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

 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

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

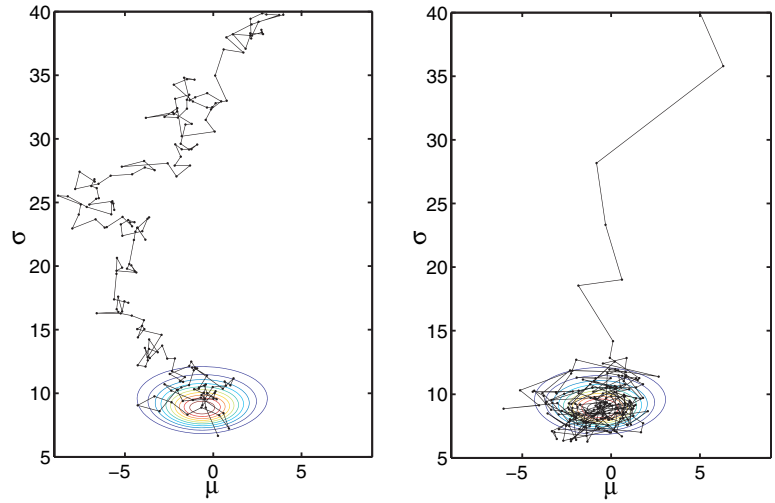

Fig. 1. Contours representing the sample estimate of $p(\mu, \sigma|X)$ where a sample of size $N = 30$ was drawn from $\mathcal{N}(X|\mu = 0, \sigma = 10)$ (both MALA and MMALA discrete diffusions were forward simulated from initial points $\mu_0 = 5$ and $\sigma_0 = 40$ with a step size $\varepsilon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

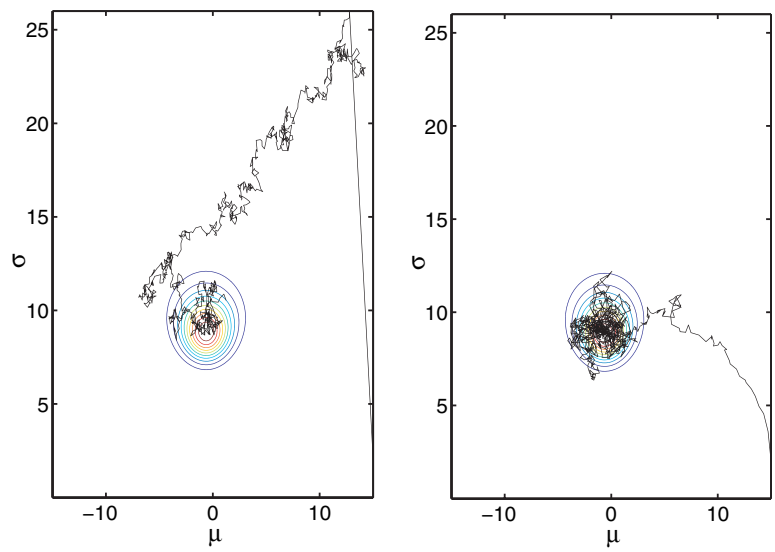

Fig. 2. Same data sample as in Fig. 1, but with $\mu_0 = 15$ and $\sigma_0 = 2$ (the step size is reduced to $\varepsilon = 0.2$ so that the MALA converges and 1000 proposal steps are taken): as in Fig. 1, from (a) it is clear that the Euclidean metric of the MALA does not exploit the hyperbolic geometry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

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

6. Riemann manifold Hamiltonian Monte Carlo methods

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

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

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

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

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

is the desired target density.

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

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

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

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

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

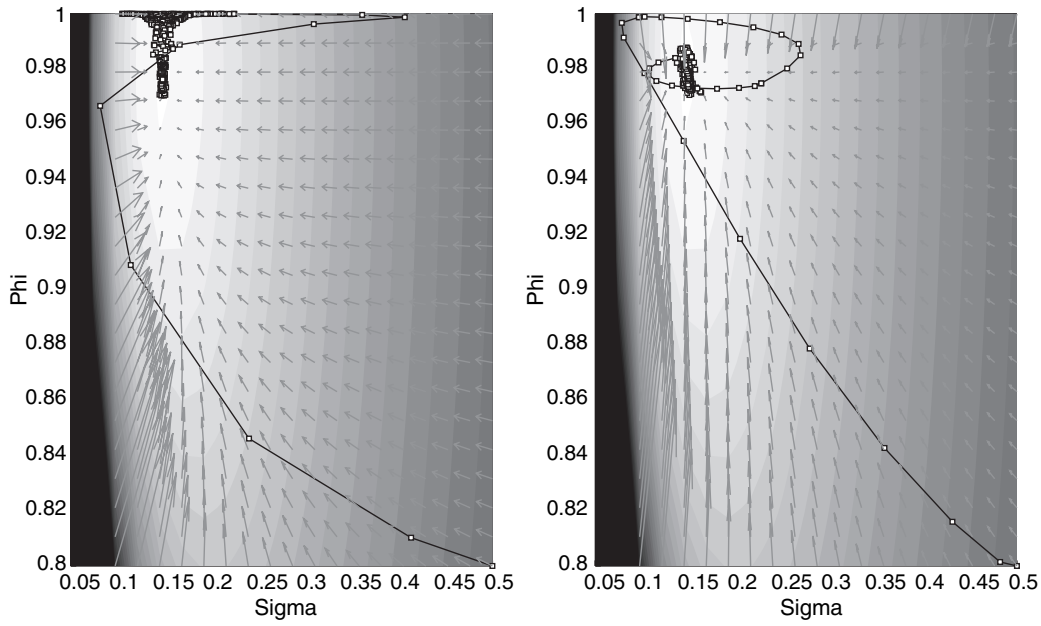

Fig. 3. Contours plotted from the stochastic volatility model investigated later in Section 8 (the latent volatilities and the parameter β are set to their true values, whereas the log-joint-probability given different values of the parameters σ and ϕ is shown by the contour plot): (a) evolution of a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

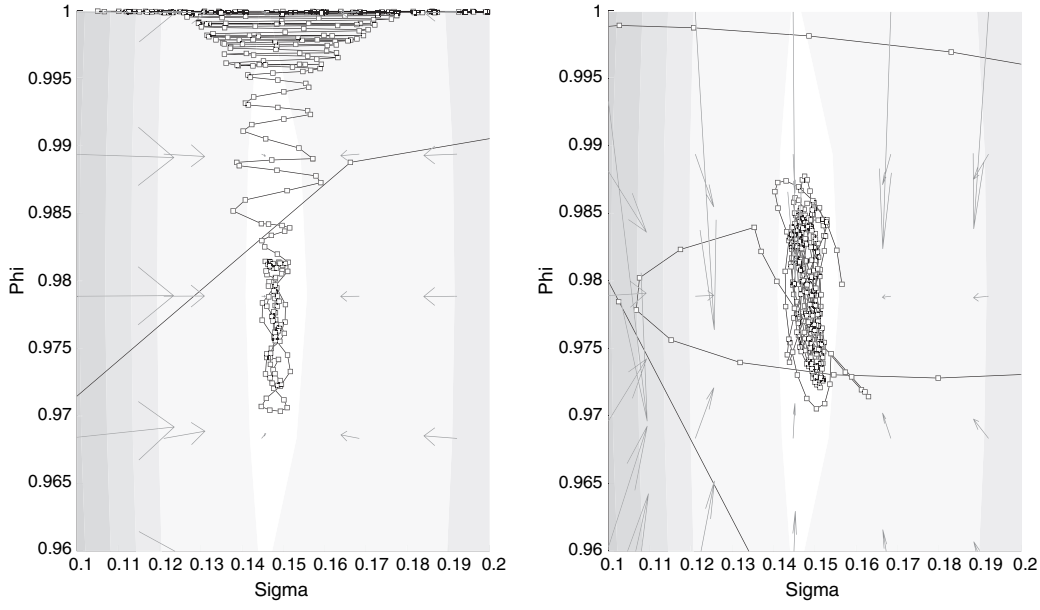

Fig. 4. Close-up of the Markov chain paths shown in Fig. 3: it is clear that RMHMC sampling effectively normalizes the gradients in each direction, whereas HMC sampling, with a unit mass matrix, exhibits stronger gradients along the horizontal direction compared with the vertical direction and therefore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

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

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

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

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

7.1. Experimental results for Bayesian logistic regression

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

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

Table 1. Summary of data sets for logistic regression

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

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

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

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

7.1.1. Metropolis–Hastings scheme

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

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

7.1.2. Auxiliary variable Gibbs sampler

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

7.1.3. Metropolis adjusted Langevin algorithm

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

7.1.4. Hamiltonian Monte Carlo sampling

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

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

7.1.5. Iterated weighted least squares

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

to the simplified MMALA without tunable step size. This method is relatively straightforward to implement and has the advantage that it requires no tuning, similarly to the auxiliary variable Gibbs sampler of Holmes and Held (2005).

7.2. Comparison of Markov chain Monte Carlo methods

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

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

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

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

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

As the number of covariates in the data set increases, so the overall performance of RMHMC sampling and the MMALA decreases owing to the increased computational burden of calculating partial derivatives of the metric tensor with respect to each of the covariates. It is clear that for logistic regression problems with a very high number of covariates, e.g. in excess of 100, the use of manifold methods will become computationally infeasible if the metric tensor is position

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

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

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

Method	*Time (s)*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	10.8	(314, 709, 979)	0.034	320
Auxiliary variables	407.5	(37.4, 1054, 1405)	10.9	1
MALA	2.7	(22.3, 576.8, 990.6)	0.122	89.3
HMC	87.3	(3197, 3612, 3982)	0.027	403
IWLS	5.15	(130, 215, 346)	0.040	272
MMALA	11.7	(702, 867, 1037)	0.0167	652
Simplified MMALA	3.2	(487, 625, 746)	0.006	1817
RMHMC	81.7	(4975, 5000, 5000)	0.016	681
RMHMC (Student t)	87.3	(1083, 1625, 2002)	0.081	134

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

Method	*Time (s)*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	23.4	(167, 613, 1015)	0.140	4.4
Auxiliary variables	618.8	(1006, 2211, 2640)	0.614	1
MALA	3.5	(95.5, 316, 667)	0.037	16.6
HMC	117.9	(3182, 3632, 3986)	0.037	16.6
IWLS	9.3	(253, 572, 918)	0.037	16.6
MMALA	42.3	(604, 766, 902)	0.070	8.8
Simplified MMALA	5.0	(435, 615, 747)	0.012	51.2
RMHMC	246.6	(4757, 5000, 5000)	0.052	11.8
RMHMC (Student t)	257.4	(3981, 4934, 5000)	0.065	9.4

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

Method	*Time (s)*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	3.8	(347, 552, 980)	0.011	19.3
Auxiliary variables	304.3	(1432, 1888, 2295)	0.212	1
MALA	1.56	(316, 550, 895)	0.005	42.4
HMC	45.7	(3265, 3605, 3893)	0.014	15.1
IWLS	2.6	(1386, 1937, 2379)	0.0019	111.6
MMALA	4.2	(1135, 1286, 1412)	0.0037	57.3
Simplified MMALA	1.9	(1046, 1160, 1300)	0.0018	117.8
RMHMC	34.4	(5000, 5000, 5000)	0.0069	30.7
RMHMC (Student t)	38.6	(3928, 4432, 4688)	0.0098	21.6

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

Method	*Time (s)*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	4.4	(418, 637, 905)	0.010	21.2
Auxiliary variables	150.9	(711, 1233, 1676)	0.212	1
MALA	1.1	(279, 524, 814)	0.0038	55.8
HMC	27.6	(3246, 3647, 4003)	0.0085	24.9
IWLS	2.4	(87, 186, 381)	0.028	7.6
MMALA	5.6	(656, 789, 903)	0.0085	24.9
Simplified MMALA	1.6	(371, 481, 617)	0.0043	49.3
RMHMC	42.2	(4862, 5000, 5000)	0.0087	24.4
RMHMC (Student t)	48.0	(2603, 2904, 3171)	0.018	11.8

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

Method	*Time (s)*	*ESS (minimum, median, maximum)*	*s/minimum ESS*	*Relative speed*
Metropolis	2.1	(59, 99, 271)	0.035	210
Auxiliary variables	139.6	(19, 44, 283)	7.35	1
MALA	0.97	(33, 58, 101)	0.029	253
HMC	24.8	(3326, 3719, 4053)	0.0076	967
IWLS	1.46	(101, 207, 328)	0.015	490
MMALA	3.3	(447, 579, 685)	0.0075	980
Simplified MMALA	1.3	(291, 403, 473)	0.0045	1633
RMHMC	28.0	(4273, 4677, 4961)	0.0065	1131
RMHMC (Student t)	31.9	(2829, 3088, 3289)	0.011	668

dependent and the same number of derivative matrices as covariates must be computed and manipulated. From a practical perspective we can impose a constant curvature manifold on the logistic regression model by employing a constant metric tensor of the form $\mathbf{G} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \alpha^{-1}\mathbf{I}$ emerging from the linear regression model. This captures the correlation structure in the covariates and as such provides an obvious decorrelating operator for the manifold methods. This will now have the same computational burden as the non-manifold methods.

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

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

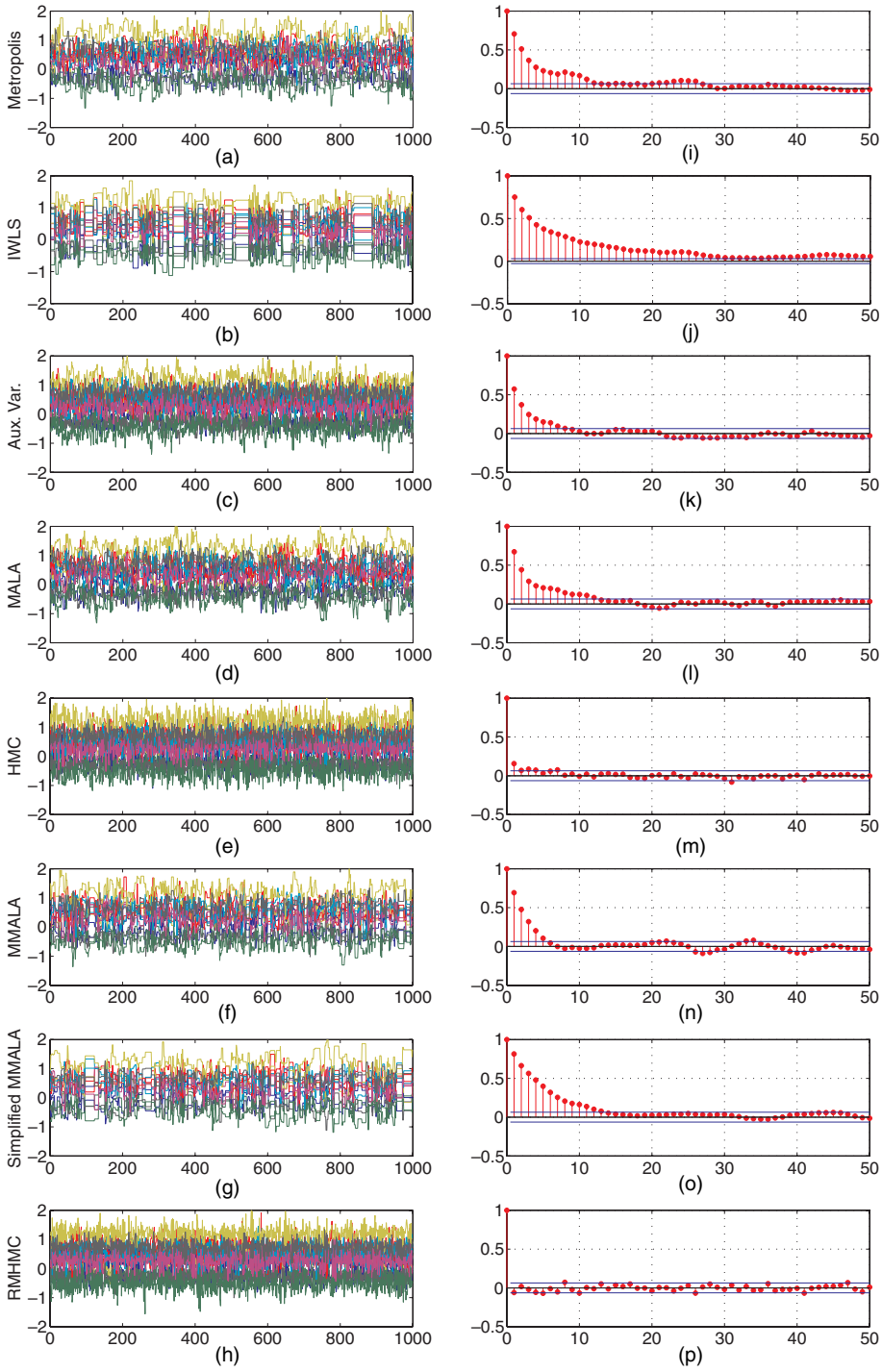

Fig. 5. Trace plots for 1000 posterior samples for the heart data set by using (a) Metropolis sampling, (b) IWLS, (c) the auxiliary variable sampler, (d) the standard MALA, (e) standard HMC sampling, (f) the MMALA, (g) the simplified MMALA and (h) RMHMC sampling: (i)–(p) corresponding auto-correlation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

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

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

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

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

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

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

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

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

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

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

Defining $\mathbf{s} = (s_1, \ldots, s_T)^{\mathrm{T}}$ 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, 20 000 posterior samples were collected after a burn-in period of 10 000 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

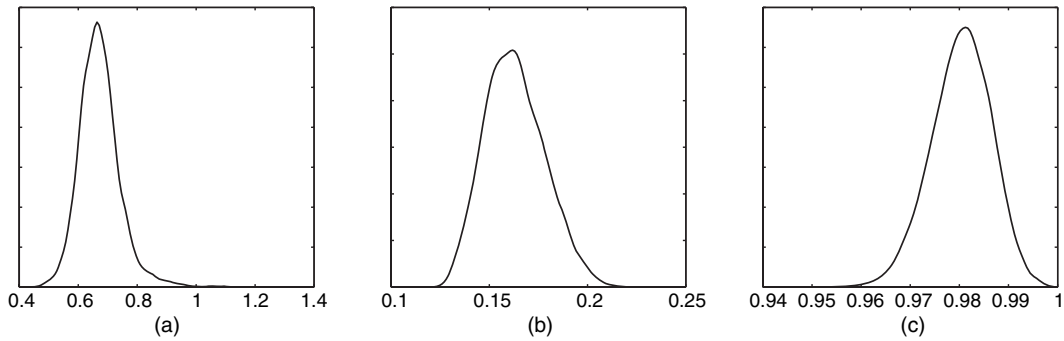

Fig. 6. Posterior marginal densities for (a) β, (b) σ and (c) ϕ, employing RMHMC sampling to draw 20 000 samples of the parameters and latent volatilities by using a simulated data set consisting of 2000 observations: the true values are $\beta = 0.65$, $\sigma = 0.15$ and $\phi = 0.98$

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 $\delta(i, i', j, j') = \sqrt{\{(i - i')^2 + (j - j')^2\}}$. The joint density is

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

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

for 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

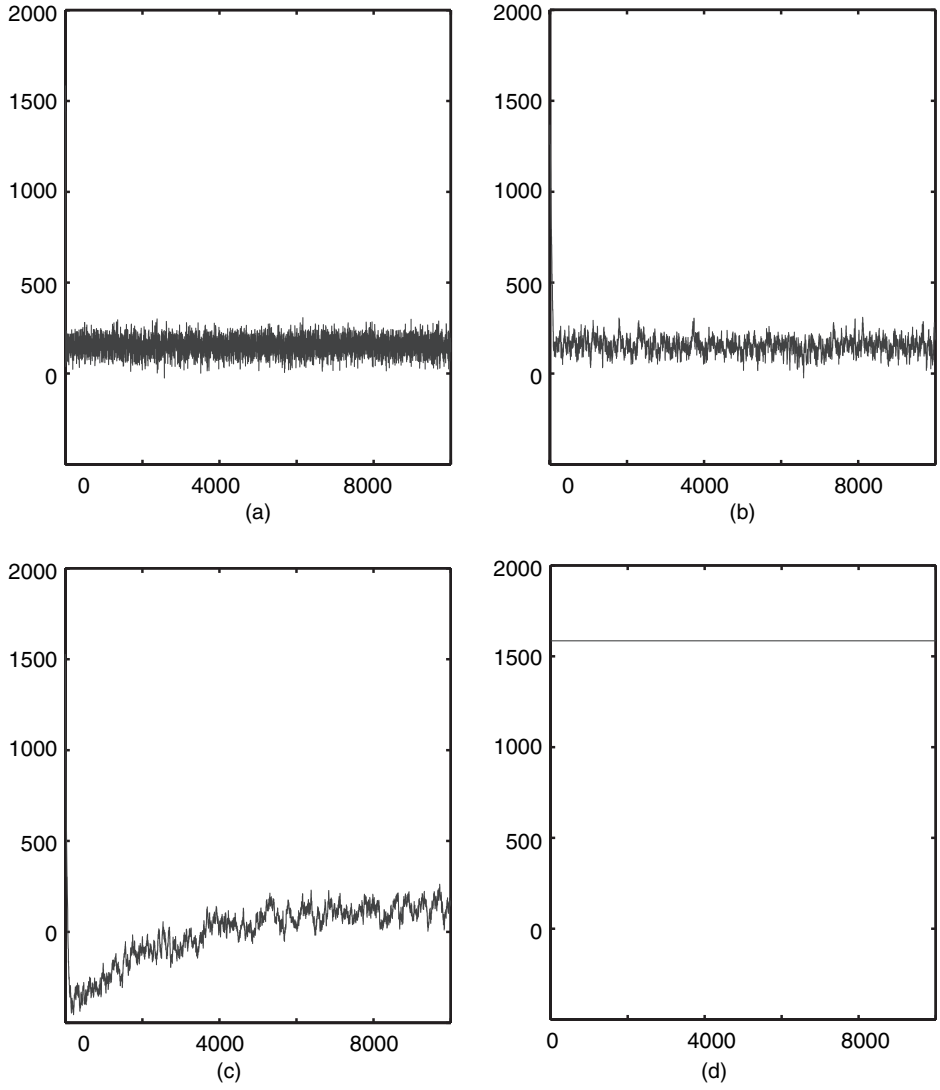

Fig. 7. Trace plots of the log-joint probability for the first 10 000 samples of the latent variables of a log-Gaussian Cox process: (a) convergence of the RMHMC scheme which can directly sample the latent variables **x** without the need for *ad hoc* reparameterizations and pilot runs for fine-tuning;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

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

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

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

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

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

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

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

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

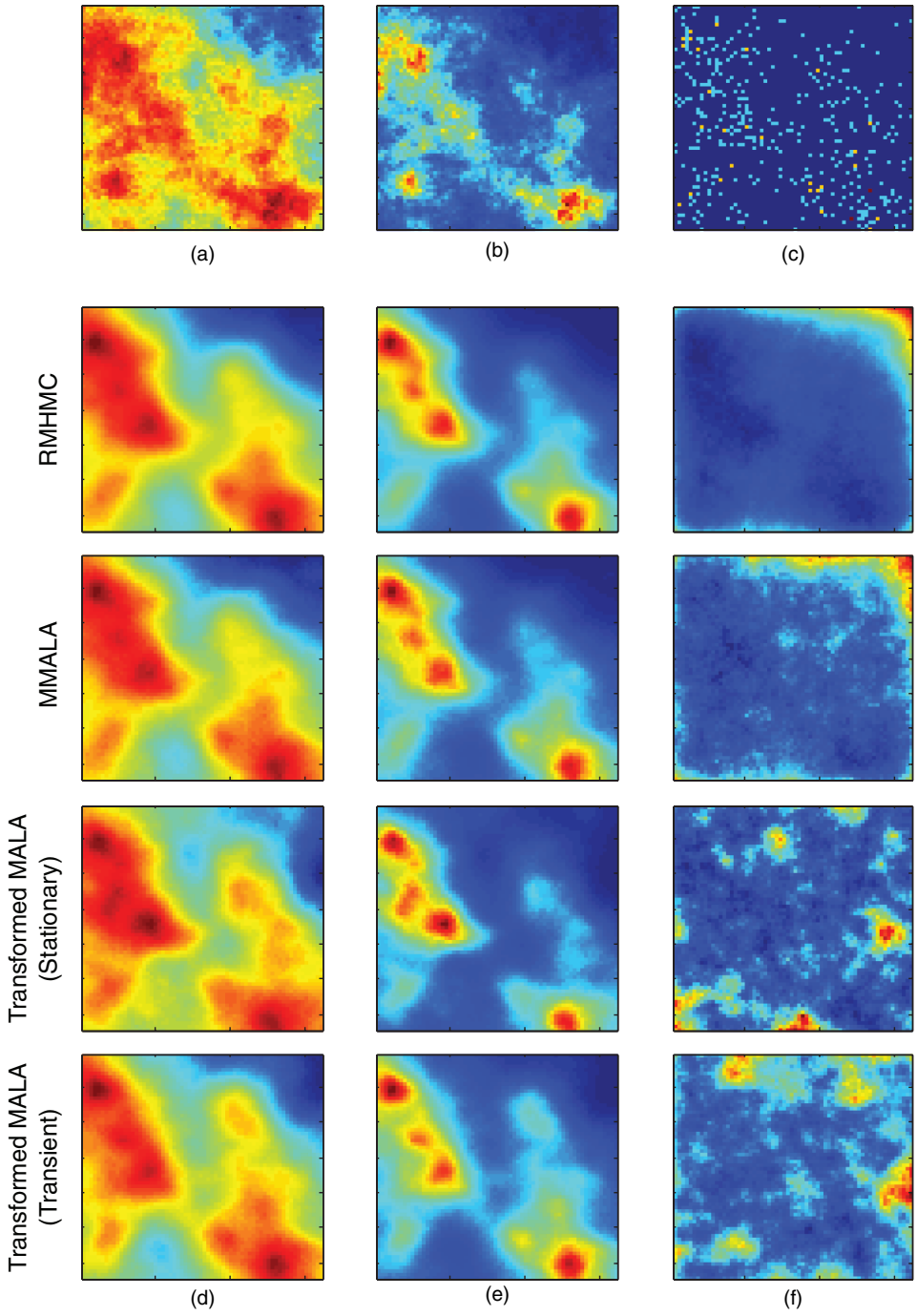

Fig. 8. Posterior latent fields and processes and associated variance, using each of the sampling methods, compared with the true latent field and process (the data employed to infer the latent field are shown in (c); RMHMC sampling produces the lowest variance estimates, which corresponds with its having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

