A General Metric for Riemannian Manifold Hamiltonian Monte Carlo

Michael Betancourt

Department of Statistics, Columbia University, New York, NY 10027, USA
betanalpha@gmail.com

Abstract. Markov Chain Monte Carlo (MCMC) is an invaluable means of inference with complicated models, and Hamiltonian Monte Carlo, in particular Riemannian Manifold Hamiltonian Monte Carlo (RMHMC), has demonstrated success in many challenging problems. Current RMHMC implementations, however, rely on a Riemannian metric that limits their application. In this paper I propose a new metric for RMHMC without these limitations and verify its success on a distribution that emulates many hierarchical and latent models.

Riemannian Manifold Hamiltonian Monte Carlo provides a powerful tool for the efficient sampling from complex distributions, but the applicability of existing approaches has been limited by the dependency on the Fisher-Rao metric. In this paper I introduce a new metric that admits a general implementation of Riemannian Manifold Hamiltonian Monte Carlo and demonstrate its efficacy on a distribution that mirrors the pathological behavior of common models.

1 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) takes advantage of symplectic geometry to yield efficient Markov transitions [3]. Augmenting an N-dimensional target density, $\pi(\mathbf{q})$, with corresponding momenta, $\mathbf{p}$, defines a joint density,

$$\begin{aligned}\pi(\mathbf{p}, \mathbf{q}) &= \pi(\mathbf{p}|\mathbf{q})\,\pi(\mathbf{q}) = \exp\left[\log \pi(\mathbf{p}|\mathbf{q})\right] \exp\left[\log \pi(\mathbf{q})\right] \\ &\propto \exp\left[-T(\mathbf{p}, \mathbf{q})\right] \exp\left[-V(\mathbf{q})\right] = \exp\left[-H(\mathbf{p}, \mathbf{q})\right].\end{aligned}$$

The Hamiltonian, $H(\mathbf{p}, \mathbf{q}) = T(\mathbf{p}, \mathbf{q}) + V(\mathbf{q})$, defines trajectories between points $\mathbf{z} = \{\mathbf{p}, \mathbf{q}\}$. Because these trajectories preserve the value of the Hamiltonian and the differential volume $\mathrm{d}^{2N}\mathbf{z}$, they also define Markovian transitions with the stationary density $\pi(\mathbf{p}, \mathbf{q})$. Alternating this Hamiltonian evolution with conditional samples of the momenta, $\mathbf{p} \sim \pi(\mathbf{p}|\mathbf{q}) \propto \exp\left[-T(\mathbf{p}, \mathbf{q})\right]$, yields an ergodic Markov chain sampling from $\mathbf{z}$ and, because the marginal of $\pi(\mathbf{p}, \mathbf{q})$ is constructed to be the target distribution, the desired samples from $\pi(\mathbf{q})$ follow by simply disregarding the momenta.

No matter the choice of the kinetic energy, $T(\mathbf{p}, \mathbf{q})$, the evolution equations incorporate the gradient of the potential, $V(\mathbf{q})$, and hence higher order information about the target distribution. This gradient guides the Markov chain along

F. Nielsen and F. Barbaresco (Eds.): GSI 2013, LNCS 8085, pp. 327–334, 2013.

regions of high probability mass and reduces random walk behavior. Note that, in practice, the Hamiltonian evolution cannot be performed analytically and we must resort to numerical integration. Error in the integration scheme introduces bias into the transitions, but this is readily avoided by considering the evolution not as a transition but rather as the proposal for a Metropolis transition [5,14].

The first [5] and still most common choice of the conditional density, $\pi(\mathbf{p}|\mathbf{q})$, is a standard gaussian, $\pi(\mathbf{p}|\mathbf{q}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \mathbf{M})$, or $T(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\mathbf{p}^T \cdot \mathbf{M}^{-1} \cdot \mathbf{p}$, where the mass matrix $\mathbf{M}$ allows for a global decorrelation and rescaling of the parameters with respect to each other. This choice, however, ultimately limits the effectiveness of HMC when applied to intricate target distributions. Because $\mathbf{p}^T \cdot \mathbf{M}^{-1} \cdot \mathbf{p}$ is a χ^2 variate, in equilibrium $\Delta T \approx N/2$ and, with the Hamiltonian conserved along each trajectory, this implies that the variation in the potential is also limited to $\Delta V \approx N/2$. When the target distribution is highly correlated, the typical set spans a potential gap much larger than this: the resulting samples become highly correlated no matter how long the trajectories are evolved [14] and the Markov chain devolves towards a random walk.

