fourth-order algorithms for solving the multi-variable Langevin equation and the Kramers equation. *Phys. Rev.* E, **63**, 016703.
Gelman, A. (2007) Struggles with survey weighting and regression modeling (with discussion). *Statist. Sci.*, **22**, 153–188.
Gelman, A. and Ghitza, Y. (2010) Who votes?: how did they vote?: and what were they thinking? *Technical Report*. Department of Political Science, Columbia University, New York.
Gelman, A. E. and Meng, X.-L. (1998) Computing normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, **13**, 163–185.
Gelman, A., Park, D., Shor, B. and Cortina, J. (2009) *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*, 2nd edn. Princeton: Princeton University Press.
Geyer, C. J. (1991) Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proc. 23rd Symp. Interface*, pp. 156–163.
Gibson, G. and Renshaw, E. (1998) Estimating parameters in stochastic compartmental models using Markov Chain methods. *IMA J. Math. Appl. Med. Biol.*, **15**, 19–40.
Giordani, P. and Kohn, R. (2010) Adaptive independent Metropolis-Hastings by fast estimation of mixtures of normals. *J. Computnl Graph. Statist.*, **19**, 243–259.
Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
Green, P. J. and Hastie, D. I. (2009) Reversible jump MCMC. *Technical Report*.
Grenander, U. and Miller, M. I. (1994) Representations of knowledge in complex systems (with discussion). *J. R. Statist. Soc.* B, **56**, 549–603.
Grenander, U., Miller, M. I. and Srivastava, A. (1998) Hilbert-Schmidt lower bounds for estimators on matrix Lie groups for ATR. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 790–802.
Guillas, S., Rougier, J., Maute, A., Richmond, A. D. and Linkletter, C. D. (2009) Bayesian calibration of the thermosphere-ionosphere electrodynamics general circulation model (TIE-GCM). *Geosci. Model Dev.*, **2**, 137–144.
Gutenkunst, R. (2008) Sloppiness, modeling, and evolution in biochemical networks. *PhD Thesis*. Cornell University, Ithaca.
Gutenkunst, R., Waterfall, J., Casey, F., Brown, K., Myers, C. and Sethna, J. (2007) Universally sloppy parameter sensitivities in systems biology models. *PLOS Comput. Biol.*, **3**, no 10, e189.
Haario, H., Laine, M., Mira, A. and Saksman, E. (2006) DRAM: efficient adaptive MCMC. *Statist. Comput.*, **16**, 339–354.
Haario, H., Saksman, E. and Tamminen, J. (1999) Adaptive proposal distribution for random walk Metropolis algorithm. *Computnl Statist.*, **14**, 375–395.
Haario, H., Saksman, E. and Tamminen, J. (2001) An adaptive Metropolis algorithm *Bernoulli*, **7**, 223–242.
He, D., Ionides, E. L. and King, A. A. (2010) Plug-and-play inference for disease dynamics: measles in large and small towns as a case study. *J. R. Soc. Interface*, **7**, 271–283.
Henderson, D. A., Boys, R. J., Proctor, C. J. and Wilkinson, D. J. (2010) Linking systems biology models to data: a stochastic kinetic model of p53 oscillations. In *The Oxford Handbook of Applied Bayesian Analysis* (eds A. O' Hagan and M. West), pp. 155–187. Oxford: Oxford University Press.
Hestenes, M. R. and Stiefel, E. (1952) Methods of conjugate gradients for solving linear systems. *J. Res. Natn Bur. Stand.*, **49**, 409–436.
Holmes, C. C. and Held, L. (2005) Bayesian auxiliary variable models for binary and multinomial regression. *Baysn Anal.*, **1**, 145–168.
Honkela, A., Raiko, T., Kuusela, M., Tornio, M. and Karhunen, J. (2010) Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *J. Mach. Learn. Res.*, **11**, 3235–3268.
Horowitz, A. M. (1991) A generalized guided monte carlo algorithm. *Phys. Lett.* B, **268**, 247–252.
Hughston, L. P. (1996) Geometry of stochastic state vector reduction. *Proc. R. Soc. Lond.*, **452**, 953–979.
Husmeier, D. (2000) The Bayesian evidence scheme for regularising probability-density estimating neural networks. *Neur. Computn*, **12**, 2685–2717.
Imai, K. and van Dyk, D. A. (2005) A bayesian analysis of the multinomial probit model using marginal data augmentation. *J. Econmetr.*, **124**, 311–334.
Ionides, E. L., Bretó, C. and King, A. A. (2006) Inference for nonlinear dynamical systems. *Proc. Natn. Acad. Sci. USA*, **103**, 18438–18443.
Jeffreys, H. (1998) *Theory of Probability*. New York: Oxford University Press.
Jupp, P. E. (2010) A van Trees inequality for estimators on manifolds. *J. Multiv. Anal.*, **101**, 1814–1825.
Kass, R. E. and Vos, P. W. (1997) *Geometrical Foundations of Asymptotic Inference*. New York: Wiley.
Kennedy, M. C. and O' Hagan, A. (2001) Bayesian calibration of computer models (with discussion). *J. R. Statist. Soc.* B, **63**, 425–464.
Koutis, I., Miller, G. L. and Peng, R. (2010) Approaching optimality for solving SDD linear systems. In *Proc. 51st A. Symp. Foundations of Computer Science*. Silver Spring: Institute of Electrical and Electronics Engineers Computer Society Press.
Kypraios, T. (2007) Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semiparametric time series models. *PhD Thesis*. Department of Mathematics and Statistics, Lancaster