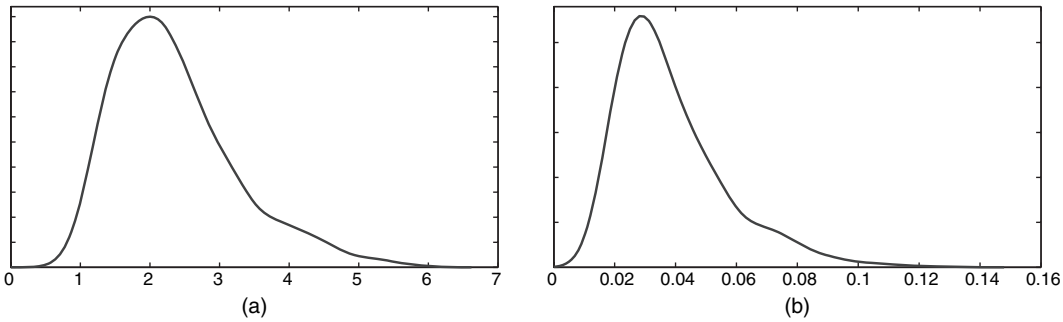

Fig. 9. Kernel density estimates of the hyperparameter samples obtained from Gibbs style sampling from the log-Gaussian Cox model: (a) true value $\sigma = 1.9$; (b) true value $\beta = 0.03$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

10.1. Experimental results for non-linear differential equations

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

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

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

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

10.1.1. Comparison of sampling schemes

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

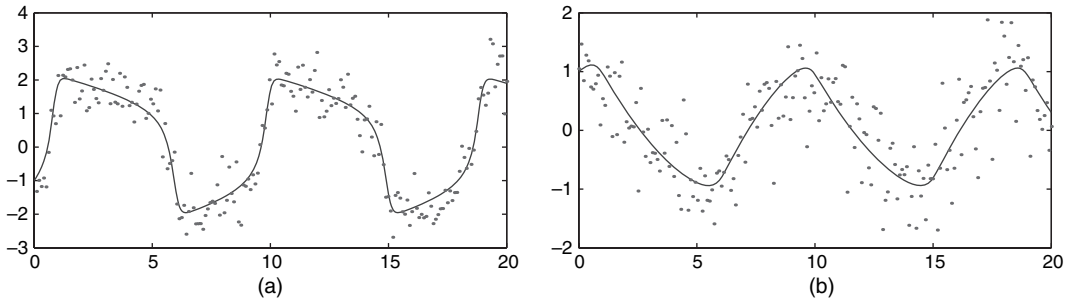

Fig. 10. Output for (a) species V and (b) species R of the Fitzhugh–Nagumo model with parameters $a = 0.2$, $b = 0.2$ and $c = 3$: •, noisy data set

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

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

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

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

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

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

11. Conclusions and discussion

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

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

Interestingly the simplified MMALA method can be seen to employ a drift term which is based on the natural gradient as defined in Amari and Nagaoka (2000) and this form of natural gradient, which is the contravariant gradient as defined in differential geometry, has been exploited in approximate Bayesian inference by Honkela *et al.* (2010). On page 319 of Robert and Casella (2004) the MALA is derived from a second-order approximation of the target density, which can also be seen to be a simplified MMALA where the metric is the negative of the Hessian matrix; a similar approach was taken in Qi and Minka (2002) when seeking to exploit the Hessian in the design of MCMC methods. The geometric perspective that is adopted in this paper provides the overarching framework that generalizes these specific approaches.

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

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

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

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

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

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

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

Acknowledgements

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

Appendix A: Expressions required for stochastic volatility model

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and the Fisher information matrix also follows in standard form as

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

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

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

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

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

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

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

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

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

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

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

Appendix C: Partial derivatives for ordinary differential equation example

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987) Hybrid Monte Carlo. *Phys. Lett.* B, **195**, 216–222.

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

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

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

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

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

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

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

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

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

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

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

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

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

Honkela, A., 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.

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

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

Discussion on the paper by Girolami and Calderhead

J. M. Sanz-Serna (*Universidad de Valladolid*)

The Metropolis adjusted Langevin algorithm (MALA) and the hybrid (or Hamilitonian) Monte Carlo (HMC) method are successful Markov chain Monte Carlo methods that use proposals based on knowledge of the target probability distribution and consequently outperform algorithms with random-walk proposals. That improvement does not come without a price tag. Both methods include a user-determined matrix (the mass matrix in the HMC method; the preconditioner in the MALA) as a 'free parameter' whose

tuning is a difficult art which in practice requires expensive trial runs. This stimulating paper by Girolami and Calderhead provides theoretical guidance into the choice of matrices and does so by exploiting two main ideas.

(a) The algorithms are generalized to incorporate a mass matrix or preconditioner that is a *function* of the state of the Markov chain. At first sight this would seem to make matters worse *vis-à-vis* the freedom in the choice of matrix.
(b) In the case of interest in Bayesian inference, where the target is a likelihood, the authors note that, once the matrix has been allowed to be state dependent, it may be chosen to coincide with the Fisher information matrix which defines a 'natural' metric to compute distances between parameterized probability measures. Endowed with such a metric, the space of parameterized distributions is a Riemannian manifold. In the case of the MALA, the resulting algorithm, the manifold MALA (MMALA), has a beautiful structure. The increment from the current state to the proposal includes a random component given by a Brownian motion on the Riemannian manifold and a deterministic component in the direction of steepest ascent in likelihood, where now 'steepest' is understood as measured by the natural metric for distributions rather than by the Euclidean distance between distribution parameter values.

The paper clearly bears out the advantages of the new algorithms, MMALA and Riemann manifold HMC (RMHMC), based on such an automatic, natural, geometric choice of the matrices and I have little doubt that it will lead to much future work. In connection with item (a) above there are obvious lines open to research. May the rather involved MMALA proposal be simplified (in ways different from that already

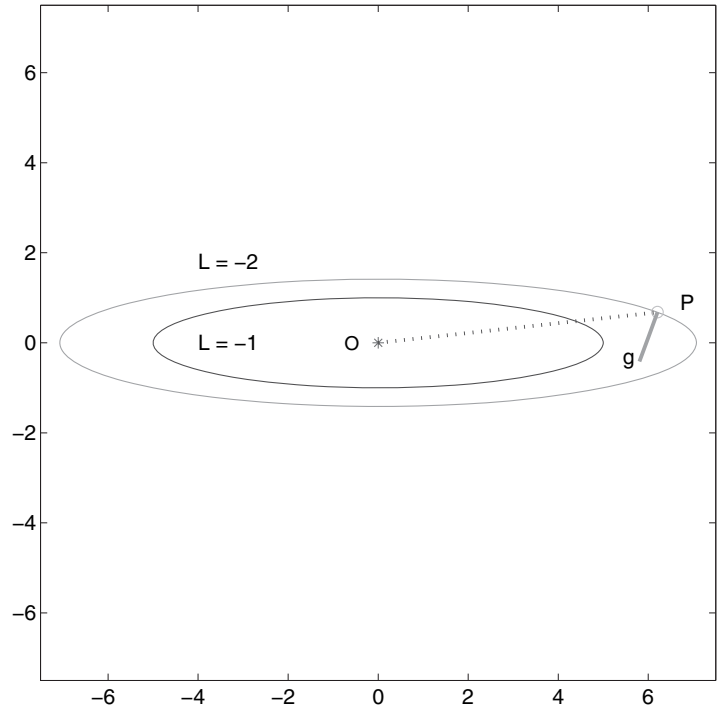

Fig. 11. The ellipses represent level sets of a real quadratic function L with a maximum at the point O; the direction of the standard Euclidean gradient g of L at P is not optimal when trying to reach O from P; the best possible direction PO is given by the product of the negative inverse Hessian and g; this is the gradient of L with respect to the metric defined by the negative Hessian; in probablity, L corresponds to the log-probablity density; in mechanics L is the negative potential and g provides the direction of the force at P; after choosing the mass matrix appropriately the acceleration will be aligned with PO; in optics the ellipses depict wave fronts in a non-isotropic medium, OP a ray emanating from O; the ray and the wave front are not orthogonal in the standard sense; Hamilton found the canonical equations guided by this analogy between optics and mechanics

considered in the paper)? For RMHMC algorithms, what is the most efficient way to integrate numerically the relevant canonical equations subject to preservation of geometric properties like reversibility and conservation of volume? Here the Hamiltonian function, although being separated in potential and kinetic components, has a variable mass matrix, a case that has not been addressed in the, by now large, body of work on geometric numerical integration as defined by Sanz-Serna (1997). The simple leapfrog or Verlet algorithm is the integrator of choice in conjunction with standard HMC methods; higher order integrators, although potentially advantageous (see the bounds in Beskos *et al.* (2010)), suffer from its more demanding computational cost per time step. For RMHMC methods where the simplest integrator is implicit, would it pay to move to higher order? In more general terms, the final success of MMALA and RMHMC methods will depend on addressing some implementation issues, particularly so when the state space has large dimensionality.

Item (b) will also attract much attention, if I am not mistaken. Are there alternative useful metrics beyond that defined by the Fisher information? The authors mention in this connection the observed Fisher information matrix or the empirical information matrix. The former, the negative Hessian of the log-probability, has the appeal of making sense not only in the context of Bayesian statistics but also for any target distribution and in fact may turn out to be useful in, say, sampling the canonical distribution in molecular simulations (I am currently experimenting with that possibility). Unfortunately the Hessian typically is not positive definite throughout the state space (for instance it is not in the illustrative example in Section 5.1), a difficulty that must be addressed. By the way, in the Bayesian context of the paper, the metric tensor equals the expected information matrix plus the negative Hessian of the log-prior: the first is necessarily positive definite; the second is not in general.

The last comments lead us to explore connections with well-known ideas from the field of optimization. There is of course a clear relationship between exploring a probability distribution and locating the maxima of the log-probability. The message of the paper is then akin to something that has been known for long in optimization: taking local steps in the direction of the current (standard) gradient is not the best way to reach the maximum of the objective function. The product of the inverse negative Hessian and the gradient provides a much better alternative as shown Fig. 11.

Something I have enjoyed when reading the paper by Girolami and Calderhead is the wide variety of ideas that contribute to shape the final algorithms, from Riemannian geometry to Bayesian statistics; from Hamiltonian dynamics to numerical geometric integration. For a non-statistician like me, it has been a privilege to study a piece of work that clearly demonstrates the inherent unity of all mathematical sciences. It gives me great pleasure to propose the vote of thanks.

Carl Edward Rasmussen (*University of Cambridge*)

I congratulate Professor Girolami and Dr Calderhead for this significant methodological development, which promises far reaching consequences for the practical application of Hamiltonian Monte Carlo (HMC) methods for inference in continuous models.

Since Neal (l993) introduced HMC methods (using the name hybrid MC) to the statistical community, and showed how they enabled the solution of otherwise intractable high dimensional inference problems, they have been adopted as a standard methodology in some fields such as machine learning. A reason for this is undoubtedly that implementation is straightforward. However, in practice, considerable experience with the method is required to make it run well. Specifically, step sizes (or equivalently fictitious masses) must be chosen for the dynamical simulation and trajectory lengths must be set to suppress random walks. Whereas an overall step size can be set by monitoring rejection rates, mixing will be slow if significantly different step sizes are appropriate for different variables in the system. Obtaining good estimates for these is often very difficult, and typically tedious and time consuming pilot runs are needed, even to obtain a reasonable diagonal mass matrix with a few distinct entries.

Various attempts have been made in the past to utilize the manifold structure of the parameter space, e.g. by Zlochin and Baram (2001), but today's paper is the first method which samples exactly from the desired invariant distribution. The method automatically provides a full mass matrix, as opposed to the hand-designed diagonal counterparts which have been used in the past. Experiments show convincingly the effectiveness of the approach.

A possible shortcoming of the paper relates to the treatment of random walks. Having provided a solution to the problem of the mass matrix (thus reducing the step size problem to fitting a single scalar, which can easily be adjusted by using the rejection rate), the final operational requirement is to determine the trajectory lengths. This is essential, to ensure that the method avoids slow random walks, which is perhaps the most compelling advantage of HMC methods. So, it would have been desirable if the authors had

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

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

$$= \mathbf{1} \wedge \frac{\pi(x^*)}{\pi(x_n)} \frac{s(x_n|x^*, r^*)}{s(x^*|x_n, r_n)}.$$

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 **f**	3 (0)	27 (17)	78 (69)	26 (24)	3 (0)	17 (6)	182 (50)
Average ESS **f**	6 (0)	102 (27)	404 (92)	112 (24)	6 (0)	51 (4)	531 (80)
Maximum ESS **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)	—	—	—	—
$10^3 \times \text{chol}(K)$	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 **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 **f** and $\boldsymbol{\theta}$ is offered by whitening the prior over **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 **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

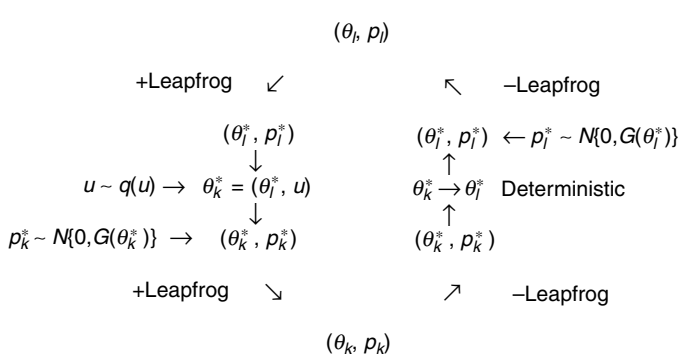

Fig. 12. Schematic diagram of the RJRMHMC 'jump' move from current state (θ_l, p_l) in model M_l to a proposed state (θ_k, p_k) in model M_k

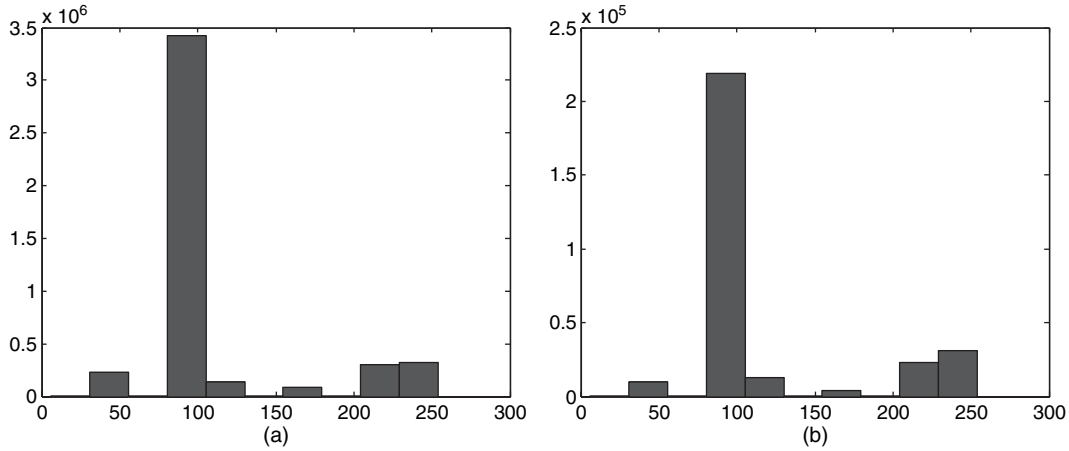

Fig. 13. Posterior model probabilities from (a) the RJRMHMC and (b) the RJMCMC algorithms

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

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

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

Table 13. Effective sample size of the proposed algorithms and Metropolis–Hastings algorithm with Gaussian proposal†

Algorithm	Metropolis–Hastings	MALA	MMALA	Simplified MMALA	Hamiltonian Monte Carlo	Riemann manifold Hamiltonian Monte Carlo
Minimum effective sample size	244 ± 24.5	36 ± 7.6	133 ± 42.6	100 ± 38.3	143 ± 18.3	26 ± 17

†The effective sample size is calculated from 3000 posterior samples after a burn-in period of 1000 samples. The low effective sample size for the Riemann manifold Hamiltonian Monte Carlo algorithm is due to the small step size required to achieve accurate integration of the Hamiltonian system.

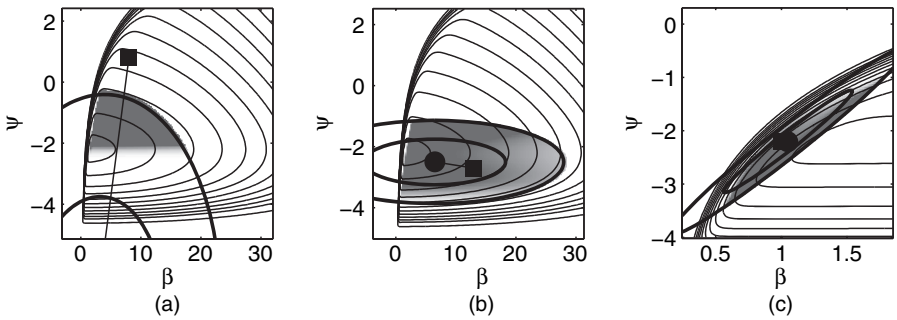

Fig. 14. Illustration of the point adaptive proposal mechanism of the simplified MMALA (■, current state; ●, mean of the proposal): the 90% and 50% contours of the Gaussian proposal are presented by the ellipses, and the shaded area is the acceptance rate for the underlying regions, with dark areas denoting a high acceptance rate

$$c_i = -\frac{\beta x_i}{2\sigma}\{\xi(\rho) + \xi(-\rho)\},$$
$$d_i = \frac{\beta x_i}{2\sigma}\xi(-\rho),$$
$$\rho = \beta x_i/\sigma,$$
$$\xi(\rho) = \mathcal{N}(\rho)/\Phi(\rho).$$

In Table 13 we compare the proposed Markov chain Monte Carlo algorithms with a componentwise adaptive Metropolis–Hastings algorithm in terms of the effective sample size. Fig. 14 also illustrates the problems associated with the skew posterior distribution and the position-dependent proposal mechanisms of the manifold Metropolis adjusted Langevin algorithm (MMALA) and simplified MMALA. From Fig. 14(c), we see that in 'steep' regions of the posterior the proposal distribution adapts to the curvature forcing the algorithm to make small steps. In contrast, in smoother regions the proposal allows for larger steps which can sometimes overshoot. This behaviour also leads to very low acceptance rates for large regions where the log-joint-likelihood is higher than the current state. This is illustrated in Figs 14(a) and 14(b) and is due to the acceptance ratio for non-symmetric proposal mechanisms.

Christian P. Robert (*Université Paris Dauphine, Ceremade and Centre de Recherche en Economie et Statistique, Paris*)

This paper is an interesting addition to recent Markov chain Monte Carlo (MCMC) literature and I am eager to see how the community will react to this potential addition to the MCMC toolbox. I am, however, wondering about the effect of the paper on MCMC practice. Indeed, although the dynamic on the level sets of

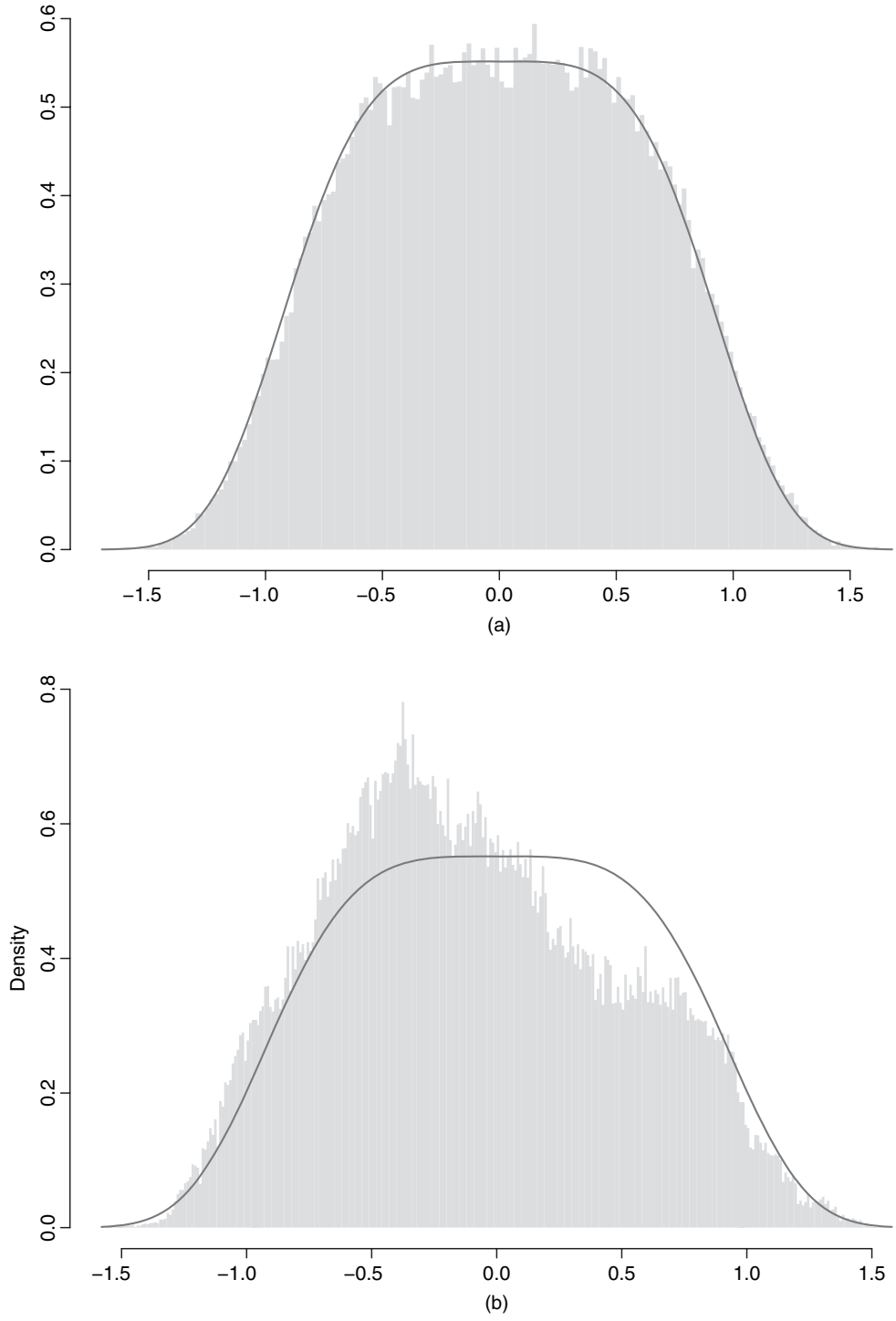

Fig. 15. Comparison of the fits of discretized Langevin diffusions to the target $f(x) \propto \exp(-x^4)$ when using a discretization step of (a) $\sigma^2 = 0.01$ and (b) $\sigma^2 = 0.001$, after $T = 10^7$ steps: this comparison illustrates the need for more time steps when using a smaller discretization step

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

is associated with Hamilton's equations, in that those moves preserve the potential $\mathcal{H}(\theta, \mathbf{p})$ and hence the target distribution at all times t, I argue that the transfer to the simulation side, i.e. the discretization part, is not necessarily useful, or at least that it does not need to be so painstaking in reproducing the continuous phenomenon.

In a continuous timeframe, the purpose of the auxiliary vector $\mathbf{p}$ is clearly to speed up the exploration of the posterior surface by taking advantage of the additional energy it provides. In the discrete time universe of simulation, however, the fact that the discretized (Euler) approximations to Hamilton's equations are not exact nor available in closed form does not present such a challenge in that approximations can be corrected by a Metropolis–Hastings step, provided of course that all terms in the Metropolis–Hastings ratio are available. However, the continuous time (physical or geometric) analogy at the core of the Hamiltonian may be unnecessarily costly when trying to carry a physical pattern in a discrete (algorithmic) time. MCMC algorithms are not set to work in continuous time and therefore the invariance and stability properties of the continuous time process that motivates the method do not carry over to the discretized version of the process. For one thing, the (continuous) time unit has no equivalent in discrete time. Therefore, the dynamics of the Hamiltonian do not tell us how long the discretized version should run, as illustrated in Fig. 15. As a result, convergence issues (of the MCMC algorithm) should not be impacted by inexact renderings of the continuous time process in discrete time. For instance, when considering the Langevin diffusion, the corresponding Langevin algorithm could equally use another scale η for the gradient than the one τ used for the noise, i.e.

$$y = x^t + \eta\nabla\pi(x) + \tau\varepsilon_t$$

rather than a strict Euler discretization where $\eta = \tau^2/2$. A few experiments run in Robert and Casella (1999), chapter 6, section 6.5, showed that using a different scale η could actually lead to improvements, even though we never pursued the matter any further.