Another issue with the simple choice above is that the inevitable numerical integration introduces a spatial scale into the system via a finite step-size. Complicated target distributions will typically exhibit multiple spatial scales depending on the particular value of the parameters, and any single choice of a step-size will generate at least some inefficiency. If the step-size is chosen to maximize efficiency, as common in adaptive schemes, regions of the target distribution with large curvature, and hence small spatial scales, can be missed entirely by the numerical trajectories.

These weaknesses can be overcome by appealing to a more sophisticated choice of the conditional density: a gaussian conditionally dependent on the $\mathbf{q}$ through a covariance matrix, $\pi(\mathbf{p}|\mathbf{q}) = \mathcal{N}(\mathbf{p}|\mathbf{0}, \boldsymbol{\Sigma}(\mathbf{q}))$, or

$$T(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\mathbf{p}^T \cdot \boldsymbol{\Sigma}^{-1}(\mathbf{q}) \cdot \mathbf{p} + \frac{1}{2}\log|\boldsymbol{\Sigma}(\mathbf{q})|.$$

Because the resulting Hamiltonian trajectories are related to geodesics on a Riemannian manifold with metric $\boldsymbol{\Sigma}(\mathbf{q})$, this choice is known as *Riemannian Manifold Hamiltonian Monte Carlo* (RMHMC) [8]. Similarly, the constant metric introduced above can be thought of as emulating dynamics on a Euclidean manifold, and to be consistent I will refer to use of the simpler Hamiltonian as *Euclidean Manifold Hamiltonian Monte Carlo* (EMHMC).

The freedom in specifying a metric admits two significant improvements: a proper choice of $\boldsymbol{\Sigma}(\mathbf{q})$ can dynamically decorrelate and rescale the target distribution to avoid inefficiencies in the numerical integration, while also yielding a dynamic determinant whose variations can compensate for much larger variations in the potential.

What, however, exactly defines a proper choice for the metric? When the target distribution is a multivariate gaussian, $V(\mathbf{q}) = \frac{1}{2}\mathbf{q}^T \cdot \mathbf{S}^{-1} \cdot \mathbf{q}$, the target distribution is standarized by taking $\boldsymbol{\Sigma}(\mathbf{q}) = \mathbf{S}^{-1}$ [14]. In a convex neighborhood any target distribution can be approximated by a multivariate gaussian, $\pi(\mathbf{q}) \approx$

$\mathcal{N}\left(\mathbf{q}|\mathbf{0}, \mathbf{H}^{-1}\right)$ or, equivalently, $V(\mathbf{q}) \approx \frac{1}{2}\mathbf{q}^T \cdot \mathbf{H} \cdot \mathbf{q}$ with the Hessian matrix $H_{ij} = \partial^2 V / \partial q^i \partial q^j$, which immediately motivates the candidate metric $\boldsymbol{\Sigma}(\mathbf{q}) = \mathbf{H}$.

This metric quickly runs into problems, however, when the target distribution is not globally convex. In neighborhoods where the Hessian is not positive-definite, for example, the conditional density $\pi(\mathbf{p}|\mathbf{q})$ becomes improper. Moreover, in the neighborhoods where the signature of the Hessian changes, the log determinant diverges and the Hamiltonian evolution becomes singular. These neighborhoods effectively partition the support of the target distribution into a disjoint union of compact neighborhoods between which the Markov chain cannot transition.

One way to avoid indefinite metrics is to take advantage of any conditioning variables, $\mathbf{y}$, in the target distribution. Marginalizing the Hessian over these conditioning variables yields the Fisher-Rao metric [2], $\Sigma_{ij} = \mathbb{E}_{\mathbf{y}}\left[\partial^2 V(\mathbf{q}|\mathbf{y}) / \partial q^i \partial q^j\right]$, which is guaranteed to be positive-semidefinite. For all but the simplest conditional distributions, however, the marginalization is unfeasible and, even when it can be performed analytically, the resulting metric can still be singular. Moreover, the marginalization removes the correlation between variables in many hierarchical and latent models, almost eliminating the effectiveness of the metric. Of course, all of this is immaterial if the target distribution lacks natural conditioning variables.

We need a means of constructing a metric from the Hessian that is not only everywhere well-behaved but also practical to compute for any given target distribution.

