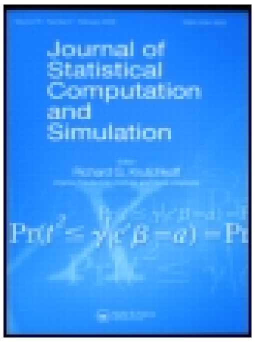

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

Geodesic Lagrangian Monte Carlo over the space of positive definite matrices: with application to Bayesian spectral density estimation

Andrew Holbrook, Shiwei Lan, Alexander Vandenberg-Rodes & Babak Shahbaba

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2017
https://doi.org/10.1080/00949655.2017.1416470

Taylor & Francis
Taylor & Francis Group

Geodesic Lagrangian Monte Carlo over the space of positive definite matrices: with application to Bayesian spectral density estimation

Andrew Holbrook[a], Shiwei Lan[b], Alexander Vandenberg-Rodes[a] and Babak Shahbaba[a]

[a]Department of Statistics, University of California, Irvine, CA, USA; [b]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA

ABSTRACT
We present geodesic Lagrangian Monte Carlo, an extension of Hamiltonian Monte Carlo for sampling from posterior distributions defined on general Riemannian manifolds. We apply this new algorithm to Bayesian inference on symmetric or Hermitian positive definite (PD) matrices. To do so, we exploit the Riemannian structure induced by Cartan's canonical metric. The geodesics that correspond to this metric are available in closed-form and – within the context of Lagrangian Monte Carlo – provide a principled way to travel around the space of PD matrices. Our method improves Bayesian inference on such matrices by allowing for a broad range of priors, so we are not limited to conjugate priors only. In the context of spectral density estimation, we use the (non-conjugate) complex reference prior as an example modelling option made available by the algorithm. Results based on simulated and real-world multivariate time series are presented in this context, and future directions are outlined.

ARTICLE HISTORY
Received 25 July 2017
Accepted 9 December 2017

KEYWORDS
HMC; Riemannian geometry; spectral analysis

1. Introduction

In this paper, we introduce geodesic Lagrangian Monte Carlo (gLMC), a methodology for Bayesian inference on a broad class of Riemannian manifolds. We illustrate this general methodology using the space of positive definite (PD) matrices as a concrete example. The resulting algorithms allow for *direct* inference on the space of PD matrices and are thus the first of their kind. As a result, gLMC facilitates better prior elicitation of covariance matrices by negating the need for conjugate priors and avoiding difficult-to-interpret transformations on variables of interest.

In statistics, PD matrices primarily appear as covariance matrices parameterizing the multivariate Gaussian model. This model is the workhorse of modern statistics and machine learning: linear regression, probabilistic principal components analysis, Gaussian Markov random fields, spectral density estimation, and Gaussian process models all rely on the multivariate Gaussian distribution. The d-dimensional Gaussian distribution is completely specified by a mean vector μ and a covariance matrix Σ in $\mathcal{S}_d^+$, the space of d-by-d

CONTACT Andrew Holbrook aholbroo@uci.edu

PD matrices. By imposing different structures on the covariance matrix, one can create different models. In some cases, it is possible to parameterize the covariance matrices in terms of a small number of parameters. However, learning of the *unstructured* covariance matrices, usually involved in inference on a large number of parameters, has remained as an issue. The conjugate Gaussian inverse-Wishart model has known deficiencies [1]. Outside of non-linear parameterizations of the Cholesky decomposition or matrix logarithm, there has not yet been a way to perform Bayesian inference directly on the space of PD matrices with flexible prior specifications using unstructured covariance matrices.

In this most general context the difficulty is in sampling from a posterior distribution on an abstract, high-dimensional manifold with boundary. It has not been clear how to propose moves from point to point within (and without leaving) this space. Our method takes advantage of the intrinsic, Riemannian geometry on the space of PD matrices. This space is incomplete under the Euclidean metric: following a straight trajectory will often result in matrices that are no longer PD. The space is, however, geodesically complete under the canonical metric: no matter how far the sampler travels along any geodesic, it never leaves the space of PD matrices. Intuitively, we redefine 'straight line' in a way that precludes leaving the set of interest. Moreover, the metric-induced geodesics provide a natural way to traverse the space of PD matrices, and these geodesics fit nicely with recent advances in Hamiltonian Monte Carlo (HMC) on manifolds [2–4].

To this end, we usegLMC, which belongs to a growing class of HMC algorithms. HMC provides an intelligent, partially deterministic method for moving around the parameter space while leaving the target distribution invariant. New Markov states are generated by numerically integrating a Hamiltonian system while Metropolis-Hastings steps account for the numerical error [5]. Riemannian manifold Hamiltonian Monte Carlo (RMHMC) adapts the proposal path by incorporating second-order information in the form of a Riemannian metric tensor [2]. Lagrangian Monte Carlo (LMC) builds on RMHMC by using a random velocity in place of RMHMC's random momentum. LMC's explicit integrator is no longer volume preserving; it therefore requires Jacobian corrections for each accept-reject step [6]. The embedding geodesic Monte Carlo (egMC) [3] is able to take the geometry of the parameter space into account while avoiding implicit integration by splitting the Hamiltonian [7] into a Euclidean and a geodesic component. Unfortunately, egMC is not applicable when a manifold's Riemannian embedding is unknown. gLMC, on the other hand, efficiently uses the same split Hamiltonian formulation as egMC but does not require an explicit Riemannian embedding (see [4], for example). This last fact makes gLMC an ideal candidate for Bayesian inference on the space of PD matrices.

gLMC allows us to treat the entire covariance matrix as one would treat any other model parameters. We are no longer restricted to use a conjugate prior distribution or to specify a low-rank structure. We illustrate applications of gLMC for PD matrices using both simulated and real-world data. First, we show that gLMC provides the same empirical results as the closed-form solution for the conjugate Gaussian inverse-Wishart model. After this, we focus on applying gLMC for Hermitian PD matrices to multivariate spectral density estimation and compare the results obtained from two different prior specifications: the inverse-Wishart prior and the complex reference prior. Then, we obtain credible intervals for the squared coherences (see Section 2) of simulated vector auto-regressive time series for which the spectral density matrix is known. Finally, we apply gLMC to learn the spectral density matrix associated with multivariate local field potentials (LFPs).

The contributions of this paper are as follows:

- gLMC, an MCMC methodology for Bayesian inference on general Riemannian manifolds, is proposed;
- to illustrate the general methodology, we provide a detailed description of gLMC on the spaces of symmetric and Hermitian PD matrices;
- for classical statisticians, the paper serves as a brief introduction to spectral density estimation and its Bayesian approach;
- the proposed algorithms are applied to Bayesian inference on (real and complex) covariance matrices based on simulated and real-world data and using a number of different prior specifications.

It should be noted that the proposed method is useful for generating samples from the posterior distribution of interest and not just a point estimate. The proposed method is for full inference of a posterior distribution defined *directly* over the space of PD matrices without limiting ourselves to conjugate priors, as such, is the first of its kind.

That said, it is sometimes sufficient for the scientist to obtain a point estimate of the covariance or spectral density matrix. In this context, regularization of the estimate is often advantageous. Regularization approaches may be interpreted as Bayesian and their corresponding point estimates are interpreted as *maximum a posteriori* (MAP) estimates. See [8] for a statistically minded survey of covariance estimation and regularization, and see [9] for a state-of-science approach to point estimation in signal processing applications.

The rest of the paper is outlined thus: in Section 2, we provide motivation for our approach in the form of a brief introduction to spectral density estimation for multivariate time series; in Section 3, we define PD matrices and show how the space of PD (symmetric or Hermitian) matrices comprises a Riemannian manifold; in Section 4, we present the gLMC methodology for Bayesian inference on *general* Riemannian manifolds; in Section refpdhmc, we detail gLMC on the spaces of symmetric and Hermitian PD matrices; in Section 6, we introduce the reader to common proper and improper priors for covariance matrices; in Section 7, we present empirical results based on simulated and real-world data.

2. Motivation: learning the spectral density matrix

Given a stationary multivariate time series $y(t) = (y_1(t), \dots, y_d(t))^{\mathsf{T}} \in \mathbb{R}^d,\ t = 1, \dots, T$, one often wants to characterize the dependencies between vector elements through time. There are multiple ways to define such dependencies, and these definitions feature either the time series directly or the Fourier-transformed series in the frequency domain. In the time domain, one characterization of dependence is provided by the lagged variance–covariance matrix Γ_ℓ. In terms of lag ℓ, this is written

$$\Gamma_\ell = \operatorname{Cov}\left(y(t),\ y(t-\ell)\right) = \mathrm{E}\left(\left(y(t)-\mu\right)\left(y(t-\ell)-\mu\right)^{\mathsf{T}}\right). \tag{1}$$

Note that Γ_ℓ and μ are invariant over time by stationarity. If the scientist has a reason to suspect that a certain lag ℓ is importanta priori, then Γ_ℓ can be a useful measure. On the other hand, it is often more scientifically tractable to think in terms of frequencies rather than lags. In neuroscience, for example, one might hypothesize that two brain regions have 'correlated' activity during the performance of a specific task, but this co-activity may be too complex to describe in terms of a simple lagged relationship. The spectral density approach lends itself naturally to this kind of question. For a full discussion, see [10].