Paul Fearnhead (*Lancaster University*)

I congratulate the authors for this stimulating paper. I can see it having a large influence within the statistics community, not only through the new Markov chain Monte Carlo algorithms that are introduced, but also through helping to promote Hamiltonian Monte Carlo (HMC) methods, and through the strong emphasis on how geometry can help us to design efficient Markov chain Monte Carlo algorithms.

I would like to comment in detail on the link between the HMC and Metropolis adjusted Langevin algorithm (MALA). Consider the simplest HMC algorithm of Section 3, with, for notational simplicity, $\mathbf{M} = \mathbf{I}$, the identity matrix.

Denote the current state $(\theta^{(n)}, \mathbf{p}^{(n)})$. Then one iteration of the HMC algorithm involves the following steps.

Step 1: set $\theta(0) = \theta^{(n)}$ and $\mathbf{p}(0) = \mathbf{p}^{(n)}$. For $j = 1, \ldots, m$, repeat the following steps (a)–(c).

(a) Set $\mathbf{p}(j - \frac{1}{2}) = \mathbf{p}(j-1) + \varepsilon\nabla\mathcal{L}\{\theta(j-1)\}/2$.
(b) Set $\theta(j) = \theta(j-1) + \varepsilon\,\mathbf{p}(j - \frac{1}{2})$.
(c) Set $\mathbf{p}(j) = \mathbf{p}(j - \frac{1}{2}) + \varepsilon\nabla\mathcal{L}\{\theta(j)\}/2$.

Step 2: set $\theta^{(n+1)} = \theta(m)$ and $\mathbf{p}^{(n*)} = \mathbf{p}(m)$ with probability

$$\min(1, \exp[-H\{\theta(m), \mathbf{p}(m)\} + H(\theta^{(n)}, \mathbf{p}^{(n)})]);$$

otherwise $\theta^{(n+1)} = \theta^{(n)}$ and $\mathbf{p}^{(n*)} = \mathbf{p}^{(n)}$.
Step 3: simulate $\mathbf{p}^{(n+1)}$ from $p(\mathbf{p})$.

Step 1 is just simulating the Hamiltonian dynamics by using the Stormer–Verlet integrator; this produces a new position $\theta(m)$ and momentum $\mathbf{p}(m)$ which are accepted or rejected in step 2. Finally step 3 involves updating the momentum.

As pointed out by the authors, $m = 1$ results in an MALA. The advantage of the HMC algorithm over the MALA is that with the MALA the momentum at each iteration of the Stormer–Verlet integrator is independent—which means that you will observe random-walk-type behaviour, with the possibility of the next move being in the opposite direction to the current move.

An alternative way of removing this behaviour is to change the update of the momentum to one where the new value is strongly dependent on the old one. So we could replace step 3 above by

Step 3: simulate $\mathbf{p}^{(n+1)}$ given $\mathbf{p}^{(n*)}$,

$$\mathbf{p}^{(n+1)} = (1-a)\mathbf{p}^{(n*)} + \sqrt{\{1-(1-a)^2\}}\mathbf{Z}_{n+1}.,$$

with a arbitrary but small. This sort of idea has been suggested before in the physics literature by Horowitz (1991). It should be straightforward to extend this type of algorithm to include the Riemann manifold ideas introduced in Section 4 and onwards.

If we set $a = \sigma^2\varepsilon/2$ and let $\varepsilon \to 0$, we find that the dynamics of this new algorithm converge to the solution of the following hypoelliptic stochastic differential equation:

$$\mathrm{d}\theta_t = p_t\,\mathrm{d}t,$$
$$\mathrm{d}p_t = \left\{\nabla\mathcal{L}(\theta_t) - \frac{\sigma^2}{2}p_t^2\right\}\mathrm{d}t + \sigma\,\mathrm{d}B_t.$$

This also gives a link with both the Hamiltonian and the Langevin dynamics, as $\sigma = 0$ gives the former, whereas $\sigma \to \infty$ gives the latter. This new algorithm corresponds to the in-between case of σ finite and non-zero.

A. Beskos (*University College London*) **and A. M. Stuart** (*University of Warwick, Coventry*)

Several applications, including inverse problems and problems related to diffusion processes, give rise to the issue of sampling a probability measure π defined via its density with respect to a Gaussian measure $\pi_0 = N(0, C)$:

$$\frac{\mathrm{d}\pi}{\mathrm{d}\pi_0}(\theta) \propto \exp\{-\Phi(\theta)\}. \tag{32}$$

Such problems are typically high dimensional, indeed often including non-parametric Bayesian estimation of functions from data. It is known that standard Markov chain Monte Carlo methods such as the Metropolis adjusted Langevin algorithm and hybrid Monte Carlo algorithm perform poorly in high dimensions as the time discretization of the underlying dynamics requires very small time steps to provide non-negligible acceptance probabilities: see Roberts and Rosenthal (2001) and Beskos *et al.* (2010a). However, for problems with the structure (32) it is now understood that carefully chosen proposals, which exploit the underlying Gaussianity, can allow time steps which do not scale poorly with dimension. It turns out that such ideas are closely related to those introduced in the paper as we now explain.

The dynamical systems underlying the Metropolis adjusted Langevin algorithm and hybrid Monte Carlo algorithm when the objective is to sample from density (32) are respectively, for positive definite matrix M,

$$\mathrm{d}\theta = -\tfrac{1}{2}M^{-1}\{C^{-1}\theta + \nabla\Phi(\theta)\} + \sqrt{M^{-1}}\,\mathrm{d}W \tag{33}$$

and

$$M\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} = C^{-1}\theta + \nabla\Phi(\theta) = 0. \tag{34}$$

Both these dynamics preserve π, in the second case provided that $\mathrm{d}\theta/\mathrm{d}t$ is distributed according to $N(0, M^{-1})$. Thus, appropriate time discretizations could provide tuned proposals. Critically, we found that appropriate choices of M combined with judicious time discretizations remove dimension dependence in the size of time steps used, hence improving decorrelation in the Markov chain Monte Carlo context. This idea is described in Beskos *et al.* (2008, 2010b): therein, it is shown that a desirable choice is $M = C^{-1}$ since it equilibrates the timescales in the dynamical systems which are generated by the Gaussian reference measure; a link is made to the idea of *preconditioning* in numerical analysis.

The choice $M = C^{-1}$ is precisely the one that the methods proposed in the paper would make in the case $\Phi = 0$. Herein, M is not constant but for $\Phi = 0$ things simplify: from a statistical perspective the likelihood is trivial in this case, so M is determined only by the prior π_0. Thus, in terms of choosing M the work in Beskos *et al.* (2008, 2010b) can be seen as following the idea in the paper, but based only on the prior. Importantly, attainment of a dimension-free time step in Beskos *et al.* (2008, 2010b) also required use of trapezoidal discretization schemes and splitting techniques for the underlying dynamics. Combining the ideas of Beskos *et al.* (2008, 2010b), which provide algorithms that are efficient in high dimensions, with the ideas in the paper which make use of the likelihood but may be difficult to deploy in high dimensions, seems an interesting direction for future research.

John Skilling (*Maximum Entropy Data Consultants, Kenmare*)

Let me start by complimenting Girolami and Calderhead on their smart and professional use of geometry. Probabilistic computation relies on proposal transitions, which can obviously be assisted by a sympathetic metric that adds power to our algorithms. The authors stress that this metric is arbitrary, and I agree.

But the popular metric is the Fisher–Rao metric, and this need not be sympathetic. Its metric is the curvature of the Kullback–Leibler (KL) '$p\log(p/q)$' formula, with $\sqrt{}$determinant as additive density. That is odd, because it is based on local differentials which do not add up in the usual linear way. That non-linearity has odd effects. For a start, it imposes a Dirichlet $\frac{1}{2}$ density on the underlying probability space. But a Dirichlet index should be a measure. Making it constant wrongly makes different digitizations have inconsistent densities.

An embedded manifold $\mathbf{p}(\theta)$ inherits its Fisher–Rao metric over θ from its Dirichlet $\frac{1}{2}$ surroundings. Suppose that the manifold is locally rough. The metric homes in on this roughness as the paths lengthen. Geodesics spend long arc lengths there and the local density increases. But that is wrong, because local roughness should have no macroscopic consequence.

Consider a periodic population with minimum at phase A and maximum at phase B, between which growth and decay have some given form (linear, logistic or whatever). The Fisher–Rao metric yields a density for A and B that is dominated with probability 1 by instantaneous growth or instantaneous decay. The (A,B) manifold *allows* sharp structure, which geometry then wrongly *insists* on. Something is deeply awry.

Returning to foundations, geometry needs a commutative scalar connection—to be called *distance*—between distributions **p** and **q**. Distance is to be found among the more general directed (from **p** to **q**) divergences. Symmetry of independence, for which the joint distribution of independent problems is required to be the ordinary direct product of the constituents, forces scalar divergence to be the standard KL formula. That is the foundation of information and entropy. Independence is necessary and sufficient for the KL formula.

However, the KL formula is non-commutative: **p** to **q** differs from **q** to **p**. The family of divergences has no commutative member. Hence *there is no general distance*! Yes, the KL formula is differentiable and its Hessian defines a Riemannian metric. But the corresponding geodesic distances are Hellinger, not KL, and disobey the vital symmetry of independence. There is macroscopic interference. Thus *Fisher–Rao geometry destroys its own foundation*.

Geometry is fine—even necessary—but it is empirical and not Fisher–Rao.

The following contributions were received in writing after the meeting.

Karim Anaya-Izquierdo (*The Open University, Milton Keynes*) **and Paul Marriott** (*University of Waterloo*)

The authors have done an excellent job of showing how to use the Riemannian structure of regular models to improve existing Markov chain Monte Carlo algorithms. Of course as Amari (1985) and others have shown (see Kass and Vos (1997) and references therein), there is much more to the geometry underlying statistical inference than the Riemannian metric tensor: indeed, as we have recently discussed in Anaya-Izquierdo *et al.* (2010), more to the geometry than the manifold structure itself. We wonder whether the authors have considered the underlying geometrical issues which occur when undertaking inferences on mixture models of various kinds. These models are no longer manifolds since they do not have a fixed dimension, they can have boundaries, and tangent spaces become tangent cones, all of which make local analysis more subtle. For mixture models the underlying geometrical object is typically a convex hull where natural moves in the space can follow geodesics defined by Amari's -1-connection. This holds for a general mixture as well as for the inferentially more tractable local mixture model; see Anaya-Izquierdo and Marriott (2007). Furthermore, for mixtures the Riemannian structure, even in simple cases such as two-component mixtures of exponentially distributed random variables, can fail to exist; see Li *et al.* (2009).

Cedric Archambeau and Guillaume Bouchard (*Xerox Research Centre Europe, Meylan*) (© Xerox)

This work addresses a fundamental problem in Markov chain Monte Carlo (MCMC) sampling. It provides a way of designing proposal densities without the need to rely on prior knowledge about the problem at hand. Simple inference problems where the entries of the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^D$ have very different scales or are highly correlated can lead to slow convergence. At present, there is no principled way to solve this problem. For example, one could resort to an *ad hoc* preconditioning as in Hamiltonian Monte Carlo (HMC) methods. Another approach is to introduce auxiliary variables to improve mixing rates, but again improvements can be fairly limited, as shown in the Bayesian treatment of logistic regression

by Holmes and Held (2005). Exploiting information geometrical concepts is an important step forward in the design of general purpose samplers. By using second-order information, the Riemann manifold HMC algorithm seems to provide the same level of improvement as Newton's method did compared with first-order gradient descent in unconstrained optimization.

The effective sample sizes reported by the authors were much larger than those obtained for Metropolis adjusted Langevin or HMC algorithms: typically a factor of 100 or more. Ignoring the computational overhead of Riemann manifold HMC sampling suggests that these numbers would be sufficient for the sampler to be a serious contender to current deterministic approximate inference algorithms, which are biased, but known to converge faster than conventional MCMC algorithms. A study of the bias–variance trade-off of Riemann manifold HMC methods, variational Bayesian inference (Attias, 1999) and expectation–propagation (Opper and Winther, 2000; Minka, 2001) with respect to their computational cost and their speed of convergence on a simple probabilistic model like Bayesian logistic regression would be interesting.

The main practical difficulty with Riemann manifold MCMC sampling is the $\mathcal{O}(D^3)$ computational cost associated with the inversion of the metric tensor $\mathbf{G}(\boldsymbol{\theta})$ and to a lesser extent to the computation of the Christoffel symbols. For high dimensional $\boldsymbol{\theta}$, it is crucial to exploit the special structure of $\mathbf{G}(\boldsymbol{\theta})$ when possible. More generally, we would have to approximate $\mathbf{G}^{-1}(\boldsymbol{\theta})$ and quantify what would be lost as a function of the step size and the curvature of the manifold. This is a challenging line of research that has been very little explored in recent years but could have a significant influence in many application domains of computational statistics, non-linear optimization and statistical machine learning.

Magali Beffy (*Institut National de la Statistique et des Etudes Economiques, Paris*) **and Christian P. Robert** (*Université Paris Dauphine, Ceremade and Centre de Recherche en Economie et Statistique, Paris*)

The paper gives a very clear geometric motivation for the use of a Hamiltonian representation. As such, it suggests an immediate generalization by extending the Hamiltonian dynamic to a more general dynamic on the level sets

$$\mathcal{H}(\theta, \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k) = -\mathcal{L}(\theta) + \tfrac{1}{2}\log\{(2\pi)^D|\mathbf{G}_1(\theta)|\} + \tfrac{1}{2}\mathbf{p}_1^{\mathrm{T}}\mathbf{G}_1(\theta)^{-1}\mathbf{p}_1 + \tfrac{1}{2}\log\{(2\pi)^D|\mathbf{G}_2(\theta)|\} + \tfrac{1}{2}\mathbf{p}_1^{\mathrm{T}}\mathbf{G}_2(\theta)^{-1}\mathbf{p}_2 + \ldots + \tfrac{1}{2}\log\{(2\pi)^D|\mathbf{G}_k(\theta)|\} + \tfrac{1}{2}\mathbf{p}_1^{\mathrm{T}}\mathbf{G}_k(\theta)^{-1}\mathbf{p}_k + \ldots$$

where the $\mathbf{p}_j$s are auxiliary vectors of the same dimension as θ and the $\mathbf{G}_j(\theta)$s are symmetric matrices. This function is then associated with the partial differential equations

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \frac{\partial}{\partial p_{ij}}\mathcal{H}(\theta, \mathbf{p}_1, \ldots, \mathbf{p}_k) = \{G_j(\theta)^{-1}\mathbf{p}_j\},$$
$$\frac{\mathrm{d}p_{ij}}{\mathrm{d}t} = -\frac{\partial}{\partial \theta_i}\mathcal{H}_j(\theta, \mathbf{p}_1, \ldots, \mathbf{p}_k)$$

in that those moves preserve the potential $\mathcal{H}(\theta, \mathbf{p}_1, \ldots, \mathbf{p}_k)$ and hence the target distribution at all times t. This generalization would allow for using a range of information matrices $\mathbf{G}_j(\theta)$ in parallel. The corresponding Riemann Hamiltonian Monte Carlo implementation is to pick one of the indices j at random and to follow the same moves as in the paper, given the separation between the different energies.

Anindya Bhadra (*Texas A&M University, College Station*)

The authors are to be congratulated for a novel design of an efficient and automatic choice of the preconditioning matrix for the Metropolis adjusted Langevin algorithm (MALA) or mass matrix for Hamiltonian Monte Carlo (HMC) schemes. The clever use of local curvature information results in possible improvements in the relative speed of convergence to the high dimensional target distribution, as demonstrated by the authors' various illustrative examples.

The full manifold MALA (MMALA) and Riemann manifold HMC schemes described by the authors require

(a) evaluations of the partial derivatives up to third order for the log-likelihood function and
(b) inversion of the position-specific metric tensor of the Riemann manifold formed by the parameter space.

In the general case, considering the absence of nice analytical properties inducing sparsity in the covariance matrix etc., these two steps are computationally intensive (as pointed out by the authors) and in many of the examples the authors are forced to resort to a simplified version of the MMALA scheme.

One interesting application would be to use the authors' approach for a Markov chain Monte Carlo (MCMC) based stochastic optimization scheme like simulated annealing. At the early stages of the algorithm, with high temperature, the use of local curvature information in the MCMC proposal should result in a high acceptance rate and the search should quickly reach a neighbourhood of the global maxima (see Fig. 1). However, once this has been achieved, the use of local curvature information is unlikely to have much further benefit at the expense of the substantial computational burden imposed by steps (a) and (b) mentioned above and a computationally simpler MCMC algorithm that does not use local curvature information (or global MALA or HMC algorithm) can be used once the temperature has cooled sufficiently to do a local search.

Numerical schemes for computing derivatives that are needed by the authors are often unstable. Ionides *et al.* (2006) proposed iterated filtering, a derivative-free maximum-likelihood-based inference technique for partially observed Markovian state space models (an example of such a model is presented by the authors in Section 8) that has been succesfully applied in many scientific applications (e.g. Laneri *et al.* (2010), Bretó *et al.* (2009) and He *et al.* (2010)). From the computational perspective, a favourable comparison of the iterated filtering with particle MCMC technique presented in Andrieu *et al.* (2010) is presented in Bhadra (2010) and iterated filtering presents a viable maximum likelihood alternative to Bayesian inference in many difficult situations.

Luke Bornn and Julien Cornebise (*University of British Columbia, Vancouver*)

We show how the proposed Riemann manifold Hamiltonian Monte Carlo (RMHMC) method can be particularly useful in the case of strong geometric features such as ridges commonly occurring in non-identifiable models. Although it has been suggested to use tempering or adaptive methods to handle these ridges (e.g. Neal (2001) and Haario *et al.* (2001)), they remain a celebrated challenge for new Monte Carlo methods (Cornuet *et al.*, 2009). RMHMC sampling, by exploiting the geometry of the surface to help to make intelligent moves along the ridge, is a brilliant advance for non-identifiability sampling issues.

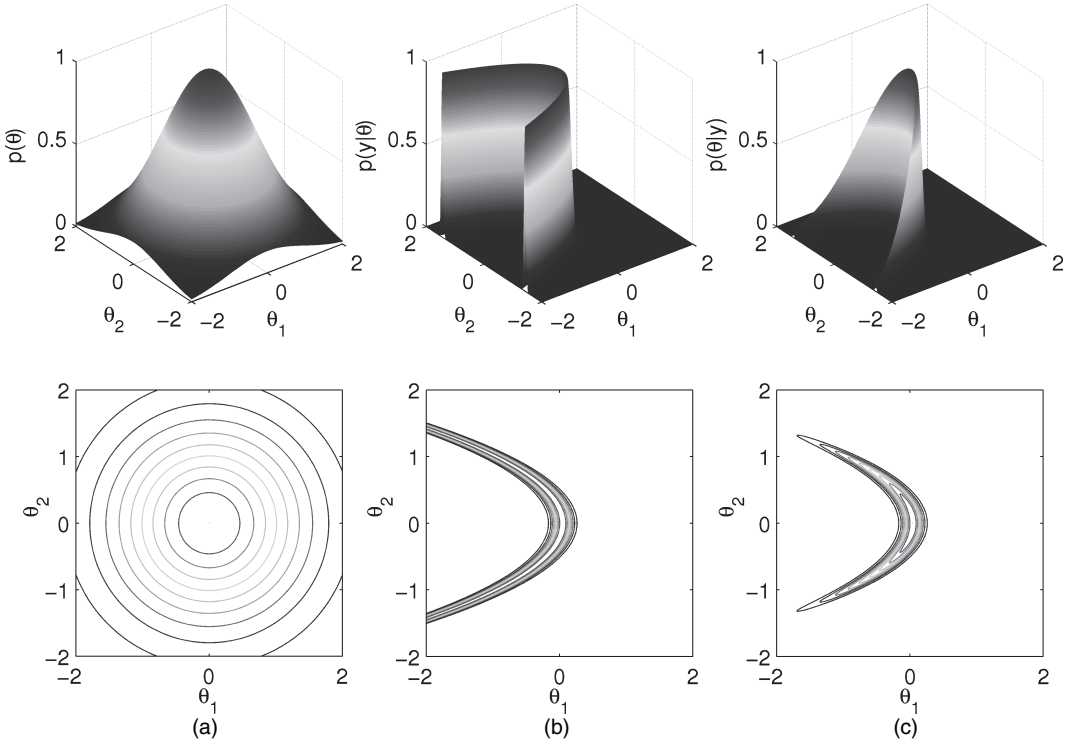

Fig. 16. (a) Prior, (b) likelihood and (c) posterior for the warped bivariate Gaussian distribution with $n = 100$ values generated from the likelihood with parameter settings $\sigma_\theta = \sigma_y = 1$: as the sample size increases and the prior becomes more diffuse, the posterior becomes less identifiable and the ridge in the posterior becomes stronger

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

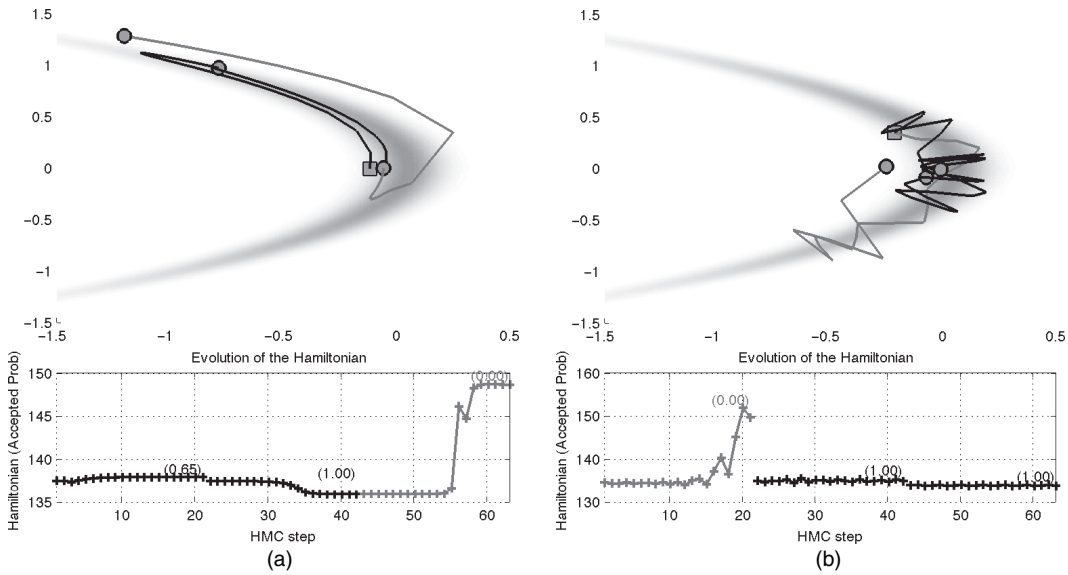

Fig. 17. Three typical consecutive trajectories of 20 leapfrog steps each, with step size 0.1, for the RMHMC and HMC algorithm, chosen to highlight two acceptances (black) and one rejection (grey), representative of the approximately 65% acceptance ratio for both the HMC and RMHMC algorithms: we see 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

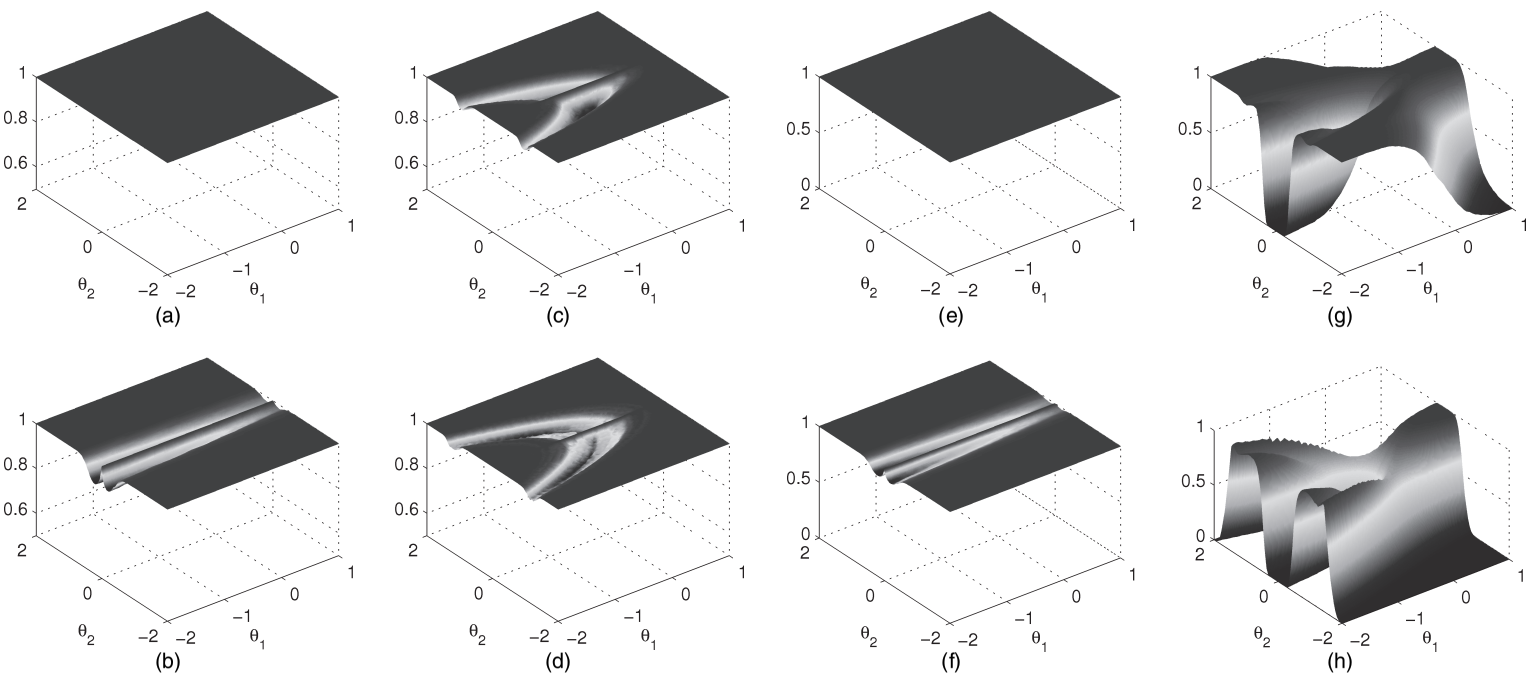

Fig. 18. Probability of a single iteration of the generalized leapfrog (a)–(d) admitting a fixed point $\mathbf{p}(\varepsilon/2)$ and (e)–(h) admitting a fixed point and having a sufficiently small derivative at the fixed point for the fixed point iterations to converge (all graphs are plotted as a function of starting point $\theta(0)$ and are shown for two step sizes $\varepsilon \in \{0.1, 1.0\}$ and two prior distribution standard deviations $\sigma_\theta \in \{0.5, 1.0\}$, i.e. varying levels of identifiability; the region of stability for the fixed point iterations becomes much smaller as the step size increases and as identifiability decreases, even creating regions with null probability of convergence): (a) $\varepsilon = 0.10$, $\sigma_\theta = 0.50$; (b) $\varepsilon = 0.10$, $\sigma_\theta = 1.00$; (c) $\varepsilon = 1.00$, $\sigma_\theta = 0.50$; (d) $\varepsilon = 1.00$, $\sigma_\theta = 1.00$; (e) $\varepsilon = 0.10$, $\sigma_\theta = 0.50$; (f) $\varepsilon = 0.10$, $\sigma_\theta = 1.00$; (g) $\varepsilon = 1.00$, $\sigma_\theta = 0.50$; (h) $\varepsilon = 1.00$, $\sigma_\theta = 1.00$

much practical insight when choosing the step size—possibly adapting it 'on the fly', when RMHMC sampling already provides a clever way to devise the direction of the moves adaptively.