2 The SoftAbs Metric

With a careful application of matrix functions, it is possible to maintain the desirable behavior of the Hessian in convex neighborhoods while avoiding its singular behavior elsewhere. Moreover, because the functions are local the resulting metric is readily implemented for general distributions.

2.1 Definition

The exponential map [15], exp, is a matrix function from the space of all matrices to the component of the general linear group, $\mathrm{GL}(n)$, connected to the identity matrix: an isomorphism of the space of positive-definite matrices. Because this mapping preserves the symmetric part of the domain, any symmetric matrix, such as the Hessian, is guaranteed to be mapped to a symmetric, positive-definite matrix admissible as a Riemannian metric.

One benefit of the exponential map is that it preserves the eigenbasis of the input matrix, $\mathbf{X}$. If $\mathbf{X} = \mathbf{Q} \cdot \boldsymbol{\lambda} \cdot \mathbf{Q}^T$ is the eigendecomposition of $\mathbf{X}$ with $\boldsymbol{\lambda} = \mathrm{Diag}\left(\lambda_i\right)$ the diagonal matrix of eigenvalues and $\mathbf{Q}$ the corresponding matrix of eigenvectors, then the exponential map yields $\exp\mathbf{X} = \mathbf{Q} \cdot \exp\boldsymbol{\lambda} \cdot \mathbf{Q}^T$. The metric $\exp\mathbf{H}$ provides the same decorrelation as the Hessian but also severely warps the eigenvalues and the corresponding rescaling of the local parameters.

By combining multiple exponential mappings, however, we can largely preserve the spectral decomposition of the Hessian. In particular, the *SoftAbs* map

$$\wr \mathbf{X} \wr \equiv [\exp(\alpha \mathbf{X}) + \exp(-\alpha \mathbf{X})] \cdot \mathbf{X} \cdot [\exp(\alpha \mathbf{X}) - \exp(-\alpha \mathbf{X})]^{-1}$$

approximates the absolute value of the eigenspectrum with a smooth function: $\wr \mathbf{X} \wr = \mathbf{Q} \cdot \wr \boldsymbol{\lambda} \wr \cdot \mathbf{Q}^T$, where

$$\wr \boldsymbol{\lambda} \wr = \text{Diag}\left(\lambda_i \frac{e^{\alpha \lambda_i} + e^{-\alpha \lambda_i}}{e^{\alpha \lambda_i} - e^{-\alpha \lambda_i}} \right) = \text{Diag}\left(\lambda_i \coth \alpha \lambda_i \right).$$

This map not only ensures that the transformed eigenvalues are positive but also regularizes any small eigenvalues that might introduce numerical instabilities.

Applying the SoftAbs map to the Hessian guarantees a well-behaved metric for RMHMC, $\wr \mathbf{H} \wr$, that preserves the desired properties of the Hessian while regularizing its numerical singularities. In a practical implementation, α limits the scaling of the integration step-size and restrains the numerical integrator from unwise extrapolations, emulating a trust region common in nonlinear optimization [4].

2.2 Implementation

In practice, exponential maps can be difficult to implement [11]; the eigendecomposition used above, for example, can suffer from numerical instabilities when applied to general matrices because of ambiguities among the eigenvectors. The Hessian, however, is symmetric and the eigenvectors are guaranteed to be orthogonal. Consequently, the eigendecomposition is well-behaved and provides a practical means of computing the SoftAbs map.

To implement the SoftAbs metric we first perform the eigendecomposition of the Hessian $\mathbf{H} = \mathbf{Q} \cdot \boldsymbol{\lambda} \cdot \mathbf{Q}^T$, and then reconstruct the metric as $\wr \mathbf{H} \wr = \mathbf{Q} \cdot \wr \boldsymbol{\lambda} \wr \cdot \mathbf{Q}^T$, with $\wr \boldsymbol{\lambda} \wr = \text{Diag}\left(\lambda_i \coth \alpha \lambda_i \right)$.