University, Lancaster. (Available from http://www.maths.nott.ac.uk/personal/tk/files/Kyp07.pdf.)

Laneri, K., Bhadra, A., Ionides, E. L., Bouma, M., Dhiman, R. C., Yadav, R. S. and Pascual, M. (2010) Forcing versus feedback: epidemic malaria and monsoon rains in NW India. *PLOS Computnl Biol.*, **6**, e1000898.

Lelievre, T., Otto, F., Rousset, M. and Stoltz, G. (2008) Long-time convergence of an Adaptive Biasing Force method. *Nonlinearity*, **21**, 1155–1181.

Li, P., Chen, J. and Marriott, P. (2009) Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 411–426.

Lindgren, F., Lindström, J. and Rue, H. (2010) An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach. *Technical Report 5*. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.

Liu, S. (2003) Mirror-jump sampling: a strategy for MCMC acceleration. *Masters Project*. Department of Computer Science, University of California, Santa Cruz.

Loh, W. (2005) Fixed-domain asymptotics for a subclass of Matern-type Gaussian random fields. *Ann. Statist.*, **33**, 2344–2394.

Lott, J. and Villani, C. (2009) Ricci curvature for metric-measure spaces via optimal transport. *Ann. Math.*, **169**, 903–991.

Machta, B. B., Transtrum, M. K., Chen, Y.-J. and Sethna, J. P. (2010) Information geometry and Bayesian priors. Unpublished.

Marin, J. M. and Robert, C. P. (2007) *Bayesian Core: a Practical Approach to Computational Bayesian Statistics*. New York: Springer.

Marin, J. and Robert, C. (2010) Importance sampling methods for Bayesian discrimination between embedded models. In *Frontiers of Statistical Decision Making and Bayesian Analysis* (eds M.-H. Chen, D. Dey, P. Müller, D. Sun and K. Ye), ch. 14. New York: Springer.

McCulloch, R. E., Polson, N. G. and Rossi, P. E. (2000) A bayesian analysis of the multinomial probit model with fully identified parameters. *J. Econmetr.*, **99**, 173–193.

Meng, X.-L. and Schilling, S. (2002) Warp bridge sampling. *J. Computnl Graph. Statist.*, **11**, 552–586.

Mengersen, K. L. and Tweedie, R. L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, **24**, 101–121.

Minka, T. (2001) A family of algorithms for approximate Bayesian inference. *PhD Thesis*. Massachusetts Institute of Technology, Cambridge.

Mira, A. and Geyer, C. J. (2000) On reversible Markov chains. *Flds Inst Communs Monte Carlo Meth.*, **26**, 93–108.

Mira, A., Solgi, R. and Imparato, D. (2010) Zero-variance Markov chain Monte Carlo for Bayesian estimators. *Technical Report*. University of Insubria, Insubria. (Available from http://arxiv.org/abs/1012.2983.)

Morris, J. S. and Carroll, R. J. (2006) Wavelet-based functional mixed models. *J. R. Statist. Soc.* B, **68**, 179–199.

Murray, I. and Adams, R. P. (2010) Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems* (eds J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor and A. Culotta), pp. 1723–1731.