David A. Campbell (*Simon Fraser University, Surrey*)
The authors propose an innovative use of posterior geometry that is poised to become a standard tool for Markov chain Monte Carlo (MCMC) practitioners.

In high dimensional cases, the metric tensor may be too computationally intensive to compute, as in the example of the log-Gaussian Cox process where 5000 samples took 90 h of computation. Since computing the metric tensor is the rate limiting step, several options exist for balancing the competing desires of improved computational load and high sampling efficiency.

Block updating may be worthwhile as a trade-off between computing a full metric tensor and ignoring the local topology. Randomly grouping parameters into blocks where between-blocks are assumed independent produces a sparse matrix and keeping the between-block matrix tensor fixed at the previous iterations calculation maintains topological information, albeit outdated. Sparsity leads to computational simplicity at the expense of ignoring structure and losing efficiency. Alternatively whitening transformations to decorrelate the parameters (Morris and Carroll, 2006) might be computationally necessary.

With dynamic systems, the geometry of the posterior surface may exacerbate mixing and hinder convergence. Take, for example, a two-parameter version of the FitzHugh–Nagumo system, where the data are a noiseless solution to the ordinary differential equation model as explored in Fig. 2 of Ramsay *et al.* (2007). The likelihood surface for this model includes local maxima, and ripples. As the authors point out, population-based methods are necessary for wider exploration of the posterior and single-chain MCMC algorithms cannot explore this posterior effectively. However, Riemann manifold Hamiltonian Monte Carlo (RMHMC) algorithms may encounter a new kind of issue in that, when the current parameters lie in a concentric ripple of local posterior maxima, RMHMC sampling encourages new samples to be selected in the direction locally tangent to the ripple. Assuming that the concentric ripples are of equal posterior height and sufficiently close together that they might be visited stepping just a little too far in the leapfrog step, then, new samples are more likely to jump to wider concentric ripples than they are to narrower concentric ripples. Clearly this is more of a concern with manifold Metropolis adjusted Langevin algorithms which genuinely use a local linear correlation assumption on the posterior topology compared with RMHMC algorithms where the metric tensor is assumed locally linear over a smaller neighbourhood. Although this does sound pathological, challenging posterior topologies including concentric ripples are common in models for oscillatory dynamics (Calderhead and Girolami, 2009).

Jiguo Cao (*Simon Fraser University, Burnaby*) **and Liangliang Wang** (*University of British Columbia, Vancouver*)
We thank Girolami and Calderhead for their significant contribution to Monte Carlo methods. The Metropolis–Hastings algorithm is very popular for sampling from the target probability density. However, as pointed out by Girolami and Calderhead, this algorithm is not efficient when the parameter space is of high dimension. They use the Riemann geometry of the parameter space to adapt the local structure, which greatly increases the sampling efficiency. In particular they develop a MATLAB computing package, which allows us to use this method directly. We find this computing package well organized and well documented.

We are particularly interested in their application on estimating parameters in non-linear differential equation models from noisy data. Differential equations have been widely used to model complex dynamic models in biology, engineering, medicine and many other areas. It is of great interest to identify the parameter values in these differential equations. A large challenge is that the parameter space has many local maxima. In addition, most differential equations do not have analytic solutions, and it is computationally expensive to solve differential equations numerically with thousands of candidates of parameter values. Therefore, it is very important to increase the sampling efficiency by using the local geometry information.

Girolami and Calderhead compared several sampling schemes and showed that the manifold Metropolis adjusted Langevin algorithm method has the best performance in their example. However, all these sampling chains are initialized on the true parameter values. In real applications, the true parameter values are never known. So it will be very interesting to compare these sampling schemes in their sampling efficiency when the sampling chains start with some random parameter values. From our point of view, this issue is more important than the efficiency of the sampling chains that start from the true mode.

Another important issue is the identifiability of the parameters in the differential equations from the noisy data. In the Fitzhugh–Nagumo example in Section 10.1, both variables (V and R) are observed, and

the simulated data are very dense. In this case, all three parameters are probably identifiable. However, when only V is observed and no data are available for R, will all these three parameters be identifiable? If yes, what is the minimum number of observations for V such that all three parameters are identifiable? Can we have some statistical quantity to tell us the identifiability of the parameters from the sample chains?

Siu A. Chin (*Texas A&M University, College Station*)
The authors have proposed an excellent idea for taking advantage of the metric structure of the target distribution to improve its sampling. This can have impact not only in statistical inferences but also in variational Monte Carlo calculations across multiple subfields of physics. Here are some discussion points.

(a) It seems that the corresponding Langevin algorithm bypasses the added difficulty of integrating the non-separable equations (16)–(18) in Hamiltonian Monte Carlo methods. It would be of interest to know their relative efficiency, especially when the Langevin algorithm can be improved to higher orders; see below.
(b) The Langevin algorithm that was used by the authors, the equation after equation (10), is only a first-order algorithm. This is useful as a proposed move in Metropolis–Hastings steps, but second- and fourth-order Langevin (Forbert and Chin, 2000) algorithms can be so accurate that the additional acceptance–rejection step may not be necessary. This is a viable alternative when there is no analytical expression for the proposed move to be accepted or rejected via Metropolis–Hastings steps. Moreover, in the fourth-order Langevin algorithm, a precondition matrix arises naturally.
(c) As noted by the authors, the basic bottleneck of the proposed idea is the general lack of an analytical Fisher metric for most target distributions. In physical applications, these target distributions are complex trial ground state wave functions. It is very difficult to compute their derivatives with respect to the variational parameters, not to mention computing their expectation values. Is there a way of estimating the required metric roughly, 'on the fly'?

A. C. C. Coolen (*King's College London*)
The paper by Girolami and Calderhead shows how information geometric ideas can be used to improve algorithms for sampling random variables from a given probability density. It answers the question of how to choose the noise covariance matrix in Langevin dynamics sampling, or the momenta covariance matrix in Hamiltonian Monte Carlo algorithms. The final result is elegant and effective, and provides protocols in which the exploration of candidate states takes into account curvature properties, with Fisher's information matrix as metric tensor.

The application to the Metropolis adjusted Langevin algorithm is conceptually straightforward, because the Langevin equation itself already ensures evolution to the desired invariant measure. The situation with Hamiltonian Monte Carlo dynamics seems different. Only for very few physical systems can it be proven that over time Hamilton's equations will sample uniformly all states with a specified energy. The connection between dynamics and uniform sampling has the status of a hypothesis, which is difficult to prove in practice. In statistical mechanics this is not an issue. There we usually study systems with many variables (of the order of 10^{24}) and tend to measure observables involving many of these, giving an extra layer of averaging that irons out weak inhomogeneous sampling. In this paper, however, we do not have 10^{24} degrees of freedom. It would be interesting to find out, via synthetic models, whether this could cause problems. The simplest scenario would be that of integrable Hamiltonians, with conserved quantities beyond the energy, where uniform sampling will certainly not occur. One would hope that the periodic randomization of momenta in the (Riemann manifold) Hamiltonian Monte Carlo algorithm will compensate for the inadequacy of Hamiltonian dynamics in such cases; is this true, in principle and in practice?

Second, the authors mention the possible limitations of their proposal in terms of computational resources. With many degrees of freedom and/or a non-trivial measure to sample from, the repeated calculation and inversion of Fisher matrices can become prohibitively painful. In that context it could be interesting to study the effect of canonical transformations of co-ordinates and momenta, which are known to leave invariant all the desirable properties of Hamilton's equations (structure, time reversibility, energy conservation and volume preservation). The sampling trajectory would by definition not be affected by such transformations; the dynamics remain the same. However, could it perhaps be that there is a (possibly problem-specific) canonical transformation for which Hamilton's equations take a form that is less demanding computationally?

Julien Cornebise (*University of British Columbia, Vancouver*) **and Gareth Peters** (*University of New South Wales, Sydney*)

The utility of Riemann manifold Metropolis adjusted Langevin (RMMALA) and Riemann manifold Hamiltonian Monte Carlo (RMHMC) methodology is its ability to adapt Markov chain proposals to the current state. Many references design adaptive Monte Carlo (MC) algorithms to learn efficient Markov chain Monte Carlo (MCMC) proposals, such as controlled MCMC methods (Haario *et al.*, 2001) which utilize a historically estimated global covariance for a random-walk Metropolis–Hastings algorithm. Similarly, Atchadé (2006) devised global adaptation in MALAs. Surveys are provided in Atchadé *et al.* (2009), Andrieu and Thoms (2008) and Roberts and Rosenthal (2009).

When the proposal remains essentially unchanged regardless of the current Markov chain state, performance may be poor if the shape of the target distribution varies widely over the parameter space. A typical illustration involves the 'banana-shaped' warped Gaussian, ironically originally utilized to illustrate the strength of an early adaptive MC method (Haario *et al.* (1999), Fig. 1). However, Haario *et al.* (2001), appendix A, showed that this original algorithm could exhibit strong bias, perhaps connected to requirements of 'diminishing adaptation' as studied in Andrieu and Moulines (2006). Recent locally adaptive algorithms satisfy this condition; for example *state-dependent proposal scalings* (Rosenthal (2010), section 3.4) fit a parametric family to the covariance as a function of the state, or the parameterized parameter space approach of *regional adaptive Metropolis algorithms* (Roberts and Rosenthal (2009), section 5).

Riemannian approaches provide strong rationale for parameterizing the proposal covariance as a function of the state—without learning, when the Fisher information matrix (FIM) (or observed FIM) can be computed or estimated. With unknown FIM, or to learn the optimal step size, it would be interesting to combine Riemann MC methods with adaption. A first step could involve a simplistic Riemann-inspired algorithm such as a centred random-walk Metropolis–Hastings algorithm via the (observed) FIM as the proposal covariance (as used in section 4.3.1 of Marin and Robert (2007))—equivalent to one step of the RMMALA without drift.

An additional use of Riemann MC methods could be within the MCMC step of particle MCMC (Andrieu *et al.*, 2010), where adaption was highly advantageous in the adaptive particle MCMC algorithms of Peters *et al.* (2010).

Another interesting extension involves considering the stochastic approximation alternative approach, based on a curvature updating criterion of Okabayashi and Geyer (2010), equation (10), for an adaptive line search. This was proposed as an alternative to the MCMC–maximum likelihood estimation of Geyer (1991) for complex dependence structures in exponential family models. In particular comparing properties of this curvature-based condition on the basis of local gradient information with adaptive RMMALA and RMHMC versions of the MCMC–maximum likelihood estimation algorithms would be instructive.

Additionally, one may consider how to extend Riemannian MC to trans-dimensional MC methods such as the reversible jump (Richardson and Green, 1997), for which adaptive extensions are rare (Green and Hastie (2009), section 4.2). How might a geometric approach be extended to explore disjoint unions of model subspaces efficiently as in Nevat *et al.* (2009)?

Finally, could such geometric tools be utilized to design the distance metric in approximate Bayesian computation (Beaumont *et al.*, 2009)?

D. R. Cox (*Nuffield College, Oxford*)

It is a pleasure to congratulate the authors on an impressive paper and on the lucid account of it given at the meeting. A long tradition of Research Section discussions, sadly in apparent decline, is the inclusion of contributions of dubious or at best borderline relevance. In that spirit the following remarks concern directly the example of the doubly stochastic point process but have broader implications. Are any of the difficulties associated with the model connected with poor fit? Why were the parameters of the Ornstein–Uhlenbeck process not estimated from first principles? Is the process really isotropic? More generally the essential idea of Markov chain Monte Carlo sampling is beautiful and has been developed over the years in a striking and imaginative way, but for making sense of empirical data it is a black box. The aspects of the data that drive the conclusions are hidden away, which makes critical assessment of the conclusions difficult. An argument that it is all taken care of automatically is in some contexts at least unconvincing. Is it feasible to develop adjuncts that would deal with these concerns?

David Draper (*University of California, Santa Cruz*)

I have three comments on this stimulating and useful paper.

(a) The authors have given us an embarrassment of riches: two new Markov chain Monte Carlo methods (the manifold Metropolis adjusted Langevin algorithm (MMALA) and the Riemann manifold Hamiltonian Monte Carlo (RMHMC) algorithm), both of which outperform existing approaches (the MALA and HMC algorithm) that are themselves substantial improvements on the old-fashioned-looking Metropolis–Hastings algorithm. From a user's viewpoint this is almost a surfeit, because the new techniques are so recent that we do not yet have really solid advice on which of the two will be better on *our* next problem: each of the RMHMC algorithm and simplified MMALA were the clear winners in two of the four examples in the paper. The authors note in Section 10 that '... the MMALA is perhaps particularly suited to settings in which there is a non-constant metric tensor which is expensive to compute', and this is the beginning of a kind of user's manual for the two new methods; can the authors at present say more about how to partition problems space into the regions {neither, only MMALA, only RMHMC, both} with respect to providing substantial Monte Carlo acceleration relative to other approaches, or is it too early to write the user's manual now, until more experience has accumulated?

(b) In Section 11 of the paper, the sentence 'The issue of the $\mathcal{O}(N^3)$ scaling is something which deserves further consideration' is an understatement! (This is relevant to defining the region {neither MMALA nor RMHMC works well} mentioned above.)

(c) Fig. 5 shows how the auto-correlations of the output time series for sampled quantities are driven by the new methods from large positive values down near 0. I wonder whether even more is possible: with high posterior correlations in the target density, one can imagine samplers that not only move smartly up and down the ridges induced by the correlations but do so by jumping back and forth across the posterior mode to induce *negative* auto-correlations. A student of mine and I (Liu, 2003; Draper and Liu, 2006) explored this several years ago with ordinary Metropolis–Hastings sampling under the name *mirror jump MCMC* sampling, and others (e.g. Peter Green, personal communication) have also thought along similar lines. Liu and I found that this succeeds remarkably well in simple problems, but it has not yet (to my knowledge) been tried in more complicated settings. Do the authors see any scope for ideas of this type to be folded into their approach, thereby achieving even bigger Monte Carlo accelerations?

Ian Dryden (*University of South Carolina, Columbia*)

This very interesting paper makes a strong case for using the Riemannian manifold structure in Markov chain Monte Carlo simulations for a variety of complex modelling scenarios.

The illustrative example Section 5.1 is not too difficult as N is quite large. Both methods would work adequately well after the burn-in period. However, more challenging would be the same model where more posterior variability is present, i.e. when N is small. Langevin diffusion in the manifold involves a large positive drift when very close to the boundary $\sigma = 0$, so one would need to be careful not to be thrown wildly away from the boundary with an Euler or other approximate scheme. Also, one needs to avoid jumping into negative σ territory. Such an issue occurs when simulating Brownian motion and Ornstein–Uhlenbeck processes on the sphere and complex projective space for particular choices of co-ordinates (Ball *et al.*, 2008). This can be avoided by using smaller time increments of course, or even perhaps some form of retrospective exact sampling in some applications (Beskos *et al.*, 2006, Ball *et al.*, 2006). However, the danger with accurately simulating near the boundary is that the step sizes may become extremely small, which could slow down the algorithm too much. A compromise is probably better.

The choice of metric is important, as mentioned by the authors. This is a common issue for practical analysis on all manifolds. One possibility is to let the algorithm inform the appropriate choice of metric, e.g. by running chains in parallel with different metrics and then switching to the metric which shows the most desirable features. Other sorts of adaptive methods could also be used, e.g. using a family of metrics indexed by parameter, and then adapting the choice of parameter according to some criterion. See Dryden *et al.* (2010) for a similar data-based choice of metric in the space of covariance matrices.

Shinto Eguchi (*Institute of Statistical Mathematics, Tokyo*)

This fascinating paper opens a new aspect on the connection between statistics and geometry. The differential geometric approach established intrinsic understandings for the structure of statistical inference such as sufficiency, consistency, Fisher information and second-order efficiency as discussed in Efron (1975), Dawid (1975) and Amari (1982). A wide class of second-order efficient estimators was shown in Eguchi (1983). However, these asymptotic considerations are significant only when the statistical model is rigorously correct. In fact these properties are fragile under model uncertainty, for example, caused by missing

data and other incomplete-data mechanisms, in which the statistical model is not perfectly confirmed; see Copas and Eguchi (2005) for a discussion on the tubular neighbourhood of the model.

In contrast, the simulation situation that the authors discuss permits us to obtain perfectly random sampling, in which the geometric consideration is much more directly persuasive than the general discussion in statistical inference. If any link with the Kullback–Leibler divergence is found, then the geometric consideration is applicable to the dualistic Riemannian geometry; see Amari and Nagaoka (2000) for a mathematical formulation. There are possibilities for solving the simulation time problem in a huge scale of computation for molecular dynamics; see Lelievre *et al.* (2008) for a stochastic framework. At least, this approach is directly connected to adaptive sampling in molecular dynamics. Also the approach could be related to the optimal transportation problem in recent progress in geometry, which is motivated by the positive solution for the Poincaré conjecture; see Lott and Villani (2009).

Andrew Gelman (*Columbia University, New York*)

I shall comment on this paper in my role as applied statistician and consumer of Bayesian computation. In the last few years, my colleagues and I have felt the need to fit predictive survey responses given multiple discrete predictors, e.g. estimating voting given ethnicity and income within each of the 50 states, or estimating public opinion about gay marriage given age, sex, ethnicity, education and state. We would like to be able to fit such models with 10 or more predictors—e.g. religion, religious attendance, marital status and urban–rural–suburban residence in addition to the factors mentioned above.

There are (at least) three reasons for fitting a model with many predictive factors and potentially a huge number of interactions among them.

(a) Deep interactions can be of substantive interest. For example, Gelman *et al.* (2009) discuss the importance of interactions between income, religion, religious attendance and state in understanding how people vote.
(b) Deep interactions can increase predictive power. For example Gelman and Ghitza (2010) show how the relationship between voter turnout and the combination of sex, ethnicity, education and state has systematic patterns that would not be captured by main effects or even two-way interactions.
(c) Deep interactions can help to correct for sampling problems. Non-response rates in opinion polls continue to rise, and this puts a premium on post-sampling adjustments. We can adjust for known differences between sampling and population by using post-stratification, but to do so we need reasonable estimates of the average survey response within narrow slices of the population (Gelman, 2007).

Our key difficulty—familiar in applied statistics but not always so clear in discussions of statistical computation—is that, although we have an idea of the sort of model we would like to fit, we are unclear on the details. Thus, our computational task is not merely to fit a single model but to try out many different possibilities. My colleagues and I need computational tools that are

(i) able to work with moderately large data sets (aggregations of surveys with total sample size in the hundreds of thousands),
(ii) able to handle complicated models with tens of thousands of latent parameters,
(iii) sufficiently flexible to fit models that we have not yet thought of and
(iv) sufficiently fast that we can fit model after model.

We all know by now that hierarchical Bayesian methods are a good way of estimating large numbers of parameters. I am excited about this paper, and others like it, because the tools therein promise to satisfy conditions. (i)–(iv) above.

Andrew Golightly and Richard J. Boys (*Newcastle University*)

We congratulate the authors for an interesting paper and an important contribution to Markov chain Monte Carlo methodology. The two algorithms they present, the manifold Metropolis adjusted Langevin algorithm (MMALA) and Riemann manifold Hamiltonian Monte Carlo (RMHMC) algorithm, remove the need to tune Metropolis–Hastings steps, and this can be very helpful when sampling from high dimensional target densities with strong correlations. Posterior distributions with high correlations can often occur in state space models; see, for example, Wilkinson and Golightly (2010) and Henderson *et al.* (2010). We therefore focus our discussion on the utility of the MMALA and RMHMC algorithm for state space models.

To fix notation consider a hidden Markov state process $\{X_n, n \geqslant 1\}$ with transition density parameterized by θ and observed indirectly through $\{Y_n, n \geqslant 1\}$. Suppose that we have observations on $\{Y_n\}$ which, for

simplicity, are conditionally independent given $\{X_n\}$ and the density associated with $Y|X$ is also parameterized by θ. The (assumed intractable) posterior density $p(\theta, \mathbf{x}|\mathbf{y})$ can be sampled by alternating between draws from $\theta|\mathbf{x}, \mathbf{y}$ and $\mathbf{x}|\theta, \mathbf{y}$. This two-step blocking approach is used successfully by the authors for the stochastic volatility model via both the MMALA and the RMHMC algorithm. Another possibility is to sample $p(\mathbf{x}|\mathbf{y}, \theta)$ via a particle independent Metropolis–Hastings scheme (Andrieu *et al.*, 2010). Both strategies are likely to perform well provided that $\theta|\mathbf{y}$ and $\mathbf{x}|\mathbf{y}$ are not highly correlated. In such scenarios a joint update can alleviate the problem. It is natural in this case to consider a proposal of the form $q_1(\theta^*|\theta)\, q_2(\mathbf{x}^*|\theta^*, \mathbf{y})$ where θ is the current value of the chain. The particle marginal Metropolis–Hastings algorithm (Andrieu *et al.*, 2010) sidesteps the issue of building an efficient proposal density $q_2(\mathbf{x}^*|\theta^*, \mathbf{y})$ by using a particle filter targeting the intractable density associated with $\mathbf{x}|\theta, \mathbf{y}$ and is simple to implement, requiring only the ability to simulate from the transition density associated with $\{X_n\}$. It therefore seems appealing to use the MMALA or RMHMC algorithm in the construction of $q_1(\theta^*|\theta)$. To construct the metric tensor $G(\theta)$, we require a closed form expression for the observed data likelihood $p(\mathbf{y}|\theta)$. This will rarely be available but particle filters do provide a numerical approximation to $p(\mathbf{y}|\theta)$ which, in turn, may be used to approximate $G(\theta)$ via the observed information obtained from numerical derivatives. However, this strategy may be computationally prohibitive in problems of any realistic size. Clearly further work will be needed to see how this scales with the dimension of the parameter θ.

Jim Griffin (*University of Kent, Canterbury*)

I congratulate the authors on an excellent paper with wide practical applications to Bayesian statistics. The methods described use gradient information from the unnormalized posterior density to improve mixing of the sampler. Methods for jointly updating parameters from highly correlated posterior distributions have received substantial attention recently. One promising approach is adaptive independent Metropolis–Hastings methods (see for example Giordani and Kohn (2010) and Roberts and Rosenthal (2009)) where an approximation to the posterior distribution is constructed using previous values generated by the sampler. It would interesting to know how these methods compare with those described in the paper. Gradient information may lead to better performance but, as the paper clearly demonstrates, the computational time involved in its calculation can be substantial. Therefore, understanding the trade-off in effective sample size and computational time is important.

An attractive feature of Markov chain Monto Carlo and Bayesian statistics is its modular nature. New models are often developed by elaborating simpler models. A sampler for a new model can also often be constructed by extending a sampler for the simpler model. In this spirit, I wonder how easy it would be to extend the samplers in the stochastic volatility example to more complicated models with leverage or jumps in prices. There are important differences in computational cost for the different samplers in that example. The Riemann manifold Hamiltonian Monte Carlo sampler, in particular, benefits from a nice form of tensor. Would that carry over to these more complicated models?