Hamiltonian evolution also requires two derivatives: the gradient of the quadratic form, $\mathbf{p}^T \cdot \wr \mathbf{H} \wr^{-1} \cdot \mathbf{p}$, and the log determinant, $\log |\wr \mathbf{H} \wr|$. The latter can be computed as [1,16]

$$\begin{aligned} \partial \left(\mathbf{p}^T \cdot \wr \mathbf{H} \wr^{-1} \cdot \mathbf{p} \right) &= \mathbf{p}^T \cdot \partial \wr \mathbf{H} \wr^{-1} \cdot \mathbf{p} \\ &= -\mathbf{p}^T \cdot \wr \mathbf{H} \wr^{-1} \cdot \partial \wr \mathbf{H} \wr \cdot \wr \mathbf{H} \wr^{-1} \cdot \mathbf{p} \\ &= -\left(\mathbf{Q}^T \cdot \mathbf{p} \right)^T \left[\mathbf{J} \circ \mathbf{Q}^T \cdot \partial \mathbf{H} \cdot \mathbf{Q} \right] \left(\mathbf{Q}^T \cdot \mathbf{p} \right), \end{aligned}$$

where $\circ$ denotes the Hadamard product and

$$J_{ij} \equiv \frac{\lambda_i \coth \alpha \lambda_i - \lambda_j \coth \alpha \lambda_j}{\lambda_i - \lambda_j}.$$

Note that when $\lambda_i = \lambda_j$, such as for the diagonal elements or degenerate eigenvalues, this becomes the derivative, $J_{ij} \to \partial / \partial \lambda_i \left(\lambda_i \coth \alpha \lambda_i \right)$.

Unfortunately, this form of the gradient is computationally inefficient, requiring $O(N^3)$ for each component of the gradient, and hence $O(N^4)$ overall. Taking advantage of the properties of the Hadamard product [10], however, the gradient can be manipulated to give

$$\partial \left(\mathbf{p}^T \lambda \mathbf{H} \lambda^{-1} \mathbf{p} \right) = -\mathrm{Tr} \left[\mathbf{Q} \cdot \mathbf{D} \cdot \mathbf{J} \cdot \mathbf{D} \cdot \mathbf{Q}^T \cdot \partial \mathbf{H} \right],$$

where $\mathbf{D} = \mathrm{Diag}\left(\left(\mathbf{Q}^T \cdot \mathbf{p} \right)_i \right)$. If the matrix $\mathbf{Q} \cdot \mathbf{D} \cdot \mathbf{J} \cdot \mathbf{D} \cdot \mathbf{Q}^T$ is first cached, then each component of the gradient can be computed in only $O(N^2)$ so that the complete gradient does not exceed the $O(N^3)$ complexity of the decomposition itself.

Similar Hadamard identities reduce the gradient of the log determinant to

$$\partial \log |\lambda \mathbf{H} \lambda| = \mathrm{Tr} \left[\mathbf{Q} \left(\mathbf{R} \circ \mathbf{J} \right) \mathbf{Q}^T \cdot \partial \mathbf{H} \right],$$

where $\mathbf{R} = \mathrm{Diag}\left(1 / \lambda_i \coth \alpha \lambda_i \right)$. Once again, caching the intermediate matrix, $\mathbf{Q} \left(\mathbf{R} \circ \mathbf{J} \right) \mathbf{Q}^T$, enables the full gradient to be computed in $O(N^3)$.

Table 1. When comparing the effective sample size of the latent variable, v, in the funnel distribution, hand-tuned EMHMC is over three times less effective than adaptively-tuned RMHMC. CPU time was measured with the `clock` function in the C++ library `time`.

Algorithm	Warm-Up Iterations	Samples	ϵ	Accept Rate	CPU Time (s)	ESS	ESS/Time ($\mathbf{s}^{-1}$)
EMHMC	10^3	10^5	0.001	0.999	1627	70.3	0.0432
RMHMC	10^3	10^3	0.21	0.946	6282	856	0.136

3 Experiments

The utility of the SoftAbs metric is best demonstrated on complex distributions. Neal's funnel distribution [13]

$$\pi(\mathbf{x}, v) = \prod_{i=1}^{n} \mathcal{N}\left(x_i | 0, e^{-v} \right) \cdot \mathcal{N}(v | 0, 9),$$

emulates many pathological features of popular distributions, such as those arising in hierarchical [6] and latent [12] models. Note that, by construction, the marginal distribution of v is simply $v \sim \mathcal{N}(0, 9)$[1], independent of n, admitting v and its marginal distribution as a simple diagnostic of bias in any sampling procedure.

In each experiment a Markov chain is randomly initialized, $q_i \sim U(-1, 1)$, and then taken through a series of warm-up iterations before sampling begins.