Neal, R. M. (1993) Probabilistic inference using Markov Chain Monte Carlo Methods. *Technical Report*. University of Toronto, Toronto. (Available from http://www.cs.toronto.edu/~radford/review.abstract.html.)

Neal, R. M. (1999) Regression and classification using Gaussian process priors (with discussion). *Baysn Statist.*, **6**, 475–501.

Neal, R. M. (2001) Annealed importance sampling. *Statist. Comput.*, **11**, 125–139.

Neal, R. M. (2003) Slice sampling. *Ann. Statist.*, **31**, 705–767.

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: Chapman and Hall–CRC Press.

Neal, P. and Roberts, G. (2005) A case study in non-centering for data augmentation: stochasic epidemics. *Statist. Comput.*, **15**, 315–327.

Nevat, I., Peters, G. W. and Yuan, J. (2009) Channel estimation in OFDM systems with unknown power delay profile using trans-dimensional MCMC via stochastic approximation. In *Proc. Vehicular Technology Conf.*, pp. 1–6. New York: Institute of Electrical and Electronics Engineers.

Nobile, A. (1998) A hybrid markov chain for the bayesian analysis of the multinomial probit model. *Statist. Comput.*, **8**, 229–242.

Nobile, A. (2000) Comment: Bayesian multinomial probit models with a normalization constraint. *J. Econmetr.*, **99**, 335–345.

Okabayashi, S. and Geyer, C. J. (2010) Long range search for maximum likelihood in exponential families. *Technical Report*.

Oliver, D., He, N. and Reynolds, A. C. (1996) Conditioning permeability fields to pressure data. In *Proc. 5th Eur. Conf. Mathematics of Oil Recovery, Sept.*

O'Neill, P. D. and Roberts, G. O. (1999) Bayesian inference for partially observed stochastic epidemics. *J. R. Statist. Soc.* A, **162**, 121–129.

Opper, M. and Winther, O. (2000) Gaussian processes for classification: mean field algorithms. *Neur. Computn*, **12**, 2655–2684.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007) A general framework for the parametrization of hierarchical models. *Statist. Sci.*, **22**, 59–73.

Pearlmutter, B. A. and Siskind, J. M. (2008) Reverse-mode AD in a functional framework: Lambda the ultimate backpropagator. *ACM Trans. Program. Lang. Syst.*, **30**, no. 2.

Peluchetti, S. and Roberts, G. O. (2008) An empirical study of the efficiency of the EA for diffusion simulation. *Technical Report.* University of Warwick, Coventry.

Peters, G. W., Hosack, G. R. and Hayes, K. R. (2010) Ecological non-linear state space model selection via adaptive particle Markov chain Monte Carlo (AdPMCMC). *Technical Report*.

Poyiadjis, G., Doucet, A. and Singh, S. S. (2010) Particle approximations of the score and observed information matrix in state-space models with application to parameter estimation. *Biometrika*, to be published.

Ramsay, J., Hooker, H., Campbell, D. and Cao, J. (2007) Parameter estimation for differential equations: a generalized smoothing approach (with discussion). *J. R. Statist. Soc.* B, **69**, 741–796.

Rasmussen, C. E. and Williams, C. (2006) *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* B, **59**, 731–792; correction, **60** (1998), 661.

Ridall, P. G., Pettitt, A. N., Friel, N., Henderson, R. and McCombe, P. (2007) Motor unit number estimation using reversible jump Markov chain Monte Carlo (with discussion). *Appl. Statist.*, **56**, 235–269.

Robert, C. and Casella, G. (1999) *Monte Carlo Statistical Methods*, 1st edn. New York: Springer.

Roberts, S. and Choudrey, R. (2005) Bayesian independent component analysis with prior constraints: an application in biosignal analysis, deterministic and statistical methods in machine learning. *Lect. Notes Comput. Sci.*, **3635**, 159–179.

Roberts, S. J., Husmeier, D., Rezek, I. and Penny, W. (1998) Bayesian approaches to Gaussian mixture modeling. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 1133–1142.

Roberts, G. O. and Rosenthal, J. S. (2001) Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, **16**, 351–367.

Roberts, G. O. and Rosenthal, J. S. (2009) Examples of adaptive MCMC. *J. Computnl Graph. Statist.*, **18**, 349–367.

Roberts, G. O. and Tweedie, R. L. (1996) Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, **2**, 341–363.