Adam Gripton and Mike Christie (*Heriot-Watt University, Edinburgh*)

We discuss the application of Calderhead and Girolami (2009) and this paper to a sample imaging problem: inference of physical parameters from radiographic data. Fig. 19 shows the construction of a simulated radiograph from a mass profile and known experimental set-up. The shells are of known radii but unknown densities. We infer the true densities of the observed object via the noisy radiograph Fig. 19(c) along with two image features: dose (multiplicative constant) and scatter (additive). We assume the noise is Gaussian, so that

$$\tilde{\mathbf{w}}_{\text{obs}} = \mathbf{w}_c + \sigma_N \tilde{Z}, \tag{36}$$

$$\Rightarrow -\log\{\mathbf{p}(\mathbf{w}_{\text{obs}}|\theta)\} = \text{constant} + \frac{1}{2\sigma_N^2} \sum_{i=1}^{n} \{\mathbf{w}_{\text{obs}} - g(x_i; \theta)\}^2. \tag{37}$$

Our studies compare performances of Riemann manifold Hamiltonian Monte Carlo (RMHMC) and population Markov chain Monte Carlo (PMCMC) with naive MCMC and randomized maximum likelihood (RML) algorithms (Oliver *et al.*, 1996) in sampling parameters about a maximum *a posteriori* point. We pay particular attention to cases of low signal-to-noise ratio with a view to studying a Poisson noise model, where RML is known to produce results outside a 3σ-bound of the true parameters. We use a simplified form for the position-specific metric tensor:

$$\mathbf{G}(\theta) = \frac{1}{\sigma^2} D^{\mathrm{T}} D, \qquad D_{ik} = \frac{\partial g}{\partial \theta_k}(x_i; \theta). \tag{38}$$

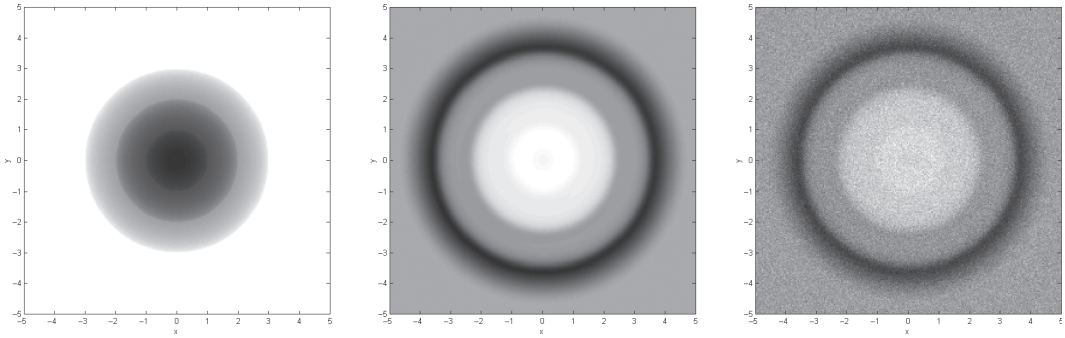

Fig. 19. (a) Mass profile and (b) clean and (c) noisy images

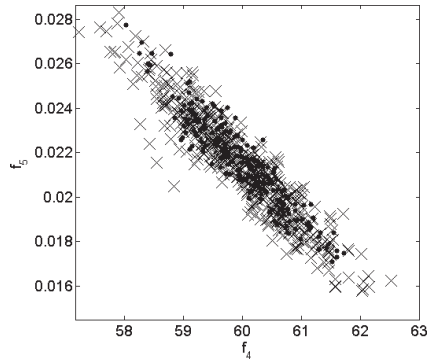

Fig. 20. PMCMC (×) *versus* RMHMC (•) algorithms

We find that this produces reasonable estimates of local covariance. We note the flexibility of the use of metric tensors in their ability to combine linearly, which is of particular use here as it can incorporate information about both prior beliefs (via the substitution $\boldsymbol{\Sigma}_{\mathrm{pr}}^{-1} \leftarrow \mathbf{G}$) and priors based on auxiliary measurements such as a known mass prior. We note that this property also allows sensible metric tensors to be computed for all temperature chains in PMCMC calculations, and their respective asymmetric proposal probability density function distributions:

$$Q_i(\theta'; \theta_0) \sim \mathcal{N}\{\theta_0, \hat{\mathbf{G}}_i^{-1}(\theta_0)\}, \tag{39}$$

where

$$\hat{\mathbf{G}}_i(\theta) = \beta_i \mathbf{G}(\theta) + (1 - \beta_i)\boldsymbol{\Sigma}_{\mathrm{pr}}^{-1}. \tag{40}$$

In Fig. 20 we present a comparison of samples from PMCMC and RMHMC sampling: in the PMCMC case we show samples from the top (posterior) chain, and in both cases we maintain the level of noise ($\sigma_n = 0.022$) used to produce Fig. 19(c). We observe the subspace of the dose f_4 and scatter f_5, which have true values of (60, 0.02) respectively. Similar results are produced, showing the efficacy of both methods in sampling about the maximum *a posteriori* point. In comparison with RML, we note that similiar distributions of results are produced with RMHMC sampling with a significant improvement in computational time. We find RMHMC sampling produces sample propositions in 0.15 s (equivalent to 360 min^{-1} at 90% acceptance), whereas RML produces maximum *a posteriori* points in 0.323 s (185 min^{-1}); iterations of PMCMC algorithms are 21200 min^{-1}. As yet, no direct comparisons can be made with RML, as the performance gain of PMCMC (and to a lesser extent RMHMC) sampling depends on the auto-correlation of the resultant samples—appropriate information regarding computational time will be provided when detailed comparisons of these methods are published.

Thiago Guerrera, **Håvard Rue and Daniel Simpson** (*Norwegian University of Science and Technology, Trondheim*)

It is our pleasure to congratulate the authors on this important contribution to the Markov chain Monte Carlo literature, which provides a way to construct an efficient scheme to sample from a complex target density that requires almost no tuning parameters. With this methodology, the authors accomplished what the recent literature on adaptive Markov chain Monte Carlo methods really has tried to do.

The observed Fisher information matrix can be employed as a metric when no closed form expression for the expected Fisher information matrix is available, although positive definiteness is no longer guaranteed. It is our experience that the observed Fisher information matrix can often be non-positive definite in regions of the parametric space 'far' from the bulk of probability mass of the target density. Have the authors any experience or, better, solutions, on this issue for problems where the expected Fisher information matrix is not analytically available?

Parallel computing is becoming increasingly important as the (computational) future is believed to be massive parallel computing. In a massive parallel environment (MPE), simulation algorithms that do run well in an MPE will naturally be preferred over algorithms that do not run so well in an MPE, even though the conclusion can be reversed running the same algorithms in a non-parallel environment. One simple example of two such algorithms can be the sequential Monte Carlo approach of Chopin (2002) compared with a single-site Gibbs sampler. Can the authors comment on the potential of the manifold Metropolis adjusted Langevin algorithm or Riemann manifold Hamiltonian Monte Carlo schemes in an MPE?

In the stochastic volatility example, a (sparse) Cholesky decomposition of the sparse metric tensor of the latent variables could have been taken instead of applying it directly to the dense inverse metric tensor. This approach would improve the performance of the manifold Metropolis adjusted Langevin algorithm scheme. We would also like to draw attention to the work by Lindgren *et al.* (2010), who construct Gaussian Markov random-field representations of Matérn fields that would improve the efficiency of the log-Gaussian Cox process example in Section 9 by using similar numerical ideas; some software is available at `http://www.r-inla.org`.

Desmond J. Higham (*University of Strathclyde, Glasgow*)

This important work deals with issues that are of immediate practical relevance, but it also raises many interesting technical questions. For example, Section 5, on the Riemann manifold and Metropolis adjusted Langevin algorithm, introduces a class of non-linear stochastic differential equations and a method for their numerical solution. As mentioned in Section 11, there is a need for theoretical analysis of long time properties such as ergodicity. Could the authors comment on the more fundamental issue of conditions that would guarantee existence or uniqueness of a solution to equation (9)? Similarly, in what sense is equation (10) regarded as a first-order integrator and what conditions would allow this to be established rigorously? Finally, would any of your computational examples currently succumb to rigorous analysis in this sense, and how large do you view the gap between what can be proved at present and what was observed in your challenging experiments?

Chris Holmes (*University of Oxford*)

The authors are to be congratulated on a truly excellent paper. I only have one very minor comment concerning some of the terminology used in Section 6. Preceding equation (19) the authors state that the Riemann manifold Hamiltonian Monte Carlo algorithm 'can once again be written as a Gibbs sampler'. As far as I am aware this does not match the usual definition of the term. The Riemann manifold Hamiltonian Monte Carlo algorithm appears to involve an accept–reject step and as such is more akin to a Metropolis-within-Gibbs or simply a Metropolis–Hastings algorithm, albeit with a beautifully efficient adaptive proposal distribution. The distinction is important. The auxiliary variable approach of Holmes and Held (2005) appears to be the only Gibbs sampler reported in the comparisons in Tables 3–7. The motivation for Holmes and Held (2005) was to develop a fully automated sampler with no user-set parameters *and* a 100% acceptance rate, i.e. a Gibbs sampler. In certain situations there is additional utility allied to this simplicity over effective sample size efficiency.

Dirk Husmeier (*Biomathematics and Statistics Scotland, Edinburgh*)

By combining concepts from physics (Hamiltonian dynamics) with Riemann geometry (the metric tensor), the authors have made a highly important contribution to current research on Markov chain Monte Carlo methods, which in my opinion will be as seminal as Green (1995). In this brief note, I shall discuss how

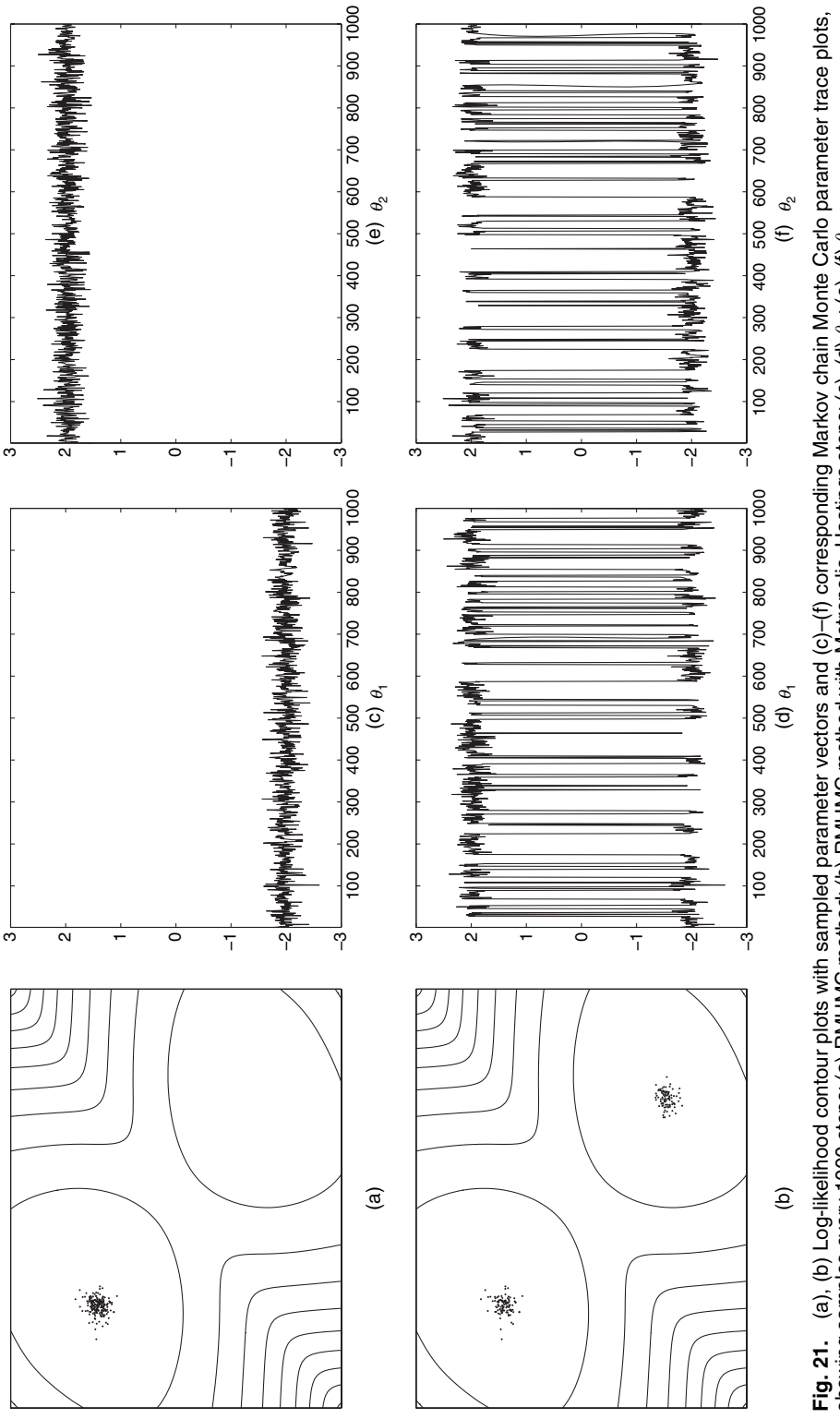

Fig. 21. (a), (b) Log-likelihood contour plots with sampled parameter vectors and (c)–(f) corresponding Markov chain Monte Carlo parameter trace plots, showing samples every 1000 steps: (a) RMHMC method; (b) RMHMC method with Metropolis–Hastings steps; (c), (d) θ_1; (e), (f) θ_2

mixing in the presence of multimodality may be further improved. For illustration, consider the Gaussian mixture

$$P(x_t|\boldsymbol{\theta}) = \frac{1}{2}\sum_{k=1}^{2}\Phi(x_t - \mu_k)$$

with parameter vector $\boldsymbol{\theta} = (\mu_1, \mu_2)$, where

$$\Phi(z) = \frac{1}{\sqrt{(2\pi)}}\exp\left(-\frac{z^2}{2}\right).$$

Given data $\mathbf{y} = \{x_1, \ldots, x_N\}$, the log-likelihood $\mathcal{L} = \Sigma_{t=1}^{N}\log\{P(x_t|\boldsymbol{\theta})\}$ is bimodal, with symmetric maxima at $\boldsymbol{\theta}^* = (\mu_1^*, \mu_2^*)$ and $\boldsymbol{\theta}^{**} = (\mu_2^*, \mu_1^*)$. Fig. 21(a) shows a contour plot for $N = 100$ and $\boldsymbol{\theta} = (2, -2)$.

To apply the Riemann manifold Hamiltonian Monte Carlo (RMHMC) method, compute the gradient

$$\frac{\partial\mathcal{L}}{\partial\mu_k} = \sum_{t=1}^{N} P(k|t, \boldsymbol{\theta})(x_t - \mu_k),$$

$$P(k|t, \boldsymbol{\theta}) = \frac{\Phi(x_t - \mu_k)}{2P(x_t|\boldsymbol{\theta})}$$

and the approximate Hessian (see Husmeier (2000) and Roberts *et al.* (1998))

$$-\frac{\partial^2\mathcal{L}}{\partial\mu_i\,\partial\mu k} = \delta_{ik}\sum_{t=1}^{N} P(k|t, \boldsymbol{\theta}).$$

Taking expectation with respect to $P(\mathbf{y}|\boldsymbol{\theta})$ leads to the approximate Fisher information matrix $\mathbf{G} = (N/2)\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

I applied the algorithm as described in Section 7.2, running 10^6 series of six Stormer–Verlet (leapfrog) numerical integration steps with step size $\varepsilon = 0.5$. Figs 21(a), 21(c) and 21(e) show that the sampled parameters are restricted to one of the two modes. To understand the reason for this poor mixing, note that, in statistical physics terms, the Hamiltonian of equation (13) contains two contributions: the 'potential energy' $V = -\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\log\{(2\pi)^D|\mathbf{G}|\}$, and the 'kinetic energy' $K = \frac{1}{2}\mathbf{p}^{\mathrm{T}}\mathbf{G}^{-1}\mathbf{p}$. For mode jumping, a potential energy barrier of $\Delta V \approx \mathcal{L}(\boldsymbol{\theta}^*) - \mathcal{L}(\mathbf{0})$ must be overcome, which in the present example is $\Delta V \approx 100$. Hamiltonian dynamics conserve the 'energy', $H = V + K = \text{constant}$, and hence require $\Delta K \approx \Delta V$ to overcome the barrier. With $\mathbf{p}$ sampled from $\mathcal{N}(\mathbf{0}, \mathbf{G})$, the probability of mode jumping is less than 10^{-20}.

As a remedy, I included a Metropolis–Hastings (MH) step with a long-tailed proposal distribution after each RMHMC step. Compared with standard MH algorithms, subsequent samples are approximately uncorrelated by virtue of the RMHMC scheme, as the authors have shown, and larger MH step sizes than reasonable in standard MH algorithms can be adopted. I tested this procedure using a Cauchy proposal distribution with scale parameter 1 and location set to the current parameter value. Figs 21(b), 21(d) and 21(f) indicate that both modes are now visited equally often, and the mixing problem has been overcome.

A more sophisticated approach is based on population MCMC sampling with annealing, which the authors have investigated in their previous work (Calderhead and Girolami, 2009). By integrating RMHMC methods into this framework, its power can be further increased.

Shiro Ikeda (*Institute of Statistical Mathematics, Tokyo*)

I congratulate the authors for their stimulating paper which reveals the importance of the Riemannian geometry of statistical models for Markov chain Monte Carlo sampling. This opens a new direction for further research.

One natural Riemannian metric of statistical models is that specified by the Fisher information matrix. It is not necessary but, as was pointed out in one of the comments, it is interesting to consider the divergence function which induces the metric.

The family of divergence functions which induce the Fisher metric has been studied. One family is known as the α-divergence family (Amari, 1985) which includes the Kullback–Leibler divergence and the Hellinger distance. Another general family has been discussed in Zhang (2004). Among the families, which divergence function is important for the current framework?

In the presentation, Girolami mentioned that the Riemannian connection plays an important role. Although the Riemannian metric that is induced by divergence functions of the above family is always the Fisher metric, the connection that is induced by them is not necessarily the Riemannian connection.

Information geometry reveals the importance of dual connections which are induced by each divergence function. As a special case, those dual connections become identical to the Riemannian connection. One such divergence function is the Hellinger distance which has symmetric divergence.

Information geometry has shown that dual connections play important roles in statistics. For example, two important connections known as *e*- and *m*-connections are dual and associated with the Kullback-Leibler divergence. We have not seen dual connections in the current paper but only the Riemannian connection. It seems that the reason comes from the reversibility which is important for this Markov chain Monte Carlo framework. This may explain why the symmetric divergence function is preferable for characterizing the framework and raises questions whether we can remove the reversibility and what happens in such a case.

Ajay Jasra (*Imperial College London*) **and Sumeetpal Singh** (*University of Cambridge*)

We congratulate the authors on a thought-provoking paper. Our comments relate to the quantities $\mathbf{G}(\theta)$ and ε. Our first remark is associated with that of the authors' on tempering on page 151. Our idea is to use $\mathbf{G}(\theta)$ in the context of sequential Monte Carlo (SMC) samplers (Del Moral *et al.*, 2006). SMC samplers simulate from a sequence of densities on the same space. For example, tempering; for $0 < \psi_1 < \ldots < \psi_q = 1$ simulate from $p_i(\theta) \propto p(\theta)^{\psi_i}, i = 1, \ldots, q$. Our idea is to introduce a sequence $\mathbf{G}_{\psi_1}(\theta), \ldots, \mathbf{G}_{\psi_q}(\theta) = \mathbf{G}(\theta)$, and then to use an SMC sampler to sample from the densities that are induced by the $\mathbf{G}_{\psi_i}$. For example, consider the model in Section 8.2. One would like to make the off-diagonal elements of $\mathbf{G}$ smaller to reduce the correlations. On inspection of $\mathbf{G}$ this is achieved by replacing ϕ with $\psi_i \phi$ (ψ_i as above). Although this is obvious for auto-regressive AR(1) models, the point is that via $\mathbf{G}(\theta)$ one could have a canonical choice for a sequence of densities, using the interpretation of the space as a Riemann manifold, to deal with high correlations within θ, whereas a tempering method may not always achieve this. (The authors' Markov chain Monte Carlo (MCMC) algorithms could be adopted in this context but it is not central to the idea.) There are few references on how this may be achieved (Chopin, 2007). Another benefit arises when computing $\mathbf{G}(\theta)$ is costly, e.g. for the model in Section 10 when T is large. If another sequence is chosen so that $\mathbf{G}_i$ is cheaper to calculate for $i < q$ and we use the manifold Metropolis adjusted Langevin algorithm (MMALA) or Riemann Hamiltonian manifold Monte Carlo algorithm in MCMC steps in the SMC sampler, then these MCMC kernels are used sparingly. For example, $\mathbf{G}_i$ corresponds to the posterior only including observations for time points $i \leqslant T$ and $q = T$ (e.g. Chopin (2002)).

At a superficial level there appears to be a connection between the authors' ideas and those in optimization (e.g. Bertsekas (1999)). For example, there are many choices for $\mathbf{G}(\theta)$ and ε if we were to interpret these as the scaling matrix of the gradient and the step size of an iterative gradient method. The MMALA for constant curvature resembles Newton's method. In this context, there are ways of dealing with a non-positive definite Hessian which if adopted could allow the use of the observed Fisher information matrix in the MMALA. In the authors' opinion, would any of these ideas proposed in the optimization literature be useful in an MCMC context?

Peter Jupp (*University of St Andrews*)

I congratulate the authors on this enjoyable paper. It is very pleasing that the geometrically natural Hamiltonian that is defined in equation (13) can prove so effective computationally.

The Hamiltonian in equation (13) is constructed by using a Riemannian metric on the parameter space Θ, and the authors mention various choices of metric. Some others arise as follows.

(a) If Θ is a vector space (or, more generally, a Lie group) and a prior is given then the Fisher information matrix of the translation model generated by the prior is a Riemannian metric on Θ. (This metric is the *information in the prior* that plays a role in the van Trees inequality of van Trees (l968); see also Jupp (2010).)

(b) A *yoke* on Θ is a smooth function on $\Theta \times \Theta$ that satisfies certain conditions on its first and second derivatives on the diagonal $\{(\theta, \theta) : \theta \in \Theta\}$. Yokes are almost the same as *contrast functions and divergences*. Every yoke on Θ gives rise to a Riemannian metric on Θ.

From the geometrical viewpoint the point $(\boldsymbol{\theta}, \mathbf{p})$ occurring in equation (13) is an element of the cotangent bundle $T^*\Theta$ of Θ and the Hamiltonian mechanics arise from a canonical symplectic form on $T^*\Theta$. Every yoke on Θ yields a symplectic form on some neighbourhood of the diagonal of $\Theta \times \Theta$. (See Barndorff-Nielsen and Jupp (1997a, b).) If this form is defined on all of $\Theta \times \Theta$ then we can consider Hamiltonian mechanics on $\Theta \times \Theta$ with the yoke as the Hamiltonian. It would be interesting to compare the behaviour of these Hamiltonians with that of those considered in the paper.

Theodore Kypraios (*University of Nottingham*)

I congratulate the authors for this very exciting piece of work on improving the existing Markov chain Monte Carlo methodology when sampling from target distributions that are of high dimension and exhibit strong correlations. This is often the situation that someone is faced with in stochastic epidemic modelling. However, there is a distinct difference compared with the example that are illustrated in the paper. The target distribution has a discontinuous probability density function.

In general, drawing (Bayesian) inference for disease outbreak data is complicated because there are various levels of inherent dependence that we need to take into account and, in addition, the epidemic process is partially observed (for example times at which people become infected are rarely known). To overcome this issue, a data augmentation approach can be employed in conjunction with Markov chain Monte Carlo sampling (see, for example, O'Neill and Roberts (1999) and Gibson and Renshaw (1998)) but standard algorithms can be problematic in various settings. Therefore, algorithms that are based on non-centred reparameterizations can offer a better alternative (see, for example, Neal and Roberts (2005) and Kypraios (2007)).

In this discussion, I would like to hear the authors' views on the applicability of Hamiltonian Monte Carlo (HMC) and Riemann manifold HMC (RMHMC) sampling in a context which shares some similarities with the most commonly used epidemic models. Suppose that we consider d-dimensional target distributions of the form

$$\pi_d(\mathbf{x}^d) = \prod_{i=1}^{d} f(x_i^d) \tag{41}$$

where

$$f(x) \propto \exp\{g(x)\} \qquad (0 < x < 1)$$

where $g(\cdot)$ is twice differentiable on [0,1] and $f(x) = 0$ otherwise. Is it obvious how someone would implement an efficient HMC algorithm? If we decide to employ an RMHMC algorithm what would the geometric structure be?

Although, in the epidemic models, we have to deal with more complex target densities where there is strong dependence between the d components, and in addition, each conditional distribution $\pi(x_i|\mathbf{x}_{-i})$ could be a piecewise exponential, I believe that model (41) should give us some insight on the applicability of the HMC and RMHMC methods for discontinuous probability density functions. Therefore, I look forward to hearing the authors' view on this.