[1] Note the use of the convention $\mathcal{N}\left(\mu, \sigma^2 \right)$.

Where noted, the integrator step-size, ϵ, is adapted with dual-averaging to yield a target Metropolis acceptance rate [9]. The number of integration steps is set by hand to approximate the half-period of the oscillating trajectories.

Autocorrelations, ρ_i, of v are computed with an initial monotone sequence estimator [7] and the effective sample size (ESS) is defined as ESS $= I \left(1 + 2 \sum_{i=1}^{I} \rho_i\right)^{-1}$, where I is the total number of generated samples.

The above procedure is applied to EMHMC with step-size adaptation, EMHMC without step-size adaption, and RMHMC with the SoftAbs metric.

3.1 EMHMC with Adaptation

Despite its simplicity, the funnel demonstrates many of the limitations of EMHMC. When adaptively tuned to the nominal acceptance rate $r = 0.65$ [14], the integrator step-size exceeds the spatial scale of the narrow neck; even though the probability mass of the mouth and neck of the funnel is comparable, the resulting trajectories overlook the neck entirely and bias resulting expectations, empirically $v \sim \mathcal{N}(1, 4)$ which is inconsistent with the true marginal $v \sim \mathcal{N}(0, 9)$, without any obvious indication.

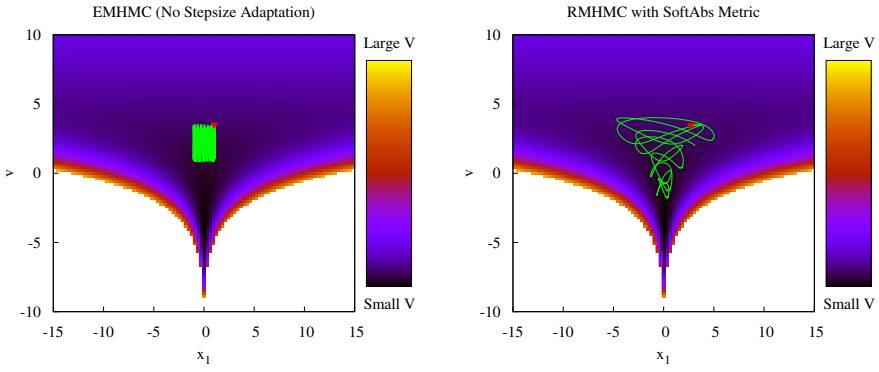
3.2 EMHMC without Adaptation

Because we know the truth in this case, we can abandon adaptive tuning and instead tune the step-size by hand; a smaller step-size ensures that the trajectories explore most of the funnel's probability mass and that the marginal distribution $p(v)$ is correct within Monte Carlo error. Unfortunately, the funnel also exhibits the limitations of a position-independent kinetic energy. The variation of the potential within the typical set is huge, and the meager variation of the kinetic energy dramatically restricts the distance of each transition (Figure 1). The EMHMC transitions struggle to cross between the mouth and neck of the funnel, and the Markov chain becomes little more than a random walk across the distribution (Figure 2).

3.3 RMHMC with the SoftAbs Metric

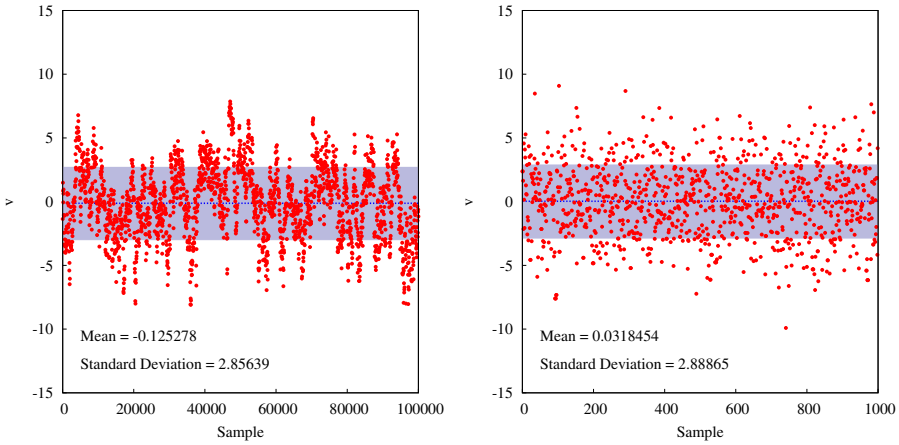
On the other hand, the SoftAbs metric, here with $\alpha = 10^6$, allows RMHMC to explore the entire distribution within a single trajectory (Figure 1). Because the metric accounts for local curvature, the step-size can be adaptively tuned[2] without introducing any bias. The huge autocorrelations of EMHMC vanish (Figure 2) and, despite the increased computation required for each transition, RMHMC yields a more efficient generation of effective samples (Table 1).