The power spectral density matrix is the Fourier transform of Γ_ℓ:

$$\Sigma(\omega) = \sum_{\ell=-\infty}^{\infty} \Gamma_\ell \exp(-i2\pi\omega\ell). \tag{2}$$

$\Sigma(\omega)$ is a Hermitian PD matrix. A Hermitian matrix M is a complex valued matrix satisfying $M^H = \bar{M}^\mathsf{T} = M$, where $\overline{(\cdot)}$ denotes taking the complex conjugate. A Hermitian matrix M is defined to be PD if $z^H M z > 0$, $\forall z \in \mathbb{C}^d \setminus \{0\}$.

A diagonal element $\Sigma_{ii}(\omega)$ is called the auto-spectrum of $y_i(t)$ at frequency ω, and an off-diagonal element $\Sigma_{ij}(\omega)$, $i \neq j$ is the cross-spectrum of $y_i(t)$ and $y_j(t)$ at frequency ω. The squared coherence is given by

$$\rho^2_{ij}(\omega) = \frac{|\Sigma_{ij}(\omega)|^2}{\Sigma_{ii}(\omega)\Sigma_{jj}(\omega)}, \tag{3}$$

where $|\cdot|$ denotes the complex modulus. There are a number of ways to estimate the spectral density matrix and, hence, the matrix of squared coherences. In this paper, we use the Whittle likelihood approximation [11]. We model the discrete Fourier transformed time series $Y(\omega_k) \in \mathbb{C}^d$ as following a (circularly-symmetric) complex multivariate Gaussian distribution:

$$Y(\omega_k) \overset{\text{ind}}{\sim} \text{CN}_d(0, \Sigma(\omega_k)), \tag{4}$$

where, for $\omega_k = k/T$ and $k = -(T/2-1), \ldots, T/2$,

$$Y(\omega_k) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\mathrm{T}} y(t) \exp(-i2\pi\omega_k t). \tag{5}$$

Three assumptions are made here. First, we assume that the $Y(\omega_k)$s are *exactly* Gaussian: this is true when the $y(t)$ follow any Gaussian process. Moreover, if $y(t)$ follow a linear process, then the $Y(\omega_k)$ are asymptotically Gaussian as T goes to infinity [12]. Second, we assume that for $\omega_k \neq \omega_{k'}$, $Y(\omega_k)$ and $Y(\omega_{k'})$ are independent, whereas [12] show that they are asymptotically uncorrelated. Third, we assume that $\Sigma(\cdot)$ is approximately piecewise constant across frequency bands, and take all $Y(\omega_k)$ to be approximately i.i.d. within a small enough frequency band. For example, if we are interested in the alpha band of neural oscillations ranging from 7.5 to 12.5 Hz, then we model

$$Y(\omega_k) \overset{\text{iid}}{\sim} \text{CN}_d(0, \Sigma_\alpha), \tag{6}$$

where Σ_α denotes the spectral density matrix shared by the entire band. For a recent use of the approximately piecewise constant assumption, see [13], where the spectrum is represented as a sum of unique AR(2) spectra, with each of the AR(2) capturing distinct frequency bands.

Thus, having obtained samples $Y(\omega)$ from a fixed frequency band, we will use gLMC over Hermitian PD matrices to perform inference on Σ. The posterior samples of Σ automatically provide samples for the distributions of the squared coherences, which can in turn elucidate dependencies between the univariate time series. Before discussing gLMC for PD matrices, we establish necessary facts regarding the space of PD matrices.

3. The space of PD matrices

Let $\mathcal{S}_d(\mathbb{C})$ denote the space of $d \times d$ Hermitian matrices, and $\mathcal{S}_d^+(\mathbb{C})$ denote its subspace of PD matrices.[1] The space of Hermitian PD matrices, $\mathcal{S}_d^+(\mathbb{C})$, may be written as a quotient space $GL(d,\mathbb{C})/U(d)$ of the complex general linear group $GL(d,\mathbb{C})$ and the unitary group $U(d)$. The general linear group is the smooth manifold for which every point is a matrix with non-zero determinant. The unitary group is the space of all complex matrices U satisfying $U^HU = UU^H = I$. This quotient space representation is rooted in the fact that every PD matrix may be written as the product $\Sigma = GG^H = GUU^HG^H$ for a unique $G \in GL(d,\mathbb{C})$ and any arbitrary unitary matrix $U \in U(d)$. For the convenience of exposition, we drop the dependence on $\mathbb{C}$ of symbols in the following of this section. Related references are [14–16]. $\mathcal{S}_d^+$ is a homogeneous space with respect to the general linear group: this means that the group acts transitively on the $\mathcal{S}_d^+$. Here the group action is given by conjugation:

$$G^*\Sigma = G\Sigma G^H. \tag{7}$$

For any $\Sigma_1, \Sigma_2 \in \mathcal{S}_d^+$, it simply takes the composition $\Sigma_2^{1/2*} \circ \Sigma_1^{-1/2*}$ to transform Σ_1 into Σ_2:

$$\Sigma_2^{1/2*} \circ \Sigma_1^{-1/2*}\Sigma_1 = \Sigma_2^{1/2*}\left(\Sigma_1^{-1/2}\Sigma_1\Sigma_1^{-1/2}\right) = \Sigma_2^{1/2}I\Sigma_2^{1/2} = \Sigma_2. \tag{8}$$

The space of Hermitian matrices, $\mathcal{S}_d$, happens to be the tangent space to the space of Hermitian PD matrices at the identity, denoted as $T_{Id}\mathcal{S}_d^+$, that is, $T_{Id}\mathcal{S}_d^+ = \mathcal{S}_d$. The action

$$\Sigma^{1/2*} : V \mapsto \Sigma^{1/2}V\Sigma^{1/2} \tag{9}$$

translates vector $V \in T_{Id}\mathcal{S}_d^+$ to its corresponding vector in $T_\Sigma\mathcal{S}_d^+$, the tangent space to the space of PD matrices at point Σ.

Élie Cartan constructed a natural Riemannian metric $g(\cdot,\cdot)$ on the tangent bundle $T\mathcal{S}_d^+$ that is invariant under group action (7). For two vectors $V_1, V_2 \in T_{Id}\mathcal{S}_d^+$, the metric is given by

$$g_I(V_1, V_2) = \mathrm{tr}(V_1V_2). \tag{10}$$

In this way the space of PD matrices is isometric to Euclidean space (equipped with the Frobenius norm) at the identity. Next define the metric at any arbitrary point Σ to be

$$g_\Sigma(V_1, V_2) = \mathrm{tr}(\Sigma^{-1}V_1\Sigma^{-1}V_2). \tag{11}$$

It is easy to check that $g_I(V_1, V_2) = g_\Sigma(\Sigma^{1/2*}V_1, \Sigma^{1/2*}V_2)$ and so $\Sigma^{1/2*}$ is a Riemannian isometry on $\mathcal{S}_d^+$.

Two geometric quantities are required for our purposes: the Riemannian metric tensor and its corresponding geodesic flow, specified by a starting point and an initial velocity vector. The computational details involving the metric tensor are presented in Section 5. Here we present the closed form solution for the geodesic flow as found in [14]. $\mathcal{S}_d^+$ is an affine symmetric space [17]. As such the geodesics under the invariant metric are generated by the one-parameter subgroups of the acting Lie group [14,16]. These one-parameter subgroups are given by the group exponential map which, at the identity, is given by the matrix exponential $\exp tG$. In order to calculate the unique geodesic curve with starting position $\Sigma(0)$ and initial velocity $V(0)$, all one needs is to translate the velocity to the identity, compute the matrix exponential, and translate it back to the point of interest. In sum, the geodesic is given by

$$\Sigma(t) = \exp_\Sigma tV(0) = \Sigma(0)^{1/2} \exp\left(t\Sigma(0)^{-1/2}V(0)\Sigma(0)^{-1/2}\right)\Sigma(0)^{1/2}. \qquad (12)$$

The corresponding flow on the tangent bundle will also be useful. This is obtained by taking the derivative with respect to t:

$$\begin{aligned} V(t) = \dot{\Sigma}(t) &= \frac{d}{dt}\exp_\Sigma tV(0) \\ &= V(0)\Sigma(0)^{-1/2}\exp\left(t\Sigma(0)^{-1/2}V(0)\Sigma(0)^{-1/2}\right)\Sigma(0)^{1/2}. \end{aligned} \qquad (13)$$

For a Lie group, the exponential map (on which the above formula is based) is a local diffeomorphism between the tangent space at a point on the manifold and the manifold itself. Given a tangent vector V at Σ, $\exp_\Sigma V$ is a point on the manifold. Incidentally, for the spaces of PD matrices, this diffeomorphism is global. The inverse of the exponential map is the logarithmic map. Whereas the exponential map on the manifold takes Hermitian matrices ($\mathcal{S}_d$) to Hermitian PD matrices ($\mathcal{S}_d^+$), the logarithmic map takes Hermitian PD matrices ($\mathcal{S}_d^+$) to Hermitian matrices ($\mathcal{S}_d$).