Krzysztof Łatuszyński, Gareth O. Roberts, Alexandre Thiéry and Katarzyna Wolny (*University of Warwick, Coventry*)

We congratulate the authors for the exciting and thought-provoking paper. They approach the fundamental problem of scaling Markov chain Monte Carlo algorithms by introducing a metric on probability distributions and altering the dynamics of the underlying stochastic differential equation accordingly to obtain more efficient, state-dependent proposals. In context of the Metropolis adjusted Langevin algorithm (MALA) we note that any choice of $\sigma(\theta)$ in

$$\mathrm{d}\theta(t) = \left[\frac{\sigma^2\{\theta(t)\}}{2}\mathcal{L}\{\theta(t)\}' + \sigma\{\theta(t)\}\,\sigma\{\theta(t)\}'\right]\mathrm{d}t + \sigma\{\theta(t)\}\,\mathrm{d}b(t) \tag{42}$$

yields a Langevin diffusion with the correct stationary distribution $p(\theta)$. By choosing $\sigma^2(\theta) = |1/\mathcal{L}(\theta)''|$ we arrive at their manifold MALA (MMALA) with the metric tensor $G(\theta)$ defined as the observed Fisher information matrix plus the negative Hessian of the log-prior (see Sections 4.2 and 5).

To have an insight into ergodicity and robustness of the above version of the MMALA, we analyse it for the family of targets $p(\theta) \propto \exp(-|\theta|^\beta)$. It provides a good benchmark for Markov chain Monte Carlo algorithms: for the random-walk Metropolis (RWM) algorithm and for the standard MALA, it is well known for which values of β (or β and ε respectively) geometrical ergodicity holds, and for which it fails (Roberts and Tweedie, 1996; Mengersen and Tweedie, 1996). This is summarized in Table 14 (see Roberts and Tweedie (1996) and Mengersen and Tweedie (1996) for details), together with the respective properties of the MMALA.

These results indicate very impressive theoretical properties. It appears that for $\beta > 1$ MMALAs can inherit the stability properties of the RWM algorithm together with the speed-up of the MALA. Moreover, it is applicable for $\beta < 1$, where the other two algorithms fail. We did our calculations by establishing drift

Table 14. Geometric ergodicity of the RWM, MALA and MMALA algorithm for target $\pi(\theta) \propto \exp(-|\theta|^{\beta})$†

Algorithm	*Ergodicity of the following conditions:*				
	$0<\beta<1$	$\beta=1$	$1<\beta<2$	$\beta=2$	$2<\beta$
RWM	N	Y	Y	Y	Y
MALA	N	Y	Y	Y*	N
MMALA	Y*	x	Y*	Y*	Y*

†N, geometric ergodicity fails; Y, geometric ergodicity holds; Y*, geometric ergodicity holds for ε sufficiently small; x, not applicable.

conditions and analysing acceptance rates for $d = 1$, and we expect to observe the same phenomenon also in higher dimensions.

V. Mansinghka (*Navia Systems, Berkeley*)
I congratulate Girolami and Calderhead on their novel and useful methods and clear, empirically grounded presentation. Hessian corrections (as well as richer geometric notions) have proved essential for improving the convergence of deterministic numerical algorithms, and the paper makes a strong case for their utility in stochastic simulation.

In fact, the manifold Metropolis adjusted Langevin algorithm (MMALA) and the Riemann manifold Hamiltonian Monte Carlo (RMHMC) algorithm may be nearly automatically applicable, owing to recent progress in the theory and practice of automatic differentiation (AD). Siskind and Pearlmutter have recently developed a general method (and a usable implementation) that can evaluate derivatives of (nearly) arbitrary real-valued programs, automatically exploiting dynamic programming (i.e. the 'back-propagation trick') where possible. Better still, it incurs only a small constant overhead over evaluating the program and can be nested to obtain efficient, accurate Hessian vector products automatically.

I have troubled graduate school memories of laboriously deriving gradients and Hessians, painstakingly manipulating the expressions to promote efficient computation. The new AD methods make this skill obsolete. They may let me automatically apply for example the MMALA by simply calling nested AD on the raw code for evaluating the joint density. Although AD systems have already been used for maximum likelihood estimators and for Hamiltonian Monte Carlo methods, the high curvature of the energy landscapes that arise in situations where AD is necessary make the MMALA and RMHMC algorithm far more appealing.

The paper also highlights a riddle that arose for me in my research on probabilistic computation. In precisely the continuous models where gradient-sensitive Markov chain Monte Carlo methods seem to help, our uncertainty suggests that the continuum may be overkill. Consider the hyperparameters describing the component prior in a mixture of Bernoulli distributions. Rather than use an adaptive sampler to avoid slow mixing Metropolis–Hasting steps, one could simply grid the unit interval, and transform the grid points to cover the positive real numbers sparsely, as in numerical integration. This discretization is amenable to Gibbs sampling, can be made exceedingly efficient (both in software and via probabilistic digital circuits) and is often sufficiently accurate precisely because of the posterior uncertainty.

Often, most of the bits of our continuous variables do not actually matter, in either our prior or our posterior. Also, in our actual machines, no continuous quantities exist and, according to information theory, in some sense none can. Although I expect the MMALA and RMHMC algorithm to make continuous variables less terrifying to modellers in the trenches, it sometimes seems to me that this line of inquiry may already beg some of the deepest questions about what makes Bayesian inference appear computationally challenging.

Jean-Michel Marin (*Université de Montpelier 2*) **and Christian P. Robert** (*Université Paris Dauphine, Ceremade and Centre de Recherche en Economie et Statistique, Paris*)
This paper is a welcome addition to the recent Markov chain Monte Carlo literature and the authors are to be congratulated for linking the two threads that are the Langevin modification of the random-

walk Metropolis–Hastings algorithm and Hamiltonian acceleration. Overall, trying to take advantage of second-order properties of the target density $\pi(\theta)$, just like the Langevin improvement takes advantage of the first order (Roberts and Tweedie, 1995; Stramer and Tweedie, 1999a, b), is a natural idea which, when implementable, can obviously speed up convergence. This is the Langevin part, which may use a fixed metric $\mathbf{M}$ or a local metric defining a Riemann manifold, $\mathbf{G}(\theta)$. Obviously, this requires that the derivation of an appropriate (observed or expected) information matrix $\mathbf{G}(\theta)$ is feasible up to some approximation level, or otherwise that sufficiently authoritative directions are given about the choice of an alternative $\mathbf{G}(\theta)$.

Whereas the logistic example that is used in the paper mostly is a toy problem (where importance sampling works extremely well, as shown in Marin and Robert (2010)), the stochastic volatility model is more challenging and the fact that the Hamiltonian scheme applies to the missing data (volatility) as well as to the three parameters of the model is quite interesting. We would thus welcome more details on the implementation of the algorithm in such a large dimension space. We wonder, however, about the appeal of this involved scheme when considering that the full conditional distribution of the volatility can be simulated exactly.

Xiao-Li Meng (*Harvard University, Cambridge*)

The following three fortune cookies (which are known generally for their seemingly thoughtful Chinglish) are my token of thanks to the authors for providing much food for thought. First, the parallelism between EM-type and Gibbs-type algorithms is so striking (van Dyk and Meng, 2010) that I have long wondered what would be a 'quadratically converging' Markov chain Monte Carlo (MCMC) algorithm to the Gibbs sampler for sampling $p(\theta|\mathbf{y})$ as Newton–Raphson (NR) iteration is to the EM algorithm for maximizing $p(\theta|\mathbf{y})$ or equivalently $\mathcal{L}(\theta) = \log\{p(\theta|\mathbf{y})\}$? In the context of maximization, infinitely many *predetermined* M can ensure that

$$\theta^{(n+1)} = \theta^{(n)} + M\,\nabla_\theta\,\mathcal{L}(\theta^{(n)}), \qquad n = 0, 1, \ldots \tag{43}$$

will converge *linearly* to a root of $\nabla_\theta\,\mathcal{L}(\theta) = 0$. NR iteration replaces M *adaptively*. with $M_n = H^{-1}(\theta^{(n)})$, where $H(\theta) = -\nabla_\theta^2\,\mathcal{L}(\theta)$ or its approximations (e.g. for quasi-NR iteration), to achieve *superlinear* convergence. Translated to the MCMC setting, the latter mean *super-geometric* convergence:

$$\| p^{(n)}(\theta|\theta^{(0)}, \mathbf{y}) - p(\theta|\mathbf{y})\| \leqslant c_{\theta^{(0)}}\rho^{q^n}, \qquad \text{for } n \text{ sufficiently large,} \tag{44}$$

using the usual notation but with the 'double power' ρ^{q^n}, where $0 \leqslant \rho < 1$ and $q < 1$. The improvement over the Metropolis adjusted Langevin algorithm or Hamiltonian Monte Carlo algorithm by replacing their 'M's with the local metric tensor $G^{-1}(\theta^{(n)})$ or its approximations seem to follow the same recipe, considering that G and H are the same or approximating each other. This analogy makes me wonder whether we are witnessing the birth of super-geometric converging Markov chain Monte Carlo methods (excluding perfect sampling, of course) or whether I simply had too much MSG (Markovianly super-geometric?) ....

Second, the authors' geometric insights reminded me of path sampling for computing normalizing constants, initially also developed by physicists as thermodynamic integration. As shown in Gelman and Meng (1998), the global (but unachievable?) optimal path is the geodesic path under the same Fisher–Rao metric as used by the authors, and the optimal path within a distribution family requires solving an Euler–Lagrange equation identical to the authors' equation for determining the Hamiltonian flow, except that the former uses unnormalized densities. One therefore may ponder whether algorithmically Riemann manifold Hamiltonian Monte Carlo sampling is a 'normalized' optimal path sampling in the sense of finding and moving along the shortest (manifold) path between the initial $p^{(1)}(\theta|\theta^{(0)}, \mathbf{y})$ and the target $p(\theta|\mathbf{y})$. Such a connection may allow borrowing strategies from the path sampling literature (e.g. by purposely going outside a given distribution family for better Monte Carlo efficiency).

Third, the authors' warped bivariate Gaussian example (from their presentation) reminded me of warp bridge sampling (Meng and Schilling, 2002), which shortens the distance by moving the (unnormalized) densities closer via *warping* them into similar manifolds *before sampling*. I wonder whether a similar strategy can be interwoven into the authors' methods for further improvements.

Antonietta Mira (*University of Lugano and Università dell'Insubria*) **and Heikki Haario** (*University of Lappeenranta*)

We compliment the authors for providing an advance in the Markov chain Monte Carlo (MCMC) literature that gives clear insight into geometrical concepts involved in Monte Carlo simulation. Following are some comments and questions.

(a) In this paper the use of derivatives of the log-target is exploited to design efficient Hamiltonian Monte Carlo (HMC) algorithms and Metropolis adjusted Langevin algorithms (MALAs). A different way of using information contained in these derivatives, in the MCMC setting, is the zero-variance (ZV) strategy of Assaraf and Caffarel (2003) and Mira *et al.* (2010). Instead of redesigning the transition kernel of the Markov chain, we substitute the original function f, whose expected value we are interested in, with a different one that has the same mean but much smaller variance. The new function is constructed by adding to f a linear combination of control variates. This well-known variance reduction technique has never been used before in MCMC sampling owing to the difficulty of constructing control variates. The only other one successful attempt that we know is by Dellaportas and Kontoyiannis (2010). We suggest combining the ZV strategy with the manifold MALA (MMALA) and the Riemann manifold HMC (RMHMC) algorithm since the required derivatives are already available, plus ZV only needs post-processing an existing Markov chain so it is easy to combine with clever samplers.
(b) How do the MMALA and RMHMC algorithm work for high dimensional multimodal targets? Here ZV is not as powerful as in other cases where we tested it, so we expect that samplers that exploit derivatives of the log-target might also break down while a combination of delayed rejection (Mira *et al.*, 2010) and adaptation (Haario *et al.*, 2006) has proved powerful here.
(c) The authors always insist on reversibility but only stationarity with respect to the target is required. There are references that prove that non-reversible samplers have better performance (Diaconis *et al.*, 2000; Mira and Geyer, 2000). Maybe there is a scope in looking for non-reversible kernels in particular for RMHMC sampling.
(d) The authors do not seem to worry about conditions that guarantee finite Fisher information: are the MMALA and RMHMC algorithm well defined if this quantity and its derivatives are infinite?
(e) We wonder why the authors use only single-component Metropolis steps in the ordinary differential equation example. In this case of three parameters, they need 15000 target evaluations to produce the 5000 samples. We implemented a vanilla random-walk Metropolis algorithm with a Gaussian proposal having covariance given by linearization via the Jacobian matrix at the maximum likelihood estimate. The results vary depending on the linearization point but, by standard covariance adaptation (Haario *et al.*, 2001), the efficiency was always enhanced so that the effective sample size of all parameters was between 300 and 400. As time and effective sample size were improved threefold a typical relative speed value for our Metropolis step is 37 (average over 10 runs; compare with Table 11). So, although other approaches are certainly needed in more complicated situations, vanilla algorithms are difficult to beat in relatively easy examples.

Iain Murray (*University of Edinburgh*) **and Ryan Prescott Adams** (*University of Toronto*)

This is an exciting piece of work. We agree with the authors that Hamiltonian Monte Carlo (HMC) methods could be used more broadly. HMC methods have a reputation as being difficult to tune and we hope that the more robust behaviour that is demonstrated here will help to alleviate this.

Dynamical methods jointly update variables, which may allow larger moves in complex models than updating variables independently. Hierarchical models often contain strong prior dependences between hyperparameters and parameters and would seem to provide a common case where HMC sampling offers benefits. Unfortunately, updating hyperparameters and parameters jointly by using standard HMC methods does not necessarily work better than updating them individually (Choo, 2000).

Consider a simple hierarchical model (Neal, 2003) where the observations are uninformative about the parameters:

$$\begin{aligned} v &\sim \mathcal{N}(0, \sigma^2), && \text{e.g. } \sigma = 3, \\ x_k &\sim \mathcal{N}\{0, \exp(v)\}, && k = 1, \ldots, K = 10, \\ \mathbf{y} &\sim p(\mathbf{y}), && \text{observations are independent of } \boldsymbol{\theta} = (v, \mathbf{x}). \end{aligned}$$

The posterior of this trivial model is equal to the prior, $p(\boldsymbol{\theta})$.

Sampling from $p(\boldsymbol{\theta})$ by using simple Markov chain methods can be difficult:

(a) sensible step sizes for $\mathbf{x}$ depend on v, and
(b) the hyperparameter v cannot move very much for fixed effects $\mathbf{x}$ (especially for large K).

One way to deal with these issues is reparameterization: rewriting the model in terms of $z_k \sim \mathcal{N}(0, 1)$ with $x_k = z_k \exp(v/2)$ gives independent variables $(v, \mathbf{z})$. Most Markov chain methods will now sample effectively

from the prior. When the distribution is reweighted by non-trivial likelihood terms, other reparameterizations may be appropriate (Christensen *et al.*, 2006; Murray and Adams, 2010).

Perhaps using a metric can replace the need for careful reparameterization, although the metric used in this paper's examples would not work: the negative Hessian $-\partial^2 \log\{p(\boldsymbol{\theta})\}/\partial\boldsymbol{\theta}^2$ is not positive definite everywhere. We could use

$$G(\boldsymbol{\theta}) = -E_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log\{p(\mathbf{y}|\boldsymbol{\theta})\}\right] - E_{\mathbf{x}|v}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log\{p(\mathbf{x}|v)\}\right] - \frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log\{p(v)\}, \tag{45}$$

which is positive definite. This metric would lead to sensible step sizes for the random effects x_k: the diffusion terms $\mathbf{z}$ are scaled by the standard deviation $\exp(v/2)$. However, the metric is diagonal, which suggests that problems with the dependences in the prior are not alleviated, at least in the weak data limit.

As highlighted in Section 4.2, there is a broader choice of metrics to be considered. We hope that algorithms that are presented here encourage renewed interest in developing new variable metric Hamiltonian methods for hierarchical models.

Matthew F. Parry (*University of Cambridge*)

The authors have added to the power of Monte Carlo methods the elegance of information geometry. The key ingredient in this paper is the recognition that in statistical models the parameter space is endowed with a Riemannian metric structure. The Fisher metric is used in the paper but, as the authors note, this choice is neither required nor necessarily possible in practice.

A feature of the 'Riemannian algorithms' that is worth mentioning for both theoretical and practical reasons is that they are *covariant*, i.e. their form is unchanged by reparameterization. As a consequence, if a particular parameterization is convenient for some reason, the form of the algorithm in this parameterization is immediate and does not have to be derived from some 'true' parameterization.

Let $p(\theta)$ be a probability density and $G_{ij}(\theta)$ be the metric (components) on the parameter space, and consider the reparameterization given by $\theta^i \to \bar{\theta}^i = \bar{\theta}^i(\theta)$, where $i, j = 1, \ldots, D$. The quantity $p(\theta)/\sqrt{\det\{G(\theta)\}}$ transforms as a scalar and consequently so does $L(\theta) := \log[p(\theta)/\sqrt{\det\{G(\theta)\}}] = \mathcal{L}(\theta) - \frac{1}{2}\log[\det\{G(\theta)\}]$. In contrast, the inverse metric transforms as

$$G^{ij}(\theta) \to \bar{G}^{ij}(\bar{\theta}) = \frac{\partial\bar{\theta}^i}{\partial\theta^k}\frac{\partial\bar{\theta}^j}{\partial\theta^l}G^{kl}(\theta), \tag{46}$$

where a sum is implied over repeated indices.

Hamiltonian algorithm

The reparameterization $\bar{\theta}^i = \bar{\theta}^i(\theta)$ is one half of a canonical transformation; the other half is the transformation of the momenta,

$$p_i \to \bar{p}_i = \frac{\partial\theta^j}{\partial\bar{\theta}^i}p_j. \tag{47}$$

It then follows from equations (46) and (47) that $K(\theta, p) := \frac{1}{2}G^{ij}p_i p_j$ transforms as a scalar. Thus the Hamiltonian $H(\theta, p) := -L(\theta) + K(\theta, p)$ is a scalar. Furthermore, the Hamiltonian remains covariant if $K \to T(L, K)$, though one will need to ensure that $\exp(-T)$ is a suitable probability density for $p|\theta$. By contrast, standard Hamiltonian Monte Carlo sampling fails to be covariant because the mass matrix is typically no longer constant after reparameterization.

Langevin algorithm

Covariance is established via Itô calculus (Hughston, 1996; Brody and Hughston, 1999): let $b^{\mathbf{a}}$, $\mathbf{a} = 1, \ldots, D$, be Brownian motions satisfying $\mathrm{d}b^{\mathbf{a}}\,\mathrm{d}b^{\mathbf{b}} = \delta^{\mathbf{ab}}\,\mathrm{d}t$, noting that $\mathbf{a}$ is a label as opposed to a tensorial index. Then the Langevin diffusion,

$$\mathrm{d}\theta^i = \frac{1}{2}G^{ij}\frac{\partial}{\partial\theta^j}L(\theta)\,\mathrm{d}t - \frac{1}{2}G^{jk}\Gamma^i_{jk}\,\mathrm{d}t + \psi^i_{\mathbf{a}}\,\mathrm{d}b^{\mathbf{a}}, \tag{48}$$

where Γ^i_{jk} is any connection and $\psi^i_{\mathbf{a}}$ is defined so that $G^{ij} = \psi^i_{\mathbf{a}}\psi^j_{\mathbf{b}}\delta^{\mathbf{ab}}$, is covariant. That is to say, if we replace all quantities in equation (48) with their barred counterparts, the equation remains valid.

As a final comment, it is worth noting that the observed Fisher information transforms as in equation (46).

W. D. Penny (*Wellcome Trust Centre for Neuroimaging, London*)

I congratulate the authors for producing what I now regard as the state of the art method for Bayesian estimation of posterior densities. The authors show that by implementing Markov chain Monte Carlo methods on the Riemann manifold one can obtain remarkable improvements on many difficult inference problems. As an applied statistician I shall restrict my comments to the applications, specifically to Section 10.1. My aim here is to correct a misnomer from spreading further within the statistics community. What follows is rather tangential to the main thrust of the paper but may be of more general appeal as it shows an interesting interaction between neuroscience, dynamical systems theory and statistical inference.

Equation (23) in the paper is described as the 'Fitzhugh–Nagumo' equation, quoting Ramsay *et al.* (2007). However, this is incorrect. The Fitzhugh–Nagumo equation is

$$\dot{V} = c\left(V - \frac{V^3}{3} + R + I\right), \tag{49}$$

$$\dot{R} = -\left(\frac{V - a + bR}{c}\right)$$

where V is voltage and R is a recovery variable. The motivation behind the Fitzhugh–Nagumo model was to create a simple two-variable system that could approximate the more complex four-variable Hodgkin–Huxley model that describes a spiking neuron. A minimal requirement for this is that in the absence of input I no spiking is produced. The above behaviour can be guaranteed if, for $I = 0$, the system has a single stable equilibrium point. Fitzhugh showed, via stability analysis, that this will be so if $1 - 2b/3 < a < 1, 0 < b < 1,$ and $b < c^2$ (Fitzhugh, 1961). Sufficiently strong input then moves the system into a stable limit cycle, producing periodic spiking. Nagumo then built an electronic circuit exhibiting these properties.

Equation (23) has no input variable, and for $I = 0$ the parameters are outside the regime specified by Fitzhugh. So, although the likelihood landscape over a and b has multiple local maxima (Ramsay *et al.*, 1997), which therefore presents a challenging problem for Bayesian inference, this has no direct relevance to biophysics. This provides an example of a more general point. Because non-linear differential equations have such a rich dynamical repertoire, a model is not properly specified unless one also provides a statement about the allowable parameter regime(s). The Bayesian paradigm is ideally suited to this context and I envisage that biophysically meaningful inferences will soon be drawn with Riemannian Markov chain Monte Carlo methods.

Anthony Pettitt (*Queensland University of Technology, Brisbane*)

This is a very interesting paper for developing fully adaptive moves in a family of Markov chain Monte Carlo algorithms. The paper shows the improved efficiency compared with the standard Metropolis adjusted Langevin algorithm in several interesting examples.

The paper concentrates on using the expected Fisher information matrix as a metric tensor. There are some models where the Fisher information is not straightforward to find such as those involving latent or hidden variables or random effects. The complete-data likelihood is available but not the observed data likelihood. One example is the motor unit estimation problem of Ridall *et al.* (2007). Other examples include situations where there are Gaussian random effects in non-linear non-Gaussian observation models and the observed data likelihood requires high dimensional integration of the Gaussian random effects. To what extent does the new algorithm work in this extended set of models?

A notoriously difficult Markov chain Monte Carlo algorithm to design so that it has reasonable acceptance rates is the reversible jump Markov chain Monte Carlo algorithm. Does the new approach have anything to offer the analyst in this context? Can a suitable Hamiltonian be defined in a higher dimensional space? Provided that the problem under study provides a Fisher information matrix for the higher dimensional model it would appear that the new approach could be extended.

Does the new algorithm have anything to offer for the designer of a Markov chain Monte Carlo algorithm which has to deal with a multimodal posterior? Since the observed information matrix is not utilized in the new algorithm how could such information be incorporated in determining an appropriate path for the new method so that multiple modes could be visited?

A challenging example which has elements of the issues raised above is the motor unit number estimation problem (Ridall *et al.*, 2007). A challenging aspect of the algorithm development is the multimodality as exemplified by Glasbey in the discussion of Ridall *et al.* (2007) which has yet to be satisfactorily dealt with. The authors' views on these issues would be greatly appreciated.

Natesh S. Pillai (*Harvard University, Cambridge*) **and Gareth O. Roberts** (*Warwick University, Coventry*)
The authors are to be congratulated for a thought-provoking paper. There are many interesting issues to be solved. In a recent paper (Beskos *et al.*, 2010) co-authored with colleagues we show that, when the mass matrix is fixed and the target density is of a product form, at stationarity, the discretization error for the Hamiltonian Monte Carlo algorithm should be of the order $N^{-1/4}$ for *any* time reversible volume preserving integrator, where N is the dimension of the target distribution. Furthermore, for the Verlet integrator, this value of the discretization error leads to an optimal acceptance probability of 0.651, confirming earlier conjectures based on simulation. This is a significant improvement over the Langevin algorithm where the optimal step size is $N^{-1/3}$ and the acceptance probability is 0.574 and the random-walk Metropolis algorithm where the corresponding quantities are N^{-1} and 0.234. We wonder what the corresponding results are in the case of Riemannian Hamiltonian Monte Carlo algorithms.

Another natural question is to consider the same problem that the authors study but in the simpler case of the random-walk Metropolis algorithm, i.e. the two proposals $y = x + Z$ *versus* $y = x + Z/\sqrt{I(x)}$ where $Z \sim N(0, I)$, and $I(\cdot)$ is the Fisher information matrix. Is the latter proposal better? Is there a strict inequality in the operator norm or in the L_2-norm?

C. R. Rao (*Hyderabad*)
I am glad to see the use of the Rao metric of a Riemannian manifold based on the Fisher information matrix in the development of a novel Markov chain Monte Carlo sampling method. The metric is especially useful in the discussion of problems of statistical inference as discussed by various authors. As observed by the authors of the present paper, other metrics may be more appropriate in certain situations. In two papers, Burbea and Rao (1982, 1984) developed a variety of metrics based on entropy functions such as Shannon entropy and on general divergence measures. They suggested a general method of obtaining a variety of metrics. They considered any C^2-function $F(\cdot, \cdot)$ on $R_+ \times R_+$ so that $F(s, t) > 0$ and $F(s, s) = 0$ for $s, t \in R$. For $p, q \in P_\mu$ we define the divergence of p and q with respect to F as

$$D_F(p, q) = \int_x F\{p(t), q(t)\} \, \mathrm{d}_\mu(t), \qquad p, q \in p_\mu.$$

Fixing $p \in P_\mu$ and letting q vary we find that

$$D_F(p, p) = \mathrm{d} D_F(p, q)|_{q=p} = 0,$$

$$d^2 D_F(p, q)|_{q=p} \geqslant 0.$$

In particular, when $p = p(t|z)$ and $q = p(t|\zeta)$ with $z, \zeta \in D$ and $p(\cdot|\cdot) \in \mathcal{F}(\mathcal{X}|D)$, we have for the (complex) Hessian, by analogy with $D_F(p, q)$,

$$\Delta_{\partial p}\{D_F(p, p)\}(z) = 4 \int_x \partial_q^2 F(p, q)|_{q=p} |\partial p|^2 \, \mathrm{d}\mu(t).$$

This, of course, is also a (Hermitian) positive definite differential quadratic form.