[2] The increased information encoded in the metric should admit a larger acceptance rate, r, for RMHMC than the EMHMC case of $r = 0.65$. Motivated by some simple experiments, here the target rate for RMHMC is set to $r = 0.95$.



(a) (b)

Fig. 1. While (a) EMHMC trajectories are limited to $\Delta V \sim (n+1)/2$ and consequently explore only a small neighborhood of the funnel, (b) RMHMC trajectories to explore the entire distribution and the dynamic decorrelation/scaling ensures that a single integrator step-size is efficient across the entire distribution.



(a) (b)

Fig. 2. Although samples of v from both (a) EMHMC and (b) RMHMC are consistent with the true marginal, $\mathcal{N}(0, 9)$, the far-reaching trajectories of RMHMC present little autocorrelation.

Nominally, the $O(N^3)$ computational burden of RMHMC is significantly worse than the $O(N)$ burden of EMHMC. The pathological behavior of distributions like the funnel, however, scales much faster, often exponentially, and the benefit of RMHMC with the SoftAbs metric only increases with dimension. Moreover, this concern ignores the burden of computing the potential itself

which, as in the case of Bayesian posteriors with many data, can overwhelm the $O(N^3)$ burden entirely.

4 Conclusions

By smoothly regularizing the eigendecomposition of the Hessian, the SoftAbs metric admits a general implementation of RMHMC robust against the many pathologies to which EMHMC can be vulnerable. Despite its apparently steep computational burden, the SoftAbs metric allows for practical inference on complex models never before deemed feasible.

Acknowledgements. I thank Bob Carpenter, Joe Formaggio, Mark Girolami, and Chris Jones for comments and suggestions, as well as the Stan team for their hospitality.

References

1. Aizu, K.: Parameter differentiation of quantum-mechanical linear operators. Journal of Mathematical Physics 4(6), 762–775 (1963)
2. Amari, S., Nagaoka, H.: Methods of information geometry, vol. 191. Amer. Mathematical Society (2007)
3. Betancourt, M., Stein, L.C.: The Geometry of Hamiltonian Monte Carlo (2011)
4. Celis, M., Dennis, J.E.: A., T.R.: A trust region strategy for nonlinear equality constrained optimization. In: Boggs, P., Byrd, R., Schnabel, R. (eds.) Numerical Optimization 1984, SIAM, Philadelphia (1985)
5. Duane, S., Kennedy, A., Pendleton, B.J., Roweth, D.: Hybrid monte carlo. Physics Letters B 195(2), 216–222 (1987)
6. Gelman, A., Carlin, J., Stern, H., Rubin, D.: Bayesian Data Analysis. Chapman & Hall/CRC Press, Boca Raton (2004)
7. Geyer, C.: Introduction to mcmc. In: Brooks, S., Gelman, A., Jones, G.L., Meng, X.L. (eds.) Handbook of MCMC. CRC Press, New York (2011)
8. Girolami, M., Calderhead, B.: Riemann manifold langevin and hamiltonian monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(2), 123–214 (2011)
9. Hoffman, M.D., Gelman, A.: The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. ArXiv e-prints (November 2011)
10. Magnus, J., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, New York (2007)
11. Moler, C., Van Loan, C.: Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Review 45(1), 3–49 (2003)
12. Murray, I., Prescott Adams, R.: Slice Sampling Covariance Hyperparameters of Latent Gaussian Models. ArXiv e-prints (June 2010)
13. Neal, R.: Slice sampling. Annals of Statistics, 705–741 (2003)
14. Neal, R.: Mcmc using hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G.L., Meng, X.L. (eds.) Handbook of MCMC. CRC Press, New York (2011)
15. Spivak, M.: A Comprehensive Introduction to Differential Geometry, vol. 1. Publish or Perish, Inc., Houston (2005)
16. Wilcox, R.M.: Exponential operators and parameter differentiation in quantum physics. Journal of Mathematical Physics 8(4), 962–982 (1967)