Together these are most of the geometric quantities required for gLMC over PD matrices. The next section presents HMC, its geometric extension RMHMC, and its Lagrangian manifestations.

4. Bayesian inference using the geodesic LMC

Given data $y_1, \ldots, y_N \in \mathbb{R}^n$, one may specify a generative model by a likelihood function, $p(y \mid q)$. In the following we allow $q \in \mathcal{M}^m$ to be an m-dimensional vector on a manifold that parameterizes the likelihood. Endowing q with a prior distribution $p(q)$ renders the posterior distribution

$$\pi(q) = p(q \mid y) = \frac{p(y \mid q)p(q)}{\int p(y \mid q)p(q)\,\mathrm{d}q}. \qquad (14)$$

The integral is often referred to as the evidence and may be interpreted as the probability of observing data y given the model. In most interesting models the evidence integral is intractable and high-dimensional models do not lend themselves to easy numerical integration. Non-quadrature sampling techniques such as importance sampling or even random walk MCMC also suffer in high dimensions. HMC is an effective sampling tool for higher dimensional models over continuous parameter spaces [5,18]. Here we discuss HMC and its geometric variants (see Section 1) in detail.

In HMC, a Hamiltonian system is constructed that consists of the parameter vector q and an auxiliary vector p of the same dimension. The negative-log transform turns the probability density functions into a potential energy function $U(q) = -\log \pi(q)$ and corresponding kinetic function $K(p)$. Thus q and p become the position and momentum of Hamiltonian function

$$H(q, p) = U(q) + K(p). \tag{15}$$

By Euler's method or extensions, the system is numerically advanced according to Hamilton's equations:

$$\begin{aligned} \frac{\mathrm{d}q}{\mathrm{d}t} &= \frac{\partial H}{\partial p}, \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= -\frac{\partial H}{\partial q}. \end{aligned} \tag{16}$$

Riemannian manifold HMC uses a slightly more complicated Hamiltonian to sample from posterior $\pi(q)$:

$$H(q, p) = -\log \pi(q) + \tfrac{1}{2} \log |G(q)| + \tfrac{1}{2} p^{\mathsf{T}} G(q)^{-1} p. \tag{17}$$

Here, $G(q)$ is the Fisher information matrix at point q (in Euclidean space) and may be interpreted as a Riemannian metric tensor induced by the curvature of the log-probability. Exponentiating and negating $H(q, p)$ reveals p to follow a Gaussian distribution centred at origin with metric tensor $G(q)$ for covariance. The corresponding system of first-order differential equations is given by

$$\begin{aligned} \frac{\mathrm{d}q}{\mathrm{d}t} &= G(q)^{-1} p, \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= \nabla_q \left(\log \pi(q) - \frac{1}{2} \log |G(q)| - \frac{1}{2} p^{\mathsf{T}} G(q)^{-1} p \right). \end{aligned} \tag{18}$$

The Hamiltonian is not separable in p and q. To get numerical solutions, one may split it into a potential term $H^{[1]}$, featuring q alone, and a kinetic term, $H^{[2]}$, featuring both variables [3,7]. The two systems are then simulated in turn. The first term is given by

$$H^{[1]}(q, p) = -\log \pi(q) + \tfrac{1}{2} \log |G(q)| \tag{19}$$

and starting at $(q(0), p(0))$ the associated system has solutions

$$q(t) = q(0) \quad \text{and} \quad p(t) = p(0) + t \nabla_q (\log \pi(q) - \tfrac{1}{2} \log |G(q)|)|_{q=q(0)}. \tag{20}$$

The second component is the quadratic form

$$H^{[2]}(q, p) = \tfrac{1}{2} p^{\mathsf{T}} G(q)^{-1} p. \tag{21}$$

The solutions to the system associated with $\mathrm{H}^{[2]}$ are given by the geodesic flow under the Levi-Civita connection with respect to metric G and with momentum $p(t) = G(q(t)) \dot{q}(t)$. There is, however, no a priori reason to restrict $G(q)$ to be the Fisher information as is done in the [2]. In fact, by allowing $G(q)$ to take on other forms, one may perform HMC on a number of manifold parameterized models.

Algorithm 1 Geodesic Lagrangian Monte Carlo

```
Let $q = q^{(k)}$ be the $k$th state of the Markov chain. The next sample is generated according
to the following procedure.
(a) Generate proposal state $q^*$:
  $v \sim N(0, G^{-1}(q))$
  $e \leftarrow -\log \pi(q) - \frac{1}{2}\log|G(q)| + \frac{1}{2}v^\mathsf{T} G(q) v$
  $q^* \leftarrow q$
  for $\tau = 1, \dots, T$ do
      $v \leftarrow v + \frac{\epsilon}{2} G(q^*)^{-1} \nabla_q \left(\log \pi(q^*) + \frac{1}{2}\log|G(q^*)|\right)$
      Progress $(q^*, v)$ along the geodesic flow for time $\epsilon$.
      $v \leftarrow v + \frac{\epsilon}{2} G(q^*)^{-1} \nabla_q \left(\log \pi(q^*) + \frac{1}{2}\log|G(q^*)|\right)$
  end for
  $e^* \leftarrow -\log \pi(q^*) - \frac{1}{2}\log|G(q^*)| + \frac{1}{2}v^\mathsf{T} G(q^*) v$
(b) Accept proposal with probability $\min\{1, \exp(e)/\exp(e^*)\}$:
  $u \sim U(0,1)$
  if $u < \exp(e - e^*)$ then
      $q \leftarrow q^*$
  end if
(c) Assign value $q$ to $q^{(k+1)}$, the $(k+1)$th state of the Markov chain.
```

4.1. Geodesic Lagrangian Monte Carlo

Byrne and Girolami [3] show how to extend the RMHMC framework to manifolds that admit a known Riemannian isometric embedding into Euclidean space. The algorithm is especially efficient when there exists a closed form linear projection of vectors in the ambient space onto the tangent space at any point. Although this embedding will always exist [19], it is rarely known. When equipped with the canonical metric, the space of PD matrices does not admit a known isometric embedding. Moreover, we are unaware of a closed-form projection onto the manifold's tangent space at a given point. We therefore opt for an intrinsic approach instead.

In the prior section, we stated that the solution to Hamilton's equations associated with the kinetic term $H^{[2]}$ is given by the geodesic flow with respect to the Levi-Civita connection. This flow is easily written in terms of the exponential map with respect to a velocity vector (as opposed to the momentum covector). Given an arbitrary covector $p \in T_q^*\mathcal{M}$, one may obtain the corresponding vector $v \in T_q\mathcal{M}$ by the one-to-one transformation $v = G^{-1}(q)p$. Hence whereas RMHMC augments the system with $p \sim N(0, G(q))$, Lagrangian Monte Carlo makes use of $v = G^{-1}(q)p \sim N(0, G^{-1}(q))$. The energy function is then given by

$$E(q, v) \propto -\log \pi(q) - \frac{1}{2}\log|G(q)| + \frac{1}{2}v^\mathsf{T} G(q) v. \qquad (22)$$

The probabilistic interpretation of the energy remains the same as in the case of RMHMC: the energy is the negative logarithm of the probability density functions of two independent random variables, one of which is the variable of interest, the other of which is the augmenting Gaussian variable. On the other hand, the physical interpretation is different.

We use the term 'energy' in order to accommodate the two physical interpretations available for Equation (22): $E(q, v)$ may be thought of either as a Hamiltonian or as a Lagrangian energy. In practice, which formulation is used is dictated by the geometric information available. The Lagrangian formulation provides efficient update equations when no closed-form geodesics are available. In this case, the Lagrangian (energy) is defined as the kinetic term T less the potential term V as follows

$$V(q, v) = \log \pi(q) + \tfrac{1}{2} \log |G(q)|, \quad \text{and} \quad T(q, v) = \tfrac{1}{2} v^{\mathsf{T}} G(q) v. \tag{23}$$

But when closed-form geodesics are available, it is useful to follow [3] and split the (now considered) Hamiltonian into two terms as in Equations (19) and (21). Within this regime, $H^{[1]} = -V$ and $H^{[2]} = T$. In analogy with Equation (20) and starting at $(q(0), v(0))$, the system defined by potential V has solution

$$q(t) = q(0), \quad v(t) = v(0) + tG(q)^{-1} \nabla_q V(q, v)|_{q=q(0)}, \tag{24}$$

and the system defined by kinetic term T has the unique geodesic path specified by starting position $q(0)$ and initial velocity $v(0)$ as a solution. The inverse metric tensor $G^{-1}(q)$ is used to 'raise the index', that is, transform the covector $\nabla_q V(q, v)$ into a vector on the tangent space at q. Thus it plays a similar function to the orthogonal projection in [3]. We call this formulation geodesic Lagrangian Monte Carlo (gLMC) and detail its steps in Algorithm 1, where the term 'Lagrangian' is used to emphasize the fact that we use velocities in place of momenta. Note [4,20] implemented the similar idea on the manifold of a d-dimensional sphere. To implement geodesic Lagrangian Monte Carlo, one must be able to compute the inverse metric tensor $G^{-1}(q)$ and the geodesic path given starting values. When the space of PD matrices is equipped with the canonical metric, $G^{-1}(q)$ is given in closed-form and the geodesic path is easily computable.