Daniel M. Roy (*Massachusetts Institute of Technology, Cambridge*)
It is my pleasure to comment on this paper by Mark Girolami and Ben Calderhead, which introduces what is likely to prove to be an important perspective on and method for Markov chain Monte Carlo simulation.

Girolami and Calderhead develop extensions to the Hamiltonian Monte Carlo algorithm and the Metropolis adjusted Langevin algorithm (MALA) that can be used to take advantage of the intrinsic geometric structure of the parameter space of a probability model. They argue convincingly through experiments that, by using the local geometry induced by the Fisher information metric, their methods produce superior results and sidestep many of the tuning problems that undermine the widespread adoption of Hamiltonian Monte Carlo and Metropolis adjusted Langevin algorithm methods.

There are several significant barriers to the adoption of these methods. First, analytic expressions for the (expected Fisher) information matrix do not exist for most models of interest. Although other metrics can be used, only the Fisher information metric is evaluated. Natural alternatives, such as the observed (Fisher) information, suffer from apparent flaws. For example, the use of the observed information at the maximum likelihood estimate would not exploit local structure and would not make sense in multimodal settings where the geometry varies considerably across modes. Given the encouraging experimental work, an important line of inquiry will be to identify alternative metrics or approximations thereof. If we are

willing to ignore the effects of local curvature in the manifold, then sample-based approximations of the local intrinsic geometry seem potentially fruitful.

Parameter rich models with thousands (if not millions) of parameters are among the most interesting as well as challenging models in use today. In these settings, the computational complexity of the methods proposed appears to be a formidable barrier. For k parameters, the worst case computational complexity of the methods are $O(k^3)$, owing to the need to compute the solutions to linear systems involving the information matrix. In the case where this matrix has special structure, fast solvers (e.g. Koutis *et al.* (2010)) can be used (as demonstrated by the authors in the stochastic volatility model). The use of iterative methods for solving linear systems (see, for example Hestenes and Stiefel (1952) and Saad (2003)), especially those that work by refining an initial estimate (which could be computed from earlier proposals), would provide a speed–accuracy knob that is lacking in exact methods. Understanding how these approximations ultimately affect the stationary distribution will be critical.

Tim Salimans (*Erasmus University Rotterdam*)

The authors are to be congratulated for making an important contribution in the development of practical and efficient Monte Carlo strategies for Bayesian inference. The ability of the proposed methods to deal with very high dimensional latent variable models is particularly exciting. In fact, the performance of these methods may be even better than advertised: in their discussion of the stochastic volatility example the authors write

> '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'.

They use the resulting Cholesky factor to sample from $\mathcal{N}\{0, G^{-1}(\theta)\}$ and to compute the determinant of $G(\theta)$. However, this costly procedure may be avoided by noting that $\mathcal{N}\{0, G^{-1}(\theta)\}$ is the posterior distribution of the state vector s in the dynamic model

$$s_{t+1} = \phi s_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \qquad s_1 \sim \mathcal{N}\{0, \sigma^2/(1-\phi^2)\},$$
$$z_t = s_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, 2)$$

where ϕ and σ^2 are the parameters of the original model and where we take all pseudo-observations z_t equal to zero. Sampling from $p(s|z) = \mathcal{N}\{0, G^{-1}(\theta)\}$ and calculating the determinant of $G(\theta)$ can now easily be done, by applying the Kalman filter and smoother (see Durbin and Koopman (200l)). Implementing this procedure, I find that the manifold Metropolis adjusted Langevin algorithm is now 25 times faster for the stochastic volatility example, which makes it competitive with the other strategies discussed for this application.

D. Schmidl and F. J. Theis (*Technische Universität München and Helmholtz Zentrum München*)

We congratulate the authors on introducing two interesting new Markov chain Monte Carlo (MCMC) sampling concepts and thoroughly evaluating them on a variety of examples. Exploitation of the likelihood's local geometry helps the manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo algorithm to improve drastically both the acceptance rate and the effective sampling size—an issue at the very heart of every MCMC sampling method. We are wondering about the application of the proposed sampling concepts on Riemannian manifolds, where the geometric structure is already well known, in particular in fields other than classical parameter inference: recent work has been applying Bayesian sampling methods to matrix factorization (MF) problems such as independent component analysis (Roberts and Choudrey, 2005) and non-negative MF (Schmidt and Laurberg, 2008; Zhong and Girolami, 2009). The high dimensionality of the parameter space as well as inherent indeterminacies are especially a major challenge to overcome for efficient MCMC sampling. Sparseness constraints on the matrices can additionally increase the problem of computing the posterior owing to a heterogenic mass distribution of the likelihood function.

In particular, all these approaches share the fact that the inference problem is defined on matrix manifolds, which may be given a Riemannian geometric structure compatible with, for example, the matrix multiplication operation. The resulting geometry does not equal the standard Euclidean one (Edelman *et al.*, 1999). To improve the convergence of classical non-Bayesian MF algorithms, this non-Euclidean structure can be efficiently used by differential geometric optimization methods (Amari *et al.*, 2000; Cardoso and Laheld, 1996; Squartini *et al.*, 2006; Theis, 2005) that exploit the intrinsic manifold structure of the problem. A similar approach is based on the geometric conjugate gradient method (Absil *et al.*, 2008). Here, the convergence of the optimization procedure is improved by working on the Riemannian

matrix manifold by using a vector transport T and its associated retraction R, such that the conjugacy requirement for the search directions is, in some sense, transported along T. It is worth mentioning that all the components of the classical conjugate gradient method can straightforwardly be lifted onto the Riemannian manifold by means of the Riemannian metric.

Thus, using knowledge about the local geometry might be a promising *ansatz* to help to overcome the aforementioned issue of efficient sampling from high dimensional spiky likelihoods. In particular, we are wondering how the computational complexity improves when applying the manifold Metropolis adjusted Langevin algorithm and Riemann manifold Hamiltonian Monte Carlo algorithm to Bayesian MF methods compared with alternative sampling methods.

Giorgos Sermaidis (*Lancaster University*)

The Metropolis adjusted Langevin algorithm is based on recognizing that the target density $p(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^D$, is the invariant distribution of a Langevin diffusion solution process to

$$\mathrm{d}\boldsymbol{\theta}(t) = \nabla_\theta \mathcal{A}\{\boldsymbol{\theta}(t)\}\,\mathrm{d}t + \mathrm{d}\mathbf{b}(t), \qquad t \geqslant 0, \tag{50}$$

where $\mathcal{A}(\cdot) := \mathcal{L}(\cdot)/2$ and $\mathbf{b}$ is standard D-dimensional Brownian motion. Since expression (50) cannot be solved analytically the Metropolis adjusted Langevin algorithm proposes values by using a first-order Euler discretization of integration step size $\varepsilon > 0$, thus inducing a trade-off between the quality of the proposal and the speed at which the state space is explored; small values of ε result in more accurate approximations to the true diffusion dynamics but at the expense of small jumps in the state space and conversely.

Nevertheless, the discretization error can be avoided and one can simulate exact draws from expression (50) by using a rejection sampling mechanism, which is known as the exact algorithm (Beskos *et al.*, 2006, 2008). In particular, let $\mathbb{W}$ denote the law of Brownian motion starting at $\boldsymbol{\theta}(0)$ and suppose that we wish to simulate the process $\boldsymbol{\theta}_T := \{\boldsymbol{\theta}(s), s \in [0, T]\}$. Assume that $h(u) := \exp\{\mathcal{A}(u) - \|u - \boldsymbol{\theta}(0)\|^2/2T\}$ is Lebesgue integrable, where $\|\cdot\|$ is the Euclidean norm, and define the so-called biased Wiener measure $\mathbb{Z}$ by $\mathrm{d}\mathbb{Z}/\mathrm{d}\mathbb{W}(\boldsymbol{\theta}_T) = \exp[\mathcal{A}\{\boldsymbol{\theta}(T)\}]$. Sampling from the finite dimensional distributions of $\mathbb{Z}$ is achieved by first sampling $\boldsymbol{\theta}(T)$ from a density proportional to $h\{\boldsymbol{\theta}(T)\}$ and filling in the remainder of the path by Brownian bridge interpolations.

Under mild assumptions, we can write the density of the law of $\boldsymbol{\theta}_T$, $\mathbb{Q}$ with respect to the law of $\mathbb{Z}$ using Girsanov's theorem as

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{Z}}(\boldsymbol{\theta}_T) \propto \exp\left[-\int_0^T \phi\{\boldsymbol{\theta}(s)\}\,\mathrm{d}s\right] \leqslant 1, \tag{51}$$

where $\phi(\boldsymbol{\theta}) := \{\|\nabla_\theta \mathcal{A}(\boldsymbol{\theta})\|^2 + \Delta_\theta \mathcal{A}(\boldsymbol{\theta})\}/2 - l$, $l \leqslant \inf_\theta\{\|\nabla_\theta \mathcal{A}(\boldsymbol{\theta})\|^2 + \Delta_\theta \mathcal{A}(\boldsymbol{\theta})\}/2$, and Δ is the Laplacian operator. Expression (51) suggests a rejection sampler where proposed paths are drawn from $\mathbb{Z}$. The ratio cannot be evaluated analytically but is recognized as the probability of zero events from an inhomogeneous Poisson process of intensity $\phi\{\boldsymbol{\theta}(s)\}$ on $[0,T]$, and thus rejection sampling can be implemented by using Poisson thinning. If $M(\boldsymbol{\theta}_T)$ is a finite dimensional random variable and r is a positive function such that $r\{M(\boldsymbol{\theta}_T)\} = r(\boldsymbol{\theta}_T) \geqslant \sup_{0 \leqslant s \leqslant T}[\phi\{\boldsymbol{\theta}(s)\}]$, then the event

$$I := \prod_{j=1}^{k} \mathbb{I}\left[\frac{\phi\{\boldsymbol{\theta}(\psi_j)\}}{r(\boldsymbol{\theta}_T)} < u_j\right],$$

where $k \sim \mathrm{Po}\{r(\boldsymbol{\theta}_T)T\}$, $u_j \sim \mathrm{Un}(0, 1)$ and $\psi_j \sim \mathrm{Un}(0, T)$, occurs with probability equal to expression (51).

The form of expression (51) implies that the acceptance probability of the algorithm decreases exponentially with T and D. However, owing to the Markov property of expression (50), the cost can be made linear in both by splitting the interval $[0,T]$ into $\mathcal{O}(TD)$ subintervals and employing the algorithm sequentially. The applicability of the algorithm is limited by how easily we can sample from $h\{\boldsymbol{\theta}(T)\}$, which generally can be as difficult as sampling from $p(\boldsymbol{\theta})$ (see Peluchetti and Roberts (2008) for classes of h where sampling can be performed efficiently).

Anuj Srivastava (*Florida State University, Tallahassee*)

Firstly, I congratulate Professor Girolami and Dr Calderhead for their excellent paper on Langevin and Hamiltonian Monte Carlo methods on spaces that form non-linear Riemannian manifolds. I believe that their work is both timely and very useful, for there is a growing need for tools for efficiently generating statistical inferences on manifolds that are increasingly cropping up in signal and image processing.

Secondly, to put a historical perspective on their effort, I wish to draw their attention to some work by Grenander's school on pattern theory. A specific example involves the use of estimation and tracking of principal subspaces of the received signal in an array signal processing environment. Srivastava (2000) formulated this as a problem of Bayesian inference on Grassmann manifolds, the manifold of all d-dimensional subspaces of an n-dimensional Euclidean space, and generated posterior samples by using a Metropolis adjusted Langevin algorithm. More specfically, they used a Markov chain Monte Carlo process $X(t)$ in which the candidates for $X(t+1)$, given $X(t)$, were generated by using stochastic gradients of the log-probability function and accepted or rejected by using an appropriate function. This function allowed the convergence results for the Metropolis–Hastings algorithm to be applicable in this set-up. Similar ideas were explored for some matrix Lie groups and their quotient spaces in a set of companion papers (Grenander *et al.*, 1998; Srivastava and Klassen, 2001). This general framework represents a modification of the original Grenander–Miller ideas on the use of jump diffusion processes on Lie groups (Grenander and Miller, 1994; Srivastava *et al.*, 2002) for sampling, where, in particular, the diffusions were implemented via the Langevin equation.

Girolami and Calderhead extend the use of the Metropolis adjusted Langevin equation to more general Riemannian manifolds, especially under arbitrary Riemannian metrics (see equation (10)). In particular, they use a natural Riemannian metric, namely the Fisher–Rao metric, for inferences associated with some parametric families.

David A. Stephens (*McGill University, Montreal*)

I congratulate the authors on their elegant presentation of an ingenious Markov chain Monte Carlo (MCMC) method. It seems that the current state of received wisdom on recommendations for the application of MCMC methods in Bayesian inference encourages the user to be *adaptive*, and to adopt a *global* (across the state space) rather than local updating strategy. The automatic tuning of Riemann manifold Hamiltonian Monte Carlo proposals dictated by the geometry of the statistical model is clearly attractive from a practical point of view. However, it is not clear to me that the automatic procedure is necessarily the optimal strategy for global exploration of the posterior, i.e. that convergence rates are optimal for precisely this kind of update. Can the authors give some theoretical reassurance on this point to expand on the empirical evidence given in the paper, and the indications given in Section 6 (page 135) and in the discussion?

The justification for the Riemann manifold Hamiltonian Monte Carlo proposals in equation (13) (page 132) is via a Hamiltonian including a kinetic energy term. Is it possible or desirable to allow the form of this latter term to be any negative log-density involving the same metric tensor? In the context of Hamiltonian dynamics, it may not be particularly sensible, but for MCMC purposes there may be advantages in doing this. The Gibbs sampler in equations (19) and (20) could certainly be implemented in a more general set-up.

The test case in Section 7 of the paper also raises the question of model or variable selection, which is an inherently discrete problem. Do the methods proposed encounter difficulties in discrete state space problems?, or if the independent Gaussian prior $\beta \sim \mathcal{N}(0, \alpha \mathbf{I})$ is replaced by an independent Laplace prior akin to the formulation of the L_1-penalized regression problem?

Finally, efficient global exploration of the posterior is often achieved by using population MCMC sampling and the exchange of information across chains. Adaptation and population methods are briefly mentioned in isolation in the paper; is it desirable to utilize the ideas in combination, and does Hamiltonian Monte Carlo sampling offer any leverage in this respect? For example, is it possible to learn about the optimal βs in the 'simple scaling' in the examples in Section 10 of the paper?

M. K. Titsias (*University of Manchester*)

The stochastic volatility and log-Gaussian Cox process models are examples of latent Gaussian models with joint density

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}),$$

where $p(\mathbf{y}|\mathbf{x})$ is a non-Gaussian likelihood and $p(\mathbf{x})$ a Gaussian prior distribution over the latent vector $\mathbf{x}$. Here, we wish to discuss only sampling the latent vector. A potential limitation of the algorithms proposed is that they require the first and second derivatives of the full joint density. This is restrictive because in certain applications we may need to deal with limited information regarding the geometry of the likelihood. For instance, the use of second derivatives of $\log\{p(\mathbf{y}|\mathbf{x})\}$ can often be undesirable because of high computational cost. In contrast, the full information geometry of the Gaussian prior can always be taken into account.

Consider the proposal distribution

$$Q(\mathbf{x}'|\mathbf{x}) \propto H(\mathbf{x}', \mathbf{x})\, p(\mathbf{x}'),$$

which proposes a new $\mathbf{x}'$ given the current $\mathbf{x}$. $H(\mathbf{x}', \mathbf{x})$ is such that its logarithm is quadratic in $\mathbf{x}'$; thus $Q(\mathbf{x}'|\mathbf{x})$ is Gaussian. By construction the proposal distribution is invariant to the Gaussian prior. $H(\mathbf{x}', \mathbf{x})$ should be set to incorporate some properties of the non-Gaussian likelihood $p(\mathbf{y}|\mathbf{x})$. Auxiliary variables can be employed for such construction. The idea is to approximate the non-Gaussian likelihood by an auxiliary Gaussian likelihood $p(\mathbf{z}|\mathbf{x})$ where $\mathbf{z}$ are auxiliary variables that can be regarded as pseudodata. The

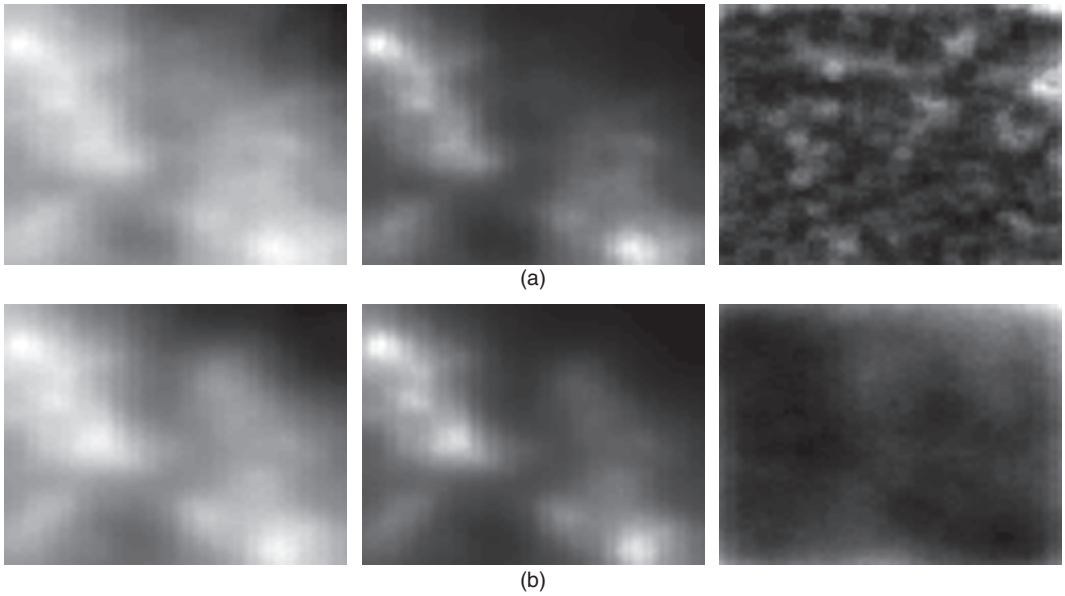

Fig. 22. Posterior Monte Carlo estimates obtained by (a) the GSRWM algorithm and (b) the GSMALA: the format of the figure and the experiment follows exactly Fig. 8; the results are obtained from 5000 posterior samples after a burn-in period in which the scale σ^2 of the proposal distribution was adapted to achieve a certain rate of acceptance; for the GSRWM algorithm, σ^2 was tuned to achieve a rate of acceptance of between 20% and 30%, whereas for the GSMALA the range was 50–60%; note that the GSMALA provides quite satisfactory results

Table 15. Effective sample sizes of the GSRWM algorithm and GSMALA on the log-Gaussian Cox point process example†

Method	*Time (s)*	*Effective sample size (minimum, median, maximum)*	*s/minimum effective sample size*
GSRWM	1311	(4, 29, 109)	327.7
GSMALA	942	(36, 205, 524)	26.1

†Note that the GSMALA that uses gradient information about the log-likelihood has significantly better performance than the GSRWM algorithm and also slightly outperforms the MMALA (see Table 10). Running times include also the adaptive phase that tunes σ^2. The GSRWM algorithm had a larger running time since it required a longer adaptive phase.

simplest choice is to use $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}\{\mathbf{z}|\mathbf{x}, (\sigma^2/2)I\}$ which says that $\mathbf{z}$ is a noisy version of $\mathbf{x}$. The sampling scheme iterates between updating $\mathbf{z}$ and $\mathbf{x}$ according to

(a) $\mathbf{z} = \mathbf{x} + (\sigma/\sqrt{2})\eta, \eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and
(b) $\mathbf{x}' \sim \{1/\mathcal{Z}(\mathbf{z})\}\,\mathcal{N}\{\mathbf{z}|\mathbf{x}', (\sigma^2/2)I\}\,p(\mathbf{x}')$ and accept or reject by using Metropolis–Hasting steps.

This iteration leaves $p(\mathbf{x}|\mathbf{y})$ invariant and implies a symmetric form for $H(\mathbf{x}', \mathbf{x})$, i.e. $H(\mathbf{x}', \mathbf{x}) = H(\mathbf{x}, \mathbf{x}')$. When the variance of the Gaussian $p(\mathbf{x})$ tends to ∞, step (b) reduces to $\mathbf{x}' = \mathbf{z} + (\sigma/\sqrt{2})\eta$ and both steps combined yield $\mathbf{x}' = \mathbf{x} + \sigma\eta$. The above algorithm can be thought of as a *Gaussian scaled random-walk Metropolis (GSRWM) algorithm*. Similarly, we can obtain a *Gaussian scaled Metropolis adjusted Langevin algorithm (GSMALA)* by sampling $\mathbf{z}$ in step (a) (while keeping (b) unchanged) according to

$$\mathbf{z} = \mathbf{x} + \frac{\sigma^2}{2}\nabla_{\mathbf{x}}\log\{p(\mathbf{y}|\mathbf{x})\} + \frac{\sigma}{\sqrt{2}}\eta.$$

This implies that the auxiliary likelihood is now

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left[\mathbf{z}|\mathbf{x} + \frac{\sigma^2}{2}\nabla_{\mathbf{x}}\log\{p(\mathbf{y}|\mathbf{x})\}, \frac{\sigma^2}{2}I\right]$$

and the scheme reduces to a standard Metropolis adjusted Langevin algorithm when the variance of $p(\mathbf{x})$ approaches ∞. When second derivatives of $\log\{p(\mathbf{y}|\mathbf{x})\}$ are easy to compute, further algorithms can be obtained by following the above framework. Preliminary results by using the GSRWM algorithm and GSMALA on the log-Gaussian Cox point process example are shown in Fig. 22 and Table 15.

Mark K. Transtrum, Yanjiun Chen, Benjamin B. Machta and James P. Sethna (*Cornell University, Ithaca*) **and Ryan Gutenkunst** (*University of Arizona, Tucson*)

The methods proposed by Girolami and Calderhead are particularly germane to error analysis in multiparameter non-linear fits to data. These fits are typically *sloppy* (Brown and Sethna, 2003; Brown, 2003; Brown *et al.*, 2004; Gutenkunst *et al.*, 2007; Daniels *et al.*, 2008), with a few well-determined 'stiff'

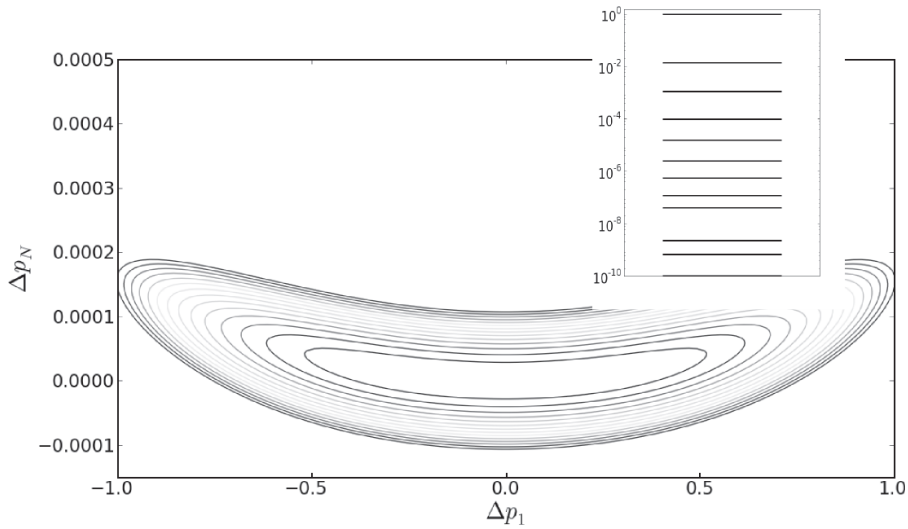

Fig. 23. Contour plot of χ^2 along the second stiffest and the sloppiest eigenvectors of the Fisher information matrix for a model of a sum of eight exponentials: note the 10^4-difference in the horizontal and vertical scales (the square root of the eigenvalue ratio); eigenvalues of the Fisher information metric (inset) at the best fit span 10 orders of magnitude; by sampling based on the Fisher information matrix at the best fit (the black curve in Fig. 24), steps are magnified along the sloppy directions, improving acceptance; however, the sloppy directions slowly rotate, reducing the acceptance ratio when the canyon no longer aligns with the eigendirections of the Fisher information; by stepping according to the local Fisher information at each step the convergence is dramatically improved (the medium grey curve in Fig. 24); we are working to incorporate geodesic improvements in this method (Transtrum, Chen and Sethna, 2010)

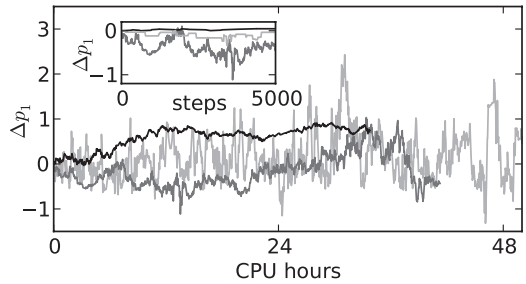

Fig. 24. Comparison of a geometrically adjusted importance sampling method (the simplified MMALA without the cost–gradient term (Gutenkunst *et al.*, 2007; Gutenkunst, 2008)) (medium grey), against two static importance sampling methods, one (black) using the metric at the best fit and one (light grey) tuned by using a principal component analysis of the final probability distribution from the black sampling: here we sample from a systems biology model of nerve growth factor signalling (Brown *et al.*, 2004); it can be seen that the geometrical method is much improved over importance sampling using only local information and a vast improvement over isotropic explorations; the globally tuned static method is faster in central processor unit time, but less efficient in number of steps (inset) than the geometrically adjusted method

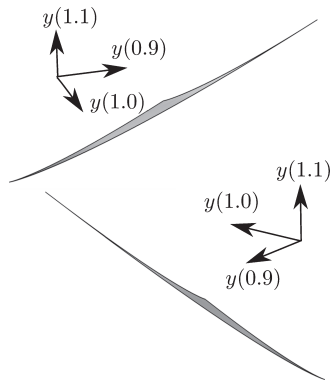

Fig. 25. The model manifold in data space takes the form of a hyper-ribbon, with a hierarchy of widths: in this simple model manifold corresponding to a model of two exponentials, $y(t) = \exp(-\theta_1 t) + \exp(-\theta_2 t)$ sampled at three nearby time points, there is a long and a thin direction, corresponding to the stiff and sloppy directions in parameter space; in higher dimensions, the hierarchy of widths is more dramatic, with the stiff directions several thousand times longer than the sloppy directions; the extrinsic curvature of the model manifold is small compared with the widths

parameter combinations and many sloppy parameter directions. These sloppy direction are thousands of times less well constrained by the data, demanding millions of times more Monte Carlo steps without clever importance sampling. Figs 23 and 24 illustrate how rotations of the sloppy directions forced us to develop a method equivalent to a simplified manifold Metropolis adjusted Langevin algorithm (MMALA) (Gutenkunst *et al.*, 2007; Gutenkunst, 2008).