Rosenthal, J. S. (2010) Optimal proposal distributions and adaptive MCMC. In *Handbook of Markov Chain Monte Carlo* (eds S. Brooks, A. Gelman, G. Jones and X.-L. Meng). Boca Raton: Chapman and Hall–CRC Press.

Saad, Y. (2003) *Iterative Methods for Sparse Linear Systems*, 2nd edn, ch. 6. Philadelphia: Society for Industrial and Applied Mathematics.

Sanz-Serna, J. M. (1997) Geometric integration. In *The State of the Art in Numerical Analysis* (eds I. S. Duff and G. A. Watson), pp. 121–143. Oxford: Clarendon.

Schmidt, M. N. and Laurberg, H. (2008) Nonnegative matrix factorization with Gaussian process priors. *Computnl Intell. Neursci.*, 1–10.

Siskind, M. and Pearlmutter, B. (2008) Nesting forward-mode AD in a functional framework. *High. Ord. Symbol. Computn*, **21**, 361–376.

Spall, J. (2005) Monte Carlo computation of the Fisher information matrix in nonstandard settings. *J. Comput. Graph. Statist.*, **14**, 889–909.

Squartini, S., Piazza, F. and Theis, F. (2006) New Riemannian metrics for speeding-up the convergence of over- and underdetermined ICA. In *Proc. Int. Symp. Circuits and Systems*, *Kos*. New York: Institute of Electrical and Electronics Engineers.

Srivastava, A. (2000) A Bayesian approach to geometric subspace estimation. *IEEE Trans Signal Process.*, **48**, 1390–1400.

Srivastava, A., Grenander, U., Jensen, G. R. and Miller, M. I. (2002) Jump-diffusion markov processes on orthogonal groups for objects recognition. *J. Statist. Planng Inf.*, **103**, 15–37.

Srivastava, A. and Klassen, E. (2001) Monte Carlo extrinsic estimators for manifold-valued parameters. *IEEE Trans. Signal Process.*, **50**, 299–308.

Stramer, O. and Tweedie, R. (1999a) Langevin-type models I: diffusions with given stationary distributions, and their discretizations. *Methodol. Comput. Appl. Probab.*, **1**, 283–306.

Stramer, O. and Tweedie, R. (1999b) Langevin-type models II: self-targeting candidates for Hastings-Metropolis algorithms. *Methodol. Comput. Appl. Probab.*, **1**, 307–328.

Theis, F. (2005) Gradients on matrix manifolds and their chain rule. *Neur. Inform. Process.*, **9**, 1–13.

Transtrum, M. K., Chen, Y.-J. and Sethna, J. P. (2010) Geodesics in Monte Carlo sampling. Unpublished.

Transtrum, M. K., Machta, B. B. and Sethna, J. P. (2010a). Why are nonlinear fits to data so challenging? *Phys. Rev. Lett.*, **104**, 1060201.

Transtrum, M. K., Machta, B. B. and Sethna, J. P. (2010b) The geometry of nonlinear least squares with applications to sloppy models and optimization. To be published.

van Trees, H. L. (1968) *Detection, Estimation and Modulation Theory, Part 1*. New York: Wiley.

Vanhatalo, J., Pietiläinen, V. and Vehtari, A. (2010) Approximate inference for disease mapping with sparse Gaussian processes. *Statist. Med.*, **29**, 1580–1607.

Vanhatalo, J. and Vehtari, A. (2007) Sparse log Gaussian processes via MCMC for spatial epidemiology. *J. Mach. Learn. Res. Wrkshp Conf. Proc.*, **1**, 73–89.

Wilkinson, D. J. and Golightly, A. (2010) Markov chain Monte Carlo algorithms for SDE parameter estimation. In *Learning and Inference in Computational Systems Biology* (eds N. Lawrence, M. Girolami, M. Rattray and G. Sanguinetti), pp. 253–275. Cambridge: MIT Press.

Zhang, J. (2004) Divergence function, duality, and convex analysis. *Neur. Computn*, **16**, 159–195.

Zhong, M. and Girolami, M. (2009) Reversible jump MCMC for non-negative matrix factorization. In *Proc. 12th Int. Conf. Artificial Intelligence and Statistics, Clearwater Beach*, vol. 5, pp. 663–670.

Zlochin, M. and Baram, Y. (2001) Manifold stochastic dynamics for Bayesian Learning. *Neur. Computn*, **13**, 2549–2572.