5. gLMC on the manifold of PD matrices

To perform gLMC on the space of PD matrices, $\mathcal{S}_d^+$, we equip the manifold with the canonical metric. In order to signify that we are no longer dealing with gLMC in its full generality, we adopt the notation of Section 3. PD matrix Σ replaces q, and symmetric or Hermitian matrix V replaces v. All other notations remain the same. As stated in the previous section, we require the inverse metric tensor $G^{-1}(\Sigma)$. To compute this quantity, we need a couple more tools provided by Moakher and Zéraï [15]. Let vech$(\cdot)$ take symmetric (Hermitian) $d \times d$ matrices to vectors of length $(d/2)(d+1)$ by stacking diagonal and subdiagonal matrix elements in the following way:

$$\text{vech}(V) = (V_{11}, V_{21}, \ldots, V_{d1}, V_{22}, \ldots, V_{d2}, \ldots, V_{dd}). \tag{25}$$

Let vec$(\cdot)$ take symmetric (Hermitian) $d \times d$ matrices to vectors of length d^2 by stacking all matrix elements:

$$\text{vec}(V) = (V_{11}, V_{21}, \ldots, V_{d1}, V_{12}, \ldots, V_{d2}, \ldots, V_{1d}, \ldots, V_{dd}). \tag{26}$$

Let D_d be the unique $d^2 \times (d/2)(d+1)$ matrix satisfying

$$\text{vec}(V) = D_d \text{vech}(V). \tag{27}$$

Algorithm 2 gLMC for symmetric PD matrices

```
Let $\Sigma = \Sigma^{(k)}$ be the $k$th state of the Markov chain. The next sample is generated according to the following procedure.
(a) Generate proposal state $\Sigma^*$:
    $\text{vech}(V) \sim N(0, G^{-1}(\Sigma))$
    $e \leftarrow -\log \pi(\Sigma) - \frac{d+1}{2} \log |\Sigma| + \frac{1}{2}\text{vech}(V)^{\mathsf{T}} G(\Sigma)\text{vech}(V)$
    $\Sigma^* \leftarrow \Sigma$
    for $\tau = 1, \dots, T$ do
        $\text{vech}(V) \leftarrow \text{vech}(V) + \frac{\epsilon}{2} G^{-1}(\Sigma^*)\text{vech}\left(\nabla_\Sigma \left(\log \pi(\Sigma^*) + \frac{d+1}{2} \log |\Sigma^*|\right)\right)$
        Progress $(\Sigma^*, V)$ along the geodesic flow for time $\epsilon$.
        $\text{vech}(V) \leftarrow \text{vech}(V) + \frac{\epsilon}{2} G^{-1}(\Sigma^*)\text{vech}\left(\nabla_\Sigma \left(\log \pi(\Sigma^*) + \frac{d+1}{2} \log |\Sigma^*|\right)\right)$
    end for
    $e^* \leftarrow -\log \pi(\Sigma^*) - \frac{d+1}{2} \log |\Sigma^*| + \frac{1}{2}\text{vech}(V)^{\mathsf{T}} G(\Sigma^*)\text{vech}(V)$
(b) Accept proposal with probability $\min\{1, \exp(e)/\exp(e^*)\}$:
    $u \sim U(0, 1)$
    if $u < \exp(e - e^*)$ then
        $\Sigma \leftarrow \Sigma^*$
    end if
(c) Assign value $\Sigma$ to $\Sigma^{(k+1)}$, the $(k+1)$th state of the Markov chain.
```

Denote D_d^+ as the Moore–Penrose inverse of D_d satisfying

$$\text{vech}(V) = D_d^+ \text{vec}(V), \tag{28}$$

with D_d^+ given by

$$D_d^+ = (D_d^{\mathsf{T}} D_d)^{-1} D_d^{\mathsf{T}}. \tag{29}$$

Then Moakher and Zéraï [15] show that the metric tensor and inverse metric tensor are given by the $(d/2)(d+1) \times (d/2)(d+1)$ dimensional matrices

$$G(\Sigma) = D_d^{\mathsf{T}} (\Sigma^{-1} \otimes \Sigma^{-1}) D_d \quad \text{and} \quad G^{-1}(\Sigma) = D_d^+ (\Sigma \otimes \Sigma) D_d^{+T}. \tag{30}$$

Finally, the determinant of $G(\Sigma)$ can be expressed in terms of Σ alone:

$$|G(\Sigma)| \propto |\Sigma|^{d+1}. \tag{31}$$

The metric tensor features in the energy function for gLMC for both symmetric and Hermitian PD matrices. For symmetric PD matrices, the energy is given by

$$\begin{aligned} E(\Sigma, V) &\propto -\log \pi(\Sigma) - \frac{1}{2} \log |G(\Sigma)| + \frac{1}{2}\text{vech}(V)^{\mathsf{T}} G(\Sigma)\text{vech}(V) \\ &\propto -\log \pi(\Sigma) - \frac{d+1}{2} \log |\Sigma| + \frac{1}{2}\text{vech}(V)^{\mathsf{T}} G(\Sigma)\text{vech}(V), \end{aligned} \tag{32}$$

but the energy for Hermitian PD matrices is slightly different. In this case, both Σ and V are complex valued, and $\text{vech}(V)$ follows a multivariate complex Gaussian distribution with covariance $G^{-1}(\Sigma)$. Therefore, the gLMC energy for Hermitian PD matrices is given by

$$\begin{aligned} E(\Sigma, V) &\propto -\log \pi(\Sigma) - \log |G(\Sigma)| + \mathrm{vech}(V)^H G(\Sigma)\mathrm{vech}(V) \\ &\propto -\log \pi(\Sigma) - (d+1)\log|\Sigma| + \mathrm{vech}(V)^H G(\Sigma)\mathrm{vech}(V), \end{aligned} \tag{33}$$

where $(\cdot)^H$ signifies the conjugate transpose. Notice that the log-determinant and quadradic terms are not multiplied by the factor 1/2. This accords with the density function of a complex Gaussian random variable. See Appendix for more details.

The metric tensor (30) and the geodesic equations (12) and (13) are the only geometric quantities required for gLMC on PD matrices. The kth iteration of the symmetric PD algorithm is shown in Algorithm 2. The kth iteration of the Hermitian PD algorithm is shown in Algorithm 3. First, one generates a Gaussian initial velocity on $T_{\Sigma^{(k)}}\mathcal{S}_d^+$ (Step 1). Then, the energy function is evaluated and stored (Step 2). Next, the system is numerically advanced using the split Hamiltonian scheme. Following Equation (24), the velocity vector V is updated one half-step with the gradient of $\mathrm{H}^{[1]}$ (Step 4). For Step 5, both Σ and V are updated with respect to $H^{[2]}$, that is, they are transported along the geodesic flow given by Equations (12) and (13):

$$[\Sigma(0), V(0)] \mapsto [\Sigma(\epsilon), V(\epsilon)]\,. \tag{34}$$

Again, the velocity vector V is updated one half-step with the gradient of $\mathrm{H}^{[1]}$ (Step 6). Finally, the energy is evaluated at the new Markov state (Step 9), and a Metropolis accept-reject step is implemented (Steps 10–12). It is important to note that, besides being over different algebraic fields, the symmetric and Hermitian instantiations only differ in their respective energies. The general implementation is the same. See Appendix for a short discussion on gradients.

6. Some priors on covariance matrices

We provide a short introduction to some well known priors for covariance matrices. This list is in no way exhaustive but is meant to hint at the choices available to practitioners. Of the priors we present, three are improper (are not well defined probability distributions) and two are proper. The real and complex versions of all five priors are shown in Table 1.

The Wishart and inverse-Wishart distributions are the most well known for PD matrices. These two distributions are popular not because they are particularly good models

Table 1. Priors for Σ and their densities up to proportionality: the first two priors are proper, that is, comprise well-defined probability distributions, the last three are not.