This sloppiness is reflected in the geometry of what we call the *model manifold*—the surface of model predictions in data space, with co-ordinates given by the parameters (Transtrum *et al.*, 2010a, b). We find that it generically takes the form of a *hyper-ribbon* (Fig. 25), with widths along different axes spanning several orders of magnitude—narrow widths corresponding to sloppy directions of the fits. We use theorems from interpolation convergence to explain this structure, and also to explain the observation that the intrinsic (Riemannian) curvature and the extrinsic curvatures are many orders of magnitude smaller than those generated by the parameter effects curvature (Bates and Watts, 1980, 1998; Bates *et al.*, 1983). Avoiding the latter by approximating geodesic co-ordinates allows us to improve algorithms to find best fits (Transtrum *et al.*, 2010a, b). We conjecture that the small intrinsic curvatures also explain why the authors' parameter-independent Riemann manifold Hamiltonian Monte Carlo algorithm outperforms the MMALA algorithm.

One should note the boundaries of the model manifold (Fig. 25); as the parameters range to ∞ or other limiting values, the predictions of non-linear models often remain finite. Unless we are careful, an efficient stochastic sampling of parameters will often suffer from *parameter evaporation*; with a flat prior, many parameters will drift to ∞ where the entropy gain overwhelms any finite cost due to poor fits. We are currently studying geometrically natural, parameter-independent choices for priors (Machta *et al.*, 2010) (generalizations of the rather singular Jeffreys prior (Jeffreys, 1998).)

Aki Vehtari (*Aalto University*) **and Jarno Vanhatalo** (*University of Helsinki*)

We have used Hamiltonian Monte Carlo (HMC) methods in many problems and, although it has been the best of the available methods, we have hoped for a better one. Although Riemann manifold HMC (RMHMC) methods did not solve our problems, the paper makes many issues clearer, for which we thank the authors.

For Gaussian processes, we have used transformation, which is close to what expected Fisher information provides. Instead of taking the expectation with respect to the prior, we use the mode of the prior (Vanhatalo and Vehtari, 2007). For Poisson models the difference is using $m_i \exp(\mu)$ instead of $m \exp\{\mu + (\Sigma)_{ii}\}$. The HMC approach that was used in Vanhatalo and Vehtari (2007) is then almost the same as RMHMC sampling in Section 9.

For efficient sampling the curvature near the posterior mode is relevant. In the spatial epidemiology data that we analysed there is large relative variation in x_i and, when a model does considerable smoothing, $m_i \exp(\mu)$ is closer to the mode of the posterior than the x_i that were used by Christensen *et al.* (2005).

In our experiments, the conditional sampling of the latent values and conditional sampling of the hyperparameters are both fast, but still sampling from the joint distribution is slow owing to a strong posterior dependence between them. When we read the earlier version of the present paper, we thought that the problem had been solved. Alas, after asking the authors about the joint sampling, they replied (what they report also in the present paper) that the joint sampling is slowed down because of the inversion of the large tensor matrix during each fixed point iteration in the integrator. Furthermore, since the expected Fisher information matrix is block diagonal, it does not include information about the dependences between the latent values and hyperparameters, reducing expectations on improvements in performance.

The authors mention that the sparse approaches that were presented by Vanhatalo and Vehtari (2007) may provide further computational efficiency. We think that it would probably be quite easy for latent value sampling. In the joint case, the tensor being block diagonal helps, but even with the sparse approach the need to invert a tensor at every fixed point iteration might slow the method too much.

Although RMHMC sampling will probably be useful in many cases, it does not remove the need to think about faster alternatives such as deterministic approximations. For example, we obtained a speed-up of several order of magnitude by using Laplace and expectation propagation methods instead of (RM)HMC for log-Gaussian processes in Vanhatalo *et al.* (2010).

Max Welling (*University of California at Irvine*)

Let me begin by congratulating the authors on their insightful paper.

The relevant unit to evaluate Markov chain Monte Carlo methods is the 'amount of mixing per unit of computation'. This is indeed what the authors measure *after* discarding the samples obtained during 'burn-in'. However, is it fair to discard this initial phase in evaluating a statistical inference method? Figs 1 and 2 visualize fast mixing behaviour by showing some sample trajectories. It is the burn-in phase that most convincingly illustrates the power of the new methods.

Now, the burn-in phase can be viewed as 'optimization with detailed balance' and, indeed, there are branches of machine learning and statistics that are solely concerned with optimization. However, it seems a waste to insist on detailed balance during burn-in. In this respect it is interesting to note that the 'simplified manifold Metropolis adjusted Langevin algorithm' is closely related to Newton methods and Fisher scoring if the extra ballast related to detailed balance is stripped off.

Let me take this one step further. How well would any Markov chain Monte Carlo method do in the face of a very large, perhaps infinite, data set? It would not make a single step, because it would need all the data to accept or reject a proposed move. So the rate at which we learn from our data is zero. It seems evident that we need methods that use only small (random) mini-batches of data to update the parameters and to increase this batch size during burn-in. Do we even need all the data when we start collecting samples? I would argue, not necessarily. It is well known that under suitable conditions parameters behave asymptotically normal. We do not need to use all the data to obtain a good estimate of the mean and

covariance structure of the posterior density and we can divide by N to scale the covariance appropriately. Asymptotic normality of parameters may in fact help to explain why the simplified manifold Metropolis adjusted Langevin algorithm was so successful in most experiments.

As a computer scientist I would like to ask the question: what can we learn from the methods proposed that can help us to solve the type of problems we are facing today, with millions to billions of data cases and models with thousands or millions of parameters? It seems we still have some distance to travel.

Ole Winther (*Technical University of Denmark, Lyngby*) **and Manfred Opper** (*Technical University of Berlin*)

Mark Girolami and Ben Calderhead's paper represents an impressive amount of effort. This paper may be the significant step towards the maturation of Langevin (gradient-based, Metropolis adjusted Langevin algorithm) and Hamiltonian (or hybrid) Monte Carlo methods for statistical inference. The paper contains several important contributions all described clearly and thoroughly illustrated. The contributions include incorporating Riemannian manifold structure in the Metropolis adjusted Langevin algorithm and Hamiltonian Monte Carlo algorithm using a local parameter-dependent metric tensor (e.g. a Bayesian Fisher matrix), generalizing the proposal mechanisms and deriving the necessary metric tensors (and derivatives) for a whole range of non-trivial examples. The log-Gaussian Cox process model is arguably the most striking example of how the methods proposed can overcome slow relaxation due to strong correlations.

The methods proposed will, like any Markov chain Monte Carlo method, not be the final algorithm as Markov chain Monte Carlo methods will always to some degree have to be tailored to the problem at hand. The methods proposed obviously have limitations in terms of possible intractability of the metric tensor and computational complexity However, the paper will surely inspire much additional work because it proposes a guiding principle of designing better proposals that take into account the correlation structure that is present in the model. It can thus serve as a starting point for approximate tractable low complexity methods.

It is an open question what metric is optimal. It is by no means obvious that the Fisher metric is always the best choice. Finally, examples where a method fails are always very instructive. The paper mentions multimodality and tempering as ways to overcome this problem. This hints that the methods proposed fail in this case. It would have been nice to see an example of this in a low dimensional bimodal problem.

The **authors** replied later, in writing, as follows.

We are most grateful to all who have contributed their comments on this paper. The overall discussion is positively brimming with exciting suggestions for further theoretical analysis, methodological advancement, algorithmic development and enhancement, as well as indications of what may be achieved in attacking emerging challenging applications. We can provide only a brief response to the many thoughtful, imaginative and valuable contributions.

Geometry and theory

Geometry

We note Eguchi's view that our proposal of exploiting geometric concepts in simulation-based statistics is more persuasive than for its adoption within statistical inference in general. Of course this is not new; previously Srivastava developed the Metropolis adjusted Langevin algorithm (MALA) on Grassmann manifolds, which is a specific instance of the more general Riemannian geometry adopted in this paper and is a good example of the potential of geometric ideas in Markov chain Monte Carlo (MCMC) research. Critchley, presumably while waiting for a bus, considers equivariance of the manifold Metropolis adjusted Langevin algorithm (MMALA) and Riemann manifold Hamiltonian Monte Carlo (RMHMC) algorithm and indeed it is Parry who demonstrates that they transform covariantly, a property that HMC methods do not enjoy. Critchley suggests consideration of the preferred point geometry, and this may be of further interest for MCMC application. For the example in Section 5.1 the preferred point metric and components of the connection, with respect to the maximum likelihood estimates, $\hat{\mu}$ and $\hat{\sigma}$, are given below, where $\Delta = \hat{\mu} - \mu$. Just how effective the preferred point geometry will be for simulation-based methods is a matter for on-going investigation:

$$\mathbf{g} = \begin{pmatrix} \dfrac{\hat{\sigma}^2}{\sigma^4} & \dfrac{2\hat{\sigma}^2\Delta}{\sigma^5} \\ \dfrac{2\hat{\sigma}^2\Delta}{\sigma^5} & \dfrac{4\hat{\sigma}^2\Delta^2}{\sigma^6} + \dfrac{2\hat{\sigma}^4}{\sigma^6} \end{pmatrix},$$

$$\partial_\mu \mathbf{g} = \begin{pmatrix} 0 & -\frac{2\hat{\sigma}^2}{\sigma^5} \\ -\frac{2\hat{\sigma}^2}{\sigma^5} & -\frac{8\hat{\sigma}^2\Delta}{\sigma^6} \end{pmatrix},$$

$$\partial_\sigma \mathbf{g} = \begin{pmatrix} -\frac{4\hat{\sigma}^2}{\sigma^5} & -\frac{10\hat{\sigma}^2\Delta}{\sigma^6} \\ -\frac{10\hat{\sigma}^2\Delta}{\sigma^6} & -\frac{24\hat{\sigma}^2\Delta^2}{\sigma^7} - \frac{12\hat{\sigma}^4}{\sigma^7} \end{pmatrix}.$$

Rao points out that a range of metrics based on more general divergence measures are also available such as the α-divergence suggested by Ikeda, and Jupp suggests an interesting route to designing Riemannian metrics based on yokes. The geometry goes beyond the metric tensor of course, and, although (implicitly) we use the Levi–Civita connection, Anaya-Izquierdo and Marriott point out that, for mixture models where the number of components K is also variable, the Amari -1 connection is natural. We wonder whether this might suggest a natural geometric structure for non-parametric Bayesian methods based on Dirichlet processes or reversible jump MCMC methods, a manifold Hamiltonian version of which is presented by Friel and Wyse. Furthermore Ikeda points out that the α-divergence induces non-Riemannian connections associated with the Riemannian metric. We consider that the study of the properties of these divergence measures in terms of the desirable characteristics of MCMC sampling (e.g. rate of convergence, geometric ergodicity and variance of Monte Carlo estimates) may then formalize the choice of geometries for MCMC algorithms. A striking example of the importance of this choice is given by Murray and Adams for the case of hierarchic Bayesian models. Both Ikeda and Critchley raise the question about the use of the dual (asymmetric) *e*- and *m*-connections and indeed this does suggest the development of non-reversible manifold MCMC sampling, as suggested by Mira and Haario, based on dual connections.

Dryden suggests that sampling close to the manifold boundary can present practical issues. This is highlighted quite forcibly by the examples provided by Stathopoulis and Filippone, and Transtrum, Gutenkunst, Chen and Sethna, although at present it is unclear how to alleviate this issue in a general way. The suggestion by Dryden of employing a number of candidate metrics which are then weighted and selected on the basis of certain features is certainly worthy of further investigation and is similar to the compound Hamiltonian suggested by Beffy and Robert, as well as Jasra and Singh.

Perhaps the most compelling practical case for the formal exploitation of local geometry in simulation-based inference is given by Transtrum, Gutenkunst, Chen and Sethna. They provide a superb illustration of the successful use of local Fisher information to improve proposal acceptance rates in models that they term *sloppy*. Schmidl and Theis present the challenge of performing sampling-based inference over matrix factorizations and it will be interesting to see how suitable the manifold-based methods are in this particular case.

Hamiltonian dynamics and geodesic equations

Cornebise and Bornn provide illustrations of RMHMC proposals where ridges are present in the sampling density. This is an excellent example of proposals that follow geometric structures such as geodesics. The fact that RMHMC Metropolis-within-Gibbs style proposals (as clarified by Holmes) follow the manifold geodesics is only briefly mentioned in the paper and we expand on this for clarity. The quadratic Hamiltonian $p_i p_j g^{ij}$ can be rewritten as $\ddot{\theta}^i + \Gamma^i_{kl}\dot{\theta}^k\dot{\theta}^l$ where Γ^i_{kl} denotes the Christoffel symbols linked to the metric g_{ij}; thus solving Hamilton's equations provides a solution of the geodesic equations. For the case where a potential field is included, i.e. $v(\boldsymbol{\theta}) + p_i p_j g^{ij}$, we define the metric $\tilde{g}_{ij} = g_{ij}\{h - v(\boldsymbol{\theta})\}$ where h is a constant value of the Hamiltonian. The Maupertuis principle states that the flows for $p_i p_j g^{-ij}$ and $v(\boldsymbol{\theta}) + p_i p_j \tilde{g}^{ij}$ coincide along energy level h. Therefore solving Hamilton's equations defined on the manifold with metric g_{ij} is equivalent to solving the geodesic equation $\ddot{\theta}^i + \tilde{\Gamma}^i_{kl}\dot{\theta}^k\dot{\theta}^l = 0$. Based on this the suggestion of Jupp to consider Hamiltonians as defined by yokes is something which we consider to be an area of research promise.

A cautionary note from Skilling certainly requires deeper consideration on the foundations of information geometry. The arguments presented regarding geodesics defining microscale roughness is something we have observed in complex *sloppy* models and demands further serious investigation.

Theory

It is most encouraging to see that preliminary analysis of the ergodicity and robustness of the MMALA by Łatuszyński, Roberts, Thiéry and Wolny suggests excellent theoretical properties. The full details of this analysis will be of importance in establishing the manifold methods within the MCMC literature. The theoretical analysis of the optimal tuning criteria for RMHMC algorithms will be challenging and

will prove to be more than a straightforward extension of the work of Beskos, Pillai, Roberts, Sanz-Serna and Stuart related to HMC sampling. We agree with Beskos and Stuart that the combination of their work on HMC sampling in Hilbert spaces and ours may prove to be very powerful for exceptionally high dimensional problems, e.g. of the order of 10^{20} as in their examples, and we note this may find interest in kernel-based models in statistical learning theory.

Higham asks about the conditions that would guarantee existence and uniqueness of solutions to the Langevin diffusion as well as its numerical solution. These details have been considered in detail in the previous work on the MALA by Roberts and collaborators. Finally we note that Sermaidis describes a scheme to simulate exact draws from a Langevin diffusion without recourse to accept–reject correction, something that Chin hints at when discussing higher order integrators.

Methodology

The discussion provides an enormous range of potential avenues for further methodological research and development with Gelman highlighting the utmost need and importance of such methodological advances in modern day applied statistics. Doucet, Jacob and Johansen suggest a very clever simulation approach for the MMALA where the score and Fisher information martix are intractable, with Golightly and Boys suggesting a somewhat similar scheme. Such strategies may be of great use in some demanding applications in physics, as pointed out by Chin, and it will be fascinating to see how this performs on a range of challenging applications such as the extensions of the stochastic volatility model suggested by Griffin, or the motor unit estimation problem of Pettitt.

It is particularly notable that the first practical generalization of RMHMC to reversible jump RMHMC sampling, so elegantly presented by Friel and Wyse, provides really promising early results. Suggestions to consider trans-dimensional manifold MCMC sampling also came from Cornebise and Peters, Pettitt, and Stephens, and interestingly the scheme of Friel and Wyse may be of importance in the types of hierarchic Bayesian models that are outlined by Gelman. Titsias presents an alternative MALA for Gaussian processes, and we wonder whether this can be extended to the MMALA. The possibility that manifold methods may have supergeometric convergence is suggested by Meng and this certainly deserves further consideration. Likewise the relationship with optimal path sampling and RMHMC geodesic-based proposals is worthy of exploration, as is the warping strategy. Both Sanz-Serna and Chin propose consideration of higher order integrators for the RMHMC algorithm and the MMALA, though at this stage it is difficult to assess their relative benefit given increased complexity. Fearnhead provides a suggestion of how to remove random-walk behaviour in the MALA, resulting in dynamics described by a hypoelliptic stochastic differential equation; however, the correctness of the resultant method needs to be confirmed.

Multimodal targets

An issue regarding *ripples* in the target density is discussed by Campbell who highlights that RMHMC sampling will not make moves tangent to the ripples whereas HMC sampling will, and the warped bivariate Gaussian (Cornebise and Bornn) is an excellent illustration of this. However, this leads onto the issue of how manifold methods are suited to sampling from multimodal distributions. This point is raised by Campbell, Stephens, Mira and Haario, Cao and Wang, Pettitt and Winther. It is Husmeier who gives a concrete example demonstrating that manifold methods themselves will fail in the face of multiple modes, but devising an appropriate mixture transition kernel can get around the issue. As suggested by Husmeier, we now give an example of manifold methods embedded within a population MCMC scheme.

We sample states $\mathbf{x}$ from a mixture of M Gaussians,

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{m}^{M} \frac{1}{M} \mathcal{N}_{\mathbf{x}}(\mu_m, \Sigma_m).$$

Denoting $R_{\mathbf{x}m} \propto (1/M)\mathcal{N}_{\mathbf{x}}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ with $\Sigma_m R_{\mathbf{x}m} = 1$, and $\mathbf{C}_{m,m'} = (\mathbf{x} - \boldsymbol{\mu}_m)(\mathbf{x} - \boldsymbol{\mu}_{m'})$, the full Hessian follows as

$$-R_{\mathbf{x}m} R_{\mathbf{x}m'} \boldsymbol{\Sigma}_m^{-1} \mathbf{C}_{m,m'} \boldsymbol{\Sigma}_{m'}^{-1} + R_{\mathbf{x}m} \boldsymbol{\Sigma}_m^{-1} \mathbf{C}_{m,m'} \boldsymbol{\Sigma}_m^{-1} - R_{\mathbf{x}m} \boldsymbol{\Sigma}_m^{-1}$$

where summation over m and m' is implicit. By assuming that the components are well separated such that for one m the $R_{\mathbf{x}m} \approx 1$ the observed Fisher information matrix then reduces to $\Sigma_m^M R_{\mathbf{x}m} \boldsymbol{\Sigma}_m^{-1}$. We first sample from a mixture of seven two-dimensional Gaussian distributions with varying means and covariances. We employ a tempering schedule with five temperatures distributed evenly between 0 and 1 then raised to the power 5 (Calderhead and Girolami, 2009). Fig. 26 shows 1000 samples collected from the posterior, after a burn-in period of 1000 iterations. The population structure allows all modes to be visited by both the

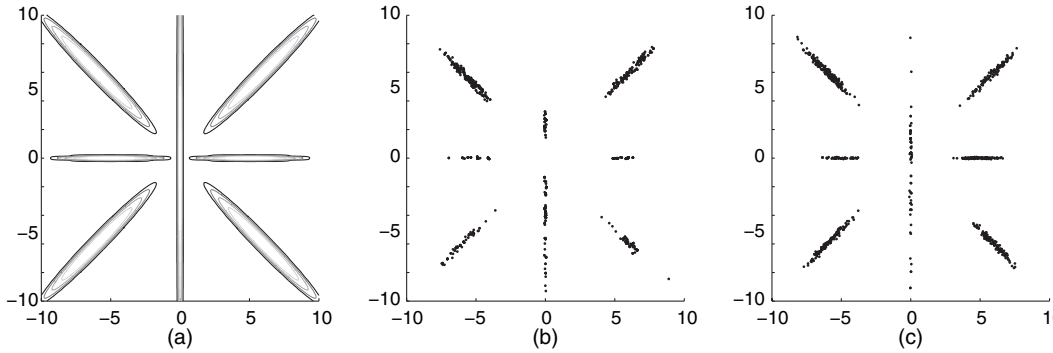

Fig. 26. (a) Contour plot of the target density and 1000 posterior samples from a mixture of seven two-dimensional Gaussian distributions with varying means and covariances by using (b) random-walk Metropolis sampling and (c) the simplified MMALA

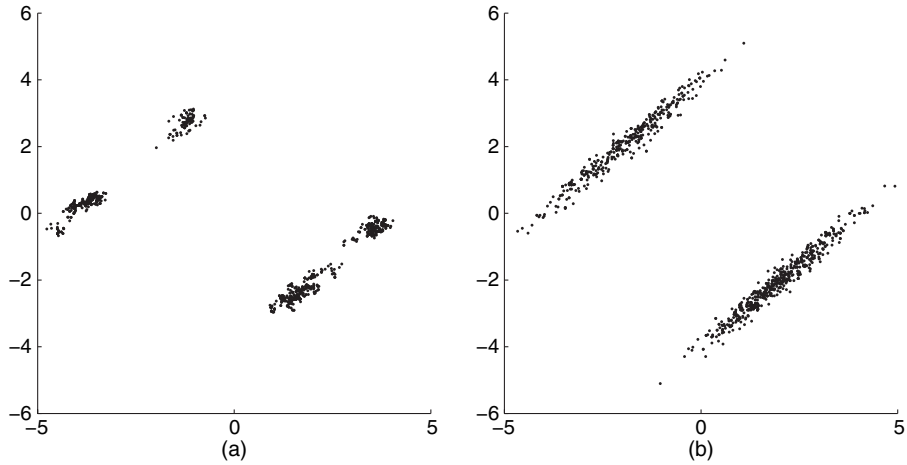

Fig. 27. 1000 posterior samples from the first two dimensions of a mixture of two strongly correlated 20-dimensional Gaussian distributions by using (a) random-walk Metropolis sampling and (b) the simplified MMALA

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 l, 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

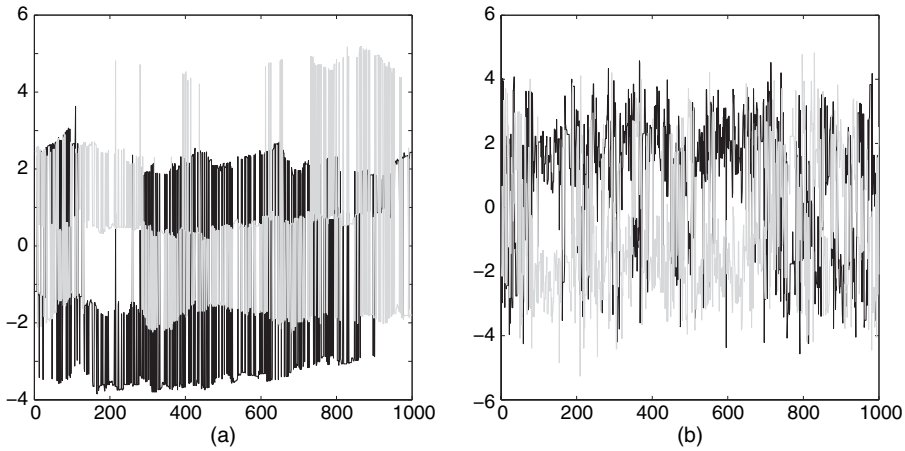

Fig. 28. Trace plots of 1000 posterior samples from the first two dimensions of a mixture of two strongly correlated 20-dimensional Gaussian distributions by using (a) random-walk Metropolis sampling and (b) the simplified MMALA

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 non-linear 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 Kontoyiannís, 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.