Prior	Real	Complex
Wishart	$\lvert\Sigma\rvert^{(n-d-1)/2}\exp(-\mathrm{tr}\{\Psi^{-1}\Sigma\}/2)$	$\lvert\Sigma\rvert^{n-d}\exp(-\mathrm{tr}\{\Psi^{-1}\Sigma\})$
inverse-Wishart	$\lvert\Sigma\rvert^{-(n+d+1)/2}\exp(-\mathrm{tr}\{\Psi\Sigma^{-1}\}/2)$	$\lvert\Sigma\rvert^{-(n+d)}\exp(-\mathrm{tr}\{\Psi\Sigma^{-1}\})$
uniform	1	1
Jeffreys	$\lvert\Sigma\rvert^{-(d+1)/2}$	$\lvert\Sigma\rvert^{-d}$
reference	$(\lvert\Sigma\rvert\prod_{i<j}(d_i-d_j))^{-1}$	$(\lvert\Sigma\rvert\prod_{i<j}(d_i-d_j)^2)^{-1}$

Notes: Σ is symmetric and Hermitian PD in the left and right columns, respectively. Note how the Wishart, inverse-Wishart, and Jeffreys priors share similar patterns moving from real to complex numbers.

Algorithm 3 gLMC for Hermitian PD matrices

Let $\Sigma = \Sigma^{(k)}$ be the kth state of the Markov chain. The next sample is generated according to the following procedure.

(a) Generate proposal state Σ^*:

```
$\mathrm{vech}(V) \sim \mathrm{CN}(0, G^{-1}(\Sigma))$
$e \leftarrow -\log \pi(\Sigma) - (d+1)\log|\Sigma| + \mathrm{vech}(V)^H G(\Sigma)\mathrm{vech}(V)$
$\Sigma^* \leftarrow \Sigma$
for $\tau = 1, \dots, T$ do
    $\mathrm{vech}(V) \leftarrow \mathrm{vech}(V) + \frac{\epsilon}{2} G^{-1}(\Sigma^*)\mathrm{vech}\left(\nabla_\Sigma\left(\log\pi(\Sigma^*) + (d+1)\log|\Sigma^*|\right)\right)$
    Progress $(\Sigma^*, V)$ along the geodesic flow for time $\epsilon$.
    $\mathrm{vech}(V) \leftarrow \mathrm{vech}(V) + \frac{\epsilon}{2} G^{-1}(\Sigma^*)\mathrm{vech}\left(\nabla_\Sigma\left(\log\pi(\Sigma^*) + (d+1)\log|\Sigma^*|\right)\right)$
end for
$e^* \leftarrow -\log\pi(\Sigma^*) - (d+1)\log|\Sigma^*| + \mathrm{vech}(V)^H G(\Sigma^*)\mathrm{vech}(V)$
```

(b) Accept proposal with probability $\min\{1, \exp(e)/\exp(e^*)\}$:

```
$u \sim U(0,1)$
if $u < \exp(e - e^*)$ then
    $\Sigma \leftarrow \Sigma^*$
end if
```

(c) Assign value Σ to $\Sigma^{(k+1)}$, the $(k+1)$th state of the Markov chain.

but because they make Bayesian inference easy for covariance matrices. The Wishart and inverse-Wishart distributions are conjugate priors for the precision and covariance matrices, respectively. This means that they provide closed-form posteriors given the data and thus obviate Monte Carlo methods. Of course, the Wishart distribution can be used as a prior for covariances (as opposed to precision matrices), but it usually is not since conjugacy is then lost. The inverse-Wishart distribution is the distribution of the inverse of a Wishart random variable. Shown in Table 1, both distributions are parameterized by symmetric or Hermitian PD matrices Ψ and scalar ν which is greater than $d-1$ for real and d for complex Σ.

The improper priors are the flat, the Jeffreys, and the reference priors. The MLE may be interpreted as the MAP estimate given the flat prior. The Jeffreys prior is the square-root determinant of the Fisher information and is parameterization invariant. When Σ is real, the Jeffreys prior is the reciprocal of the density of the Hausdorff measure with respect to the Lebesgue measure [3]; it can therefore be interpreted as the flat prior with respect to the Hausdorff measure. The reference prior is designed to prevent estimates from being ill-conditioned: as may be seen in Figure 1, it favours eigenvalues that are close together. This corresponds to better frequentist estimation properties [8]. For an introduction to the complex Wishart and inverse-Wishart distributions, see [22]. For the complex Jeffreys and reference priors, see [23,24], respectively.

Again, these priors are not intended to form a comprehensive list but give an idea of the kinds of choices that statisticians might make when choosing a prior for a covariance. For more recent developments in this area, see [25,26].

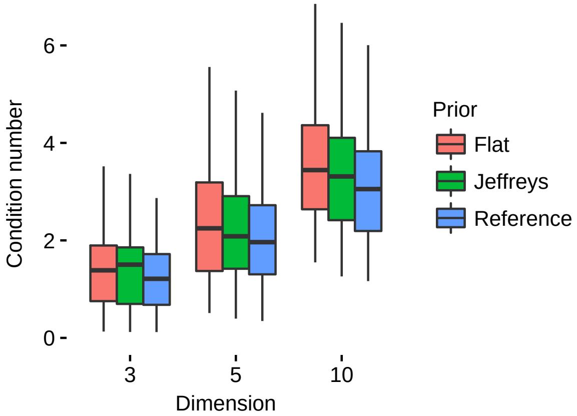

Figure 1. Median condition number by dimension and prior specification: box plots describe distributions of 100 median condition numbers for each dimension and prior. Each point is the median from 200 posterior samples based on independent data and using gLMC. The reference prior is designed to yield smaller condition numbers than Jeffreys prior and hence better asymptotics [21].

7. Results

This section features empirical validation of the gLMC algorithm as well as an application to learning the spectral density matrix for vector time series. For empirical validation, we present quantile–quantile (Q–Q) and trace plots comparing the gLMC sample to the closed-form solution made available by the conjugate prior. We then use gLMC for Hermitian PD matrices to learn the spectral density matrices of both simulated and LFP time series. We use the posteriors thus obtained to get credible intervals on the squared coherences for the vector time series.

7.1. Empirical validation

Before applying gLMC to spectral density estimation, we demonstrate validity by comparing samples from empirical posterior distributions of the Gaussian inverse-Wishart model obtained by gLMC and the closed-form solution. Note that our objective is to show that our proposed method provides valid results, similar to those obtained based on conjugate priors, while creating a flexible framework for eliciting and specifying prior distributions directly over the space of PD matrices. To this end, we compare element-wise distributions with Q–Q plots and whole-matrix distributions with two global matrix summaries. The comparisons based on 10,000 samples by the different sampling methods over 3-by-3 PD matrices are illustrated in Figure 2. The first 200 samples are discarded, and, for better visualization, every tenth sample is kept. For the global matrix summaries, we also include samples from an indirect approach, the log-Cholesky parameterization of the PD matrix (currently used in Stan software implementations [27]). Starting with a lower-triangular matrix L, one obtains a PD matrix by exponentiating the diagonal elements of L to get $\tilde{L}$ and then evaluating $\Sigma = \tilde{L}\tilde{L}^{\mathrm{T}}$. Given a distribution over PD matrices, one may obtain a distribution over lower-triangular matrices using the inverses of these transforms and their corresponding Jacobians.

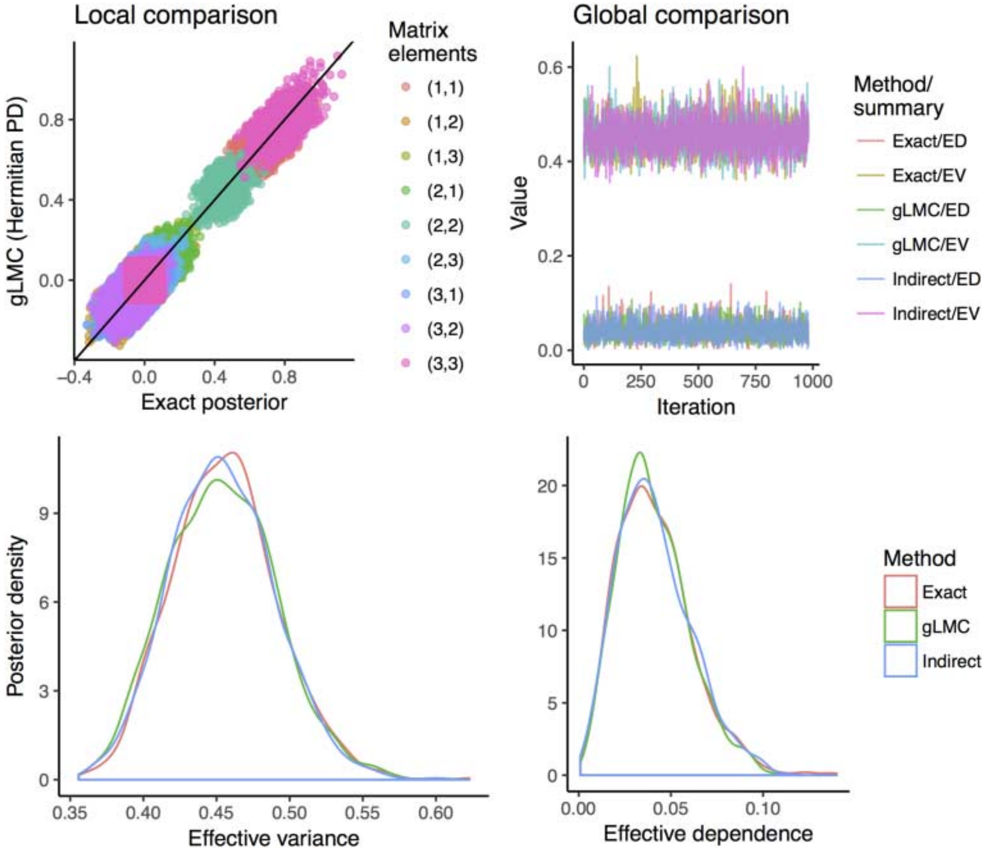

Figure 2. These figures provide empirical validation for the well-posedness of gLMC for PD matrices. On the top-left is a Q–Q plot comparing the gLMC (for Hermitian PD matrices) posterior sample with that of the closed-form posterior for the complex Gaussian inverse-Wishart model. Both real and imaginary elements are included, and points are jittered for visibility. On the top-right are posterior samples of 'global' matrix summaries pertaining both to gLMC (for symmetric PD matrices), the closed-form 'exact' solution, and the 'indirect' log-Cholesky parameterization. These summaries are the effective variance (EV) and the effective dependence (ED), built off the covariance matrix and the correlation matrix, respectively. On the bottom are posterior density plots of the same matrix summaries.

On the top-left panel of Figure 2, a Q–Q plot is used to compare the gLMC Hermitian PD posterior sample to the closed-form posterior. The Q–Q plot is the gold standard for comparing two scalar distributions using empirical samples because full quantile agreement corresponds to equality of cumulative distribution functions. Points are jittered for easy visualization, and each colour specifies a different matrix element. Note that some colours appear twice: these double appearances correspond to real and imaginary matrix elements. For example, pink appears at zero as well as the upper right of the plot: this colour corresponds to a diagonal matrix element since, on account of the matrix being Hermitian, its imaginary part is fixed at zero. Most importantly, all matrix elements fit tightly around the line $y = x$, suggesting a perfect match in quantiles between empirical distributions.

On the top-right panel of Figure 2, we present samples obtained from two whole-matrix summaries using gLMC, the 'exact' closed-form posterior, and the 'indirect' log-Cholesky parameterization. These summaries are the EV and the ED:

$$\mathrm{EV}(\Sigma) = |\Sigma|^{1/d}, \quad \text{and} \quad \mathrm{ED}(\Sigma) = 1 - |\mathrm{corr}(\Sigma)|^{1/d}. \tag{35}$$

The EV is the geometric mean of the eigenvalues of the matrix Σ. It provides a dimension free summary of the total variance encoded in the matrix. The ED gets its name because the determinant of a correlation matrix is inversely related to the magnitude of the individual correlations that make up the off-diagonals. In addition to seeing that element-wise distributions match, one would also like to know that their joint distributions correspond. The EV and ED are good summaries of global matrix features and here provide empirical evidence for the validity of gLMC for PD matrices. As we can see, the three methods have similar posterior distributions of EV and ED (Figure 2, bottom panels).

7.2. Learning the spectral density

An important benefit of gLMC is that it enables practitioners to specify prior distributions other than the inverse-Wishart on PD matrices based on needs dictated by the problems at hand. gLMC improves modelling flexibility. We use the problem of Bayesian spectral density estimation to demonstrate the possibility and advantage of using non-conjugate priors. The spectral density matrix $\Sigma(\omega)$ and its coherence matrix $R(\omega)$ are defined in Section 2. In the context of stationary, multivariate time series, the coherences that make up the off-diagonals of $R(\omega)$ provide a lag-free measure of dependence between univariate time series at a given frequency ω. Hence, these coherences are among the more interpratable parameters of the spectral density matrix.

We compare posterior inference for these coherences between two models with different priors: the first model uses the complex inverse-Wishart prior; the second uses the complex reference prior [24]. The reference prior is an improper prior that has been proposed as an alternative to Jeffrey's prior for its superior eigenvalue shrinkage (which improves asymptotic efficiency of estimators). We use the reference prior to emphasize the flexibility allowed by gLMC but not as a modelling suggestion. Svensson and Nordenvaad [24] provide a Gibbs sampling routine based on the eigen-decomposition of the covariance matrix. The reference prior's form is provided in Appendix A.1.

We apply gLMC to learning the spectral density matrix for three distinct 4-dimensional time series. The first is a simulated first-order vector-autoregressive (VAR1) time series with block structure consisting of two independent, 2-dimensional VAR1 time series. The second time series is also VAR1 but with dependencies allowed between all four of the scalar time series of which it is composed. The third time series comes from LFP recorded in the CA1 region of a rat hippocampus [28,29].

Figure 3 shows the first 100 samples from both VAR1 time series along with 95% posterior intervals for the complex moduli of the coherences. The time series are simulated with the form:

$$y(t) = \Phi y(t-1) + \epsilon_t, \quad y(1) = \epsilon_1, \quad \epsilon_t \sim N_4(0, I), \quad t = 1, \ldots, 15000, \tag{36}$$

where the eigenvalues of transition matrix Φ are bounded with absolute value less than 1 to induce stationarity. The first 10,000 data points are discarded to allow time for mixing. The first row of Figure 3 belongs to the block VAR1: Φ is a randomly constructed,

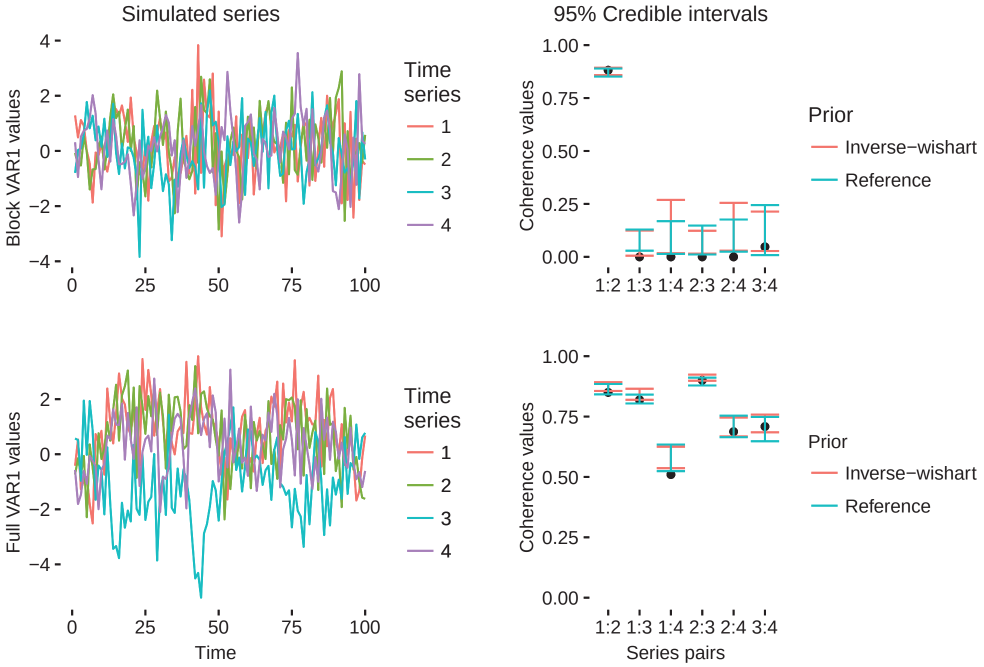

Figure 3. Two 4-dimensional VAR1 time series and credible intervals for their six corresponding coherences measured at 20–40 Hz: the top row belongs to a block VAR1 process characterized by two independent 2-dimensional VAR1 time series; the bottom row belongs to a full VAR1 process. The left column shows the first 100 samples of both time series, each of which totals 5000 samples in length. The right column shows credible intervals from posteriors obtained using the inverse-Wishart and reference priors.

block-diagonal matrix, so the first two scalar time series are independent from the second two. The second row of Figure 3 belongs the the second VAR1, all the scalar time series of which are dependent on all the others. Here Φ is also a randomly constructed matrix but is not block-diagonal. The intervals corresponding to the inverse-Wishart prior model are given in orange. The intervals corresponding to the reference prior model are given in blue. The true coherences are represented by black points and are obtained using the following closed-form formula for the spectral density of a VAR1 process [10]:

$$\Sigma(\omega) = \left(I - \Phi\, \mathrm{e}^{-2\pi i\omega}\right)^{-1} Q \left(I - \Phi\, \mathrm{e}^{-2\pi i\omega}\right)^{-H}. \tag{37}$$

Here Q is the covariance matrix of the additive noise ϵ_t, and $(\cdot)^{-H}$ denotes the inverse conjugate transpose. For the block VAR1 example, both models capture the true, non-null coherences (i.e. those given on the far left and the far right), but neither captures the null coherences. This is more than satisfactory, since coherences equal to zero imply the identity for a covariance matrix. By looking closely, one can see that the first and second time series (orange and green) are indeed strongly dependent on each other, as interval ‘1:2’ suggests. For the full VAR1 example, both models capture five out of six true coherences, but the reference prior model gets closer to the truth than the inverse-Wishart model does.

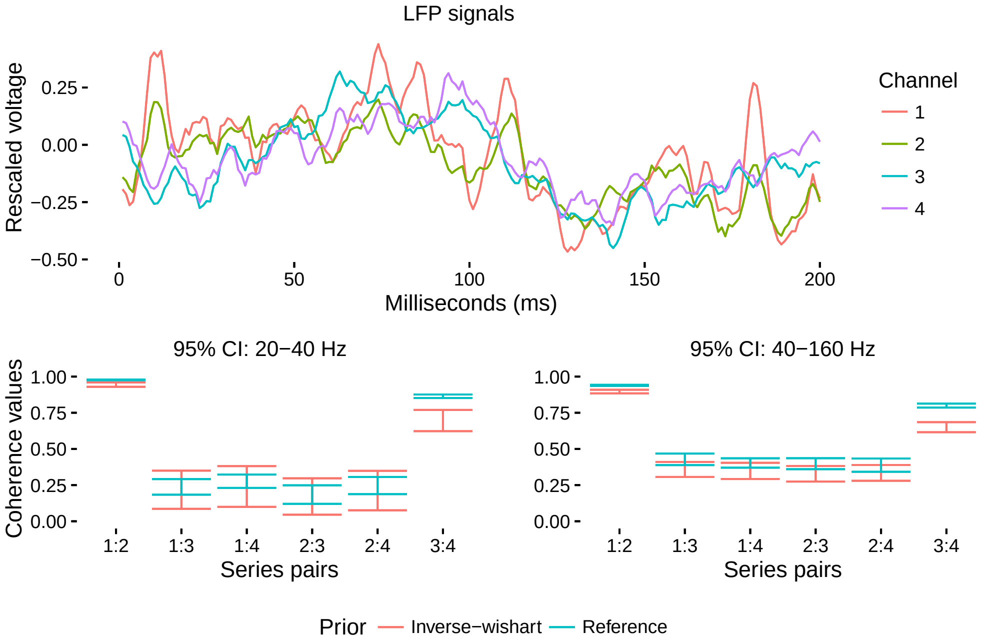

Figure 4. A 4-dimensional LFP signal with credible intervals for six coherences measured at 20–40 Hz (left) and 40–160 Hz (right). First 200 samples are shown for ease of visualization; the multi-dimensional time series totals 4000 samples in length. Coherence profiles are remarkably similar between the two frequency bands considered.

We use the same tools to detect coherences between LFP signals simultaneously recorded from the CA1 region of a rat hippocampus prior to a memory experiment [28]. Two of the LFP signals are recorded on one end of the CA1 axis, and the other two LFP signals are recorded at the opposite end. Figure 4 shows the first 200 of 4000 samples (recorded at 1000 Hz) and 95% credible intervals for the coherences at two different frequency bands: 20–40 Hz and 40–160 Hz. *The spatial discrepancy is reflected in the posterior distributions of the individual coherences.* Both bands show similar coherence patterns, where spatial location appears to dictate strength of coherence: the leftmost and rightmost pairs are closer to each other in space, while thecentre pairs are farther from each other. This reflects what is apparent in the top of Figure 4, where the first and second time series (orange and green) are dependent, and the third and fourth time series (blue and purple) are dependent. These correspond to the intervals labelled ‘1:2’ and ‘3:4’, respectively. The credible intervals are smaller for the 20–40 Hz band because that band has only 1/6 the data of the 40–160 Hz band. Between prior models, the intervals differ more for the 40–160 Hz band. This is counter-intuitive since the influence of the prior distribution is often assumed to diminish with the size of the data set. One question is whether this surprising result is related to the reference prior’s being the prior that is ‘maximally dominated by the data’ [30]. These differences – differences between posterior distributions for the two prior models – communicate that other prior distributions might provide tangible differences between results in spectral analysis and that it would be useful to understand which prior distributions are appropriate in which contexts.

8. Discussion

We presented gLMC an MCMC methodology for Bayesian inference on general Riemannian manifolds. We outlined its relationship to other geometric extensions of HMC and showed how to apply gLMC to both symmetric and Hermitian PD matrices. We demonstrated empirical validity using both element-wise and whole-matrix comparisons against the conjugate inverse-Wishart model. Finally, we applied gLMC on Hermitian PD matrices to Bayesian spectral density estimation. The algorithm proved effective for detecting true coherences of simulated time series, as well as recovering spatial discrepancies between real-world LFP signals.

We see three branches of inquiry stemming from this work: the first is algorithmic; the second, theoretical; the third, methodological. First, what variations of HMC might help extend gLMC over PD matrices into higher dimensions? There are multiple such extensions that are orthogonal to gLMC. Examples are windowed HMC, geometric extensions to the NUTs algorithm, shortcut MCMC, and look-ahead HMC [5,31,32]. Auto-tuning will prove useful: even within the same dimension, different samples will dictate different numbers of leapfrog steps and step-sizes. From the theoretical standpoint, the canonical metric on the space of PD matrices is closely related to the Fisher information metric on covariance matrices: how should one characterize this intersection between information geometry and Riemannian symmetric spaces, and how might this relationship inform Bayes estimator properties or future variations on gLMC? Methodologically, much work needs to be done in prior elicitation for Bayesian spectral density estimation. Which priors on Hermitian PD matrices should be used for which problems, what are the costs and benefits, and are there priors over symmetric PD matrices that need to be complexified (cf. [25,26])? A clear delineation will be useful for practitioners in Bayesian time series research.

Note

1. In this section we focus on the space of Hermitian PD matrices, since the class of symmetric matrices belongs to the broader class of Hermitian matrices. If the reader is primarily interested in the smaller class, then she is free to substitute $\mathbb{R}$ for $\mathbb{C}$, transpose $(\cdot)^{\mathsf{T}}$ for conjugate transpose $(\cdot)^{H}$, and the orthogonal group $O(d)$ for the unitary group $U(d)$.

Acknowledgments

The authors would like to thank Hernando Ombao and Norbert Fortin for their helpful discussions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

AH is supported by NIH grant [T32 AG000096]. SL is supported by the Defense Advanced Research Projects Agency (DARPA) funded program Enabling Quantification of Uncertainty in Physical Systems (EQUiPS), contract W911NF-15-2-0121. AV and BS are supported by National Institutes of Health [grant R01-AI107034] and National Science Foundation [grant DMS-1622490].

ORCID

Andrew Holbrook http://orcid.org/0000-0002-3558-200X
Shiwei Lan http://orcid.org/0000-0002-9167-3715

References

[1] Tokuda T, Goodrich B, VanMechelen I, et al. Visualizing distributions of covariance matrices. New York (NY): Dept Statist, Columbia Univ;2011 (Tech Rep.).

[2] Girolami M, Calderhead B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J R Stat Soc Ser B. 2011;73(2):123–214.

[3] Byrne S, Girolami M. Geodesic monte carlo on embedded manifolds. Scand J Statist. 2013;40(4):825–845.

[4] Lan S, Zhou B, Shahbaba B. Spherical Hamiltonian Monte Carlo for constrained target distributions. In: JMLR workshop and conference proceedings; Vol. 32; NIH Public Access; 2014. p. 629.

[5] Neal RM. Mcmc using Hamiltonian dynamics. In: Steve Brooks, Andrew Gelman, Galin Jones, Xiao-Li Meng, eds. Handbook of Markov Chain Monte Carlo. Vol. 2. London: CRC Press; 2011. p. 113–162.

[6] Lan S, Stathopoulos V, Shahbaba B, et al. Markov chain Monte Carlo from Lagrangian dynamics. J Comput Graph Stat. 2015;24(2):357–378.

[7] Shahbaba B, Lan S, Johnson WO, et al. Split Hamiltonian Monte Carlo. Stat Comput. 2014;24(3):339–349.

[8] Pourahmadi M. Covariance estimation: the GLM and regularization perspectives. Stat Sci. 2011;26:369–387.

[9] Abramovich YI, Besson O. On the expected likelihood approach for assessment of regularization covariance matrix. IEEE Signal Process Lett. 2015;22(6):777–781.

[10] Wang Y, Ho HL. Statistical analysis of electroencephalograms. Boca Raton: CRC Press; 2016; p. 523–565.

[11] Whittle P. The analysis of multiple stationary time series. J R Stat Soc Ser B. 1953;15:125–139.

[12] Brockwell PJ, Davis RA. Time series: theory and methods. New York: Springer Science & Business Media; 2013.

[13] Gao X, Shahbaba B, Fortin N, et al. Evolutionary state-space model and its application to time-frequency analysis of local field potentials; 2016. Available from: arXiv preprint arXiv:161007271.

[14] Pennec X, Fillard P, Ayache N. A Riemannian framework for tensor computing. Int J Comput Vis. 2006;66(1):41–66.

[15] Moakher M, Zéraï M. The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. J Math Imaging Vis. 2011;40(2):171–187.

[16] Helgason S. Differential geometry, lie groups, and symmetric spaces, Vol. 80. New York: Academic Press; 1979.

[17] Koh SS. On affine symmetric spaces. Trans Am Math Soc. 1965;119(2):291–309.

[18] Duane S, Kennedy AD, Pendleton BJ, et al. Hybrid Monte Carlo. Phys Lett B. 1987;195(2): 216–222.

[19] Nash J. The imbedding problem for Riemannian manifolds. Ann Math. 1956;63:20–63.

[20] Lan S, Shahbaba B. Sampling constrained probability distributions using spherical augmentation. Cham: Springer International Publishing; 2016. p. 25–71.

[21] Yang R, Berger JO. Estimation of a covariance matrix using the reference prior. Ann Statist. 1994;22:1195–1211.

[22] Shaman P. The inverted complex Wishart distribution and its application to spectral estimation. J Multivar Anal. 1980;10(1):51–59.

[23] Svensson L, Lundberg M. On posterior distributions for signals in gaussian noise with unknown covariance matrix. IEEE Trans Signal Process. 2005;53(9):3554–3571.

[24] Svensson L, Nordenvaad ML. The reference prior for complex covariance matrices with efficient implementation strategies. IEEE Trans Signal Process. 2010;58(1):53–66.
[25] Schwartzman A. Lognormal distributions and geometric averages of symmetric positive definite matrices. Int Stat Rev. 2015;84(3):456–486.
[26] Fazayeli F, Banerjee A. The matrix generalized inverse gaussian distribution: properties and applications; 2016. Available from: arXiv preprint arXiv:160403463.
[27] Carpenter B, Gelman A, Hoffman M, et al. Stan: a probabilistic programming language. J Stat Softw. 2016;20:1–37.
[28] Allen TA, Salz DM, McKenzie S, et al. Nonspatial sequence coding in ca1 neurons. J. Neurosci. 2016;36(5):1547–1563.
[29] Andrew H, Vandenberg-Rodes A, Fortin N, Shahbaba B. A Bayesian supervised dual-dimensionality reduction model for simultaneous decoding of LFP and spike train signals. Stat. 2017;6(1):53–67.
[30] Berger JO, Bernardo JM, Sun D. The formal definition of reference priors. Ann Statist. 2009;37:905–938.
[31] Sohl-Dickstein J, Mudigonda M, DeWeese MR. Hamiltonian Monte Carlo without detailed balance; 2014. Available from: arXiv preprint arXiv:14095191
[32] Betancourt M. Generalizing the no-u-turn sampler to Riemannian manifolds; 2103. Available from: arXiv preprint arXiv:13041920.
[33] Magnus JR, Neudecker H, Matrix differential calculus with applications in statistics and econometrics;1995.
[34] Hjorungnes A, Gesbert D. Complex-valued matrix differentiation: techniques and key results. IEEE Trans Signal Process. 2007;55(6):2740–2746.

Appendix. Real and complex matrix derivatives

The derivative of a univariate, real valued function with respect to a matrix is most cleanly calculated using the matrix differential. This is true whether $f : M_p(\mathbb{R}) \mapsto \mathbb{R}$ or $f : M_p(\mathbb{C}) \mapsto \mathbb{R}$, that is, whether f is a function over real $p \times p$ matrices or complex $p \times p$ matrices. As an example, we consider the multivariate Gaussian distribution with mean 0 and covariance Σ. First, let f be the probability density function over real valued Gaussian random vectors $y_n \in \mathbb{R}^d$, $n = 1, \dots, N$. Let Y be the $d \times N$ concatenation of these N i.i.d. random variables. Then the log density is given by

$$\begin{aligned}\log f(Y^N, \Sigma) &\propto -\frac{N}{2} \log|\Sigma| - \frac{1}{2} \sum_{n=1}^{N} y_n^{\mathsf{T}} \Sigma^{-1} y_n \\ &= -\frac{N}{2} \log|\Sigma| - \frac{1}{2} \mathrm{tr}\{\Sigma^{-1} Y Y^{\mathsf{T}}\}. \end{aligned} \tag{A1}$$

We apply the matrix differential to (A1) using two general formulas:

$$d \log|\Sigma| = \mathrm{tr}\{\Sigma^{-1} d\Sigma\}, \quad \text{and} \quad d\Sigma^{-1} = -\Sigma^{-1}\,(d\Sigma)\Sigma^{-1}, \tag{A2}$$

rendering

$$\begin{aligned} d \log f(Y, \Sigma) &= -\frac{N}{2} \mathrm{tr}\{\Sigma^{-1} d\Sigma\} + \frac{1}{2} \mathrm{tr}\left\{\Sigma^{-1}(d\,\Sigma)\Sigma^{-1} Y Y^{\mathsf{T}}\right\} \\ &= -\frac{N}{2} \mathrm{tr}\{(d\Sigma)\Sigma^{-1}\} + \frac{1}{2} \mathrm{tr}\left\{(d\Sigma)\Sigma^{-1} Y Y^{\mathsf{T}} \Sigma^{-1}\right\}. \end{aligned} \tag{A3}$$

Finally, we relate the matrix differential to the gradient with the fact that, for an arbitrary function g,

$$dg(\Sigma) = \mathrm{tr}\{(d\Sigma)A\} \quad \Longleftrightarrow \quad \nabla_\Sigma\, g(\Sigma) = A. \tag{A4}$$

This gives the final form of the gradient of the log density function with respect to covariance Σ:

$$\nabla_\Sigma \log f(Y, \Sigma) = -\frac{N}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} Y Y^{\mathsf{T}} \Sigma^{-1}. \tag{A5}$$

For more on the matrix differential, see [33]. The complex matrix differential is treated in [34] and has a similar form real valued functions. The log density of the multivariate complex Gaussian with mean 0 is given by

$$\begin{aligned}\log f(Y,\Sigma) &\propto -N\log|\Sigma| - \sum_{n=1}^{N} y_n^H \Sigma^{-1} y_n \\ &= -N\log|\Sigma| - \operatorname{tr}\{\Sigma^{-1}YY^H\},\end{aligned} \tag{A6}$$

where $(\cdot)^H$ denotes the conjugate transpose. Note that the log density is scaled by a factor of two compared to the real case. The resulting gradient is

$$\nabla_\Sigma \log f(Y,\Sigma) = -N\Sigma^{-1} + \Sigma^{-1}YY^H\Sigma^{-1}. \tag{A7}$$

A.1 The complex reference prior

Gradients of prior probabilities are calculated in a similar way. We demonstrate for the complex reference prior. Let $\lambda_i,\ i = 1,\dots,d$ be the decreasing eigenvalues of Hermitian PD matrix Σ. Then the complex reference prior has the following form:

$$p(\Sigma) \propto \frac{\mathrm{d}\Sigma}{|\Sigma|\prod_{k<j}(\lambda_k-\lambda_j)^2}. \tag{A8}$$

To use the above approach for deriving the matrix derivatives, we need to be able to write the differential $\mathrm{d}\lambda_i$ in terms of the matrix differential $\mathrm{d}\Sigma$. Magnus et al. [33] provides the formula when all eigenvalues are distinct:

$$d\lambda_i = \operatorname{tr}\left(\sum_{j=1}^{d} V_{ij}^{-1}\Sigma^{j-1}d\Sigma\right), \tag{A9}$$

where V is the Vandermonde matrix:

$$V^{\mathsf{T}} = \begin{vmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-2} & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-2} & \lambda_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{n-2} & \lambda_n^{n-1} \end{vmatrix}. \tag{A10}$$

We now calculate the gradient of the log of the complex reference prior:

$$\begin{aligned} d\log p(\Sigma) &= -d\log|\Sigma| - 2\sum_{k<j} d\log(\lambda_k-\lambda_j) \\ &= -\operatorname{tr}(\Sigma^{-1}d\Sigma) - 2\sum_{k<j}\frac{d\lambda_k - d\lambda_j}{\lambda_k-\lambda_j} \\ &= -\operatorname{tr}(\Sigma^{-1}d\Sigma) - 2\sum_{k<j}\operatorname{tr}\left(\sum_{i=1}^{d}\left(V_{ki}^{-1}-V_{ji}^{-1}\right)\Sigma^{i-1}d\Sigma\right)/(\lambda_k-\lambda_j). \end{aligned} \tag{A11}$$

Combining this with Equations (A2) and (A4) renders matrix gradient

$$\nabla_\Sigma \log p(\Sigma) \propto -\Sigma^{-1} - 2\sum_{k<j}\left(\sum_{i=1}^{d}\left(V_{ki}^{-1}-V_{ji}^{-1}\right)\Sigma^{i-1}\right)/(\lambda_k-\lambda_j). \tag{A12}$$