### Title

Flexible Bayesian Dynamic Modeling of Correlation and Covariance Matrices




### Journal

Bayesian Analysis, 15(4)

### ISSN

1936-0975



### Authors

Lan, Shiwei
Holbrook, Andrew
Elias, Gabriel A
et al.



### Publication Date

2020-12-01

### DOI

10.1214/19-ba1173

Peer reviewed

**ORIGINAL ARTICLE**

**Journal Section**

# Flexible Bayesian Dynamic Modeling of Covariance and Correlation Matrices

**Shiwei Lan[1]* | Andrew Holbrook[2] | Norbert J. Fortin[3] | Hernando Ombao[2,4] | Babak Shahbaba[2]**

[1]Department of Computing + Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125

[2]Department of Statistics, University of California-Irvine, Irvine, CA 92697

[3]Center for the Neurobiology of Learning and Memory, Department of Neurobiology and Behavior, University of California-Irvine, Irvine, CA 92697

[4]Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

**Correspondence**
Shiwei Lan, Department of Computing + Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125
Email: slan@caltech.edu

**Funding information**
SL is supported by ONR N00014-17-1-2079.
AH is supported by NIH grant T32 AG000096.
NJF is supported by NSF awards IOS-1150292 and BCS-1439267 and Whitehall Foundation award 2010-05-84.
HO is supported by NSF DMS 1509023 and NSF SES 1461534.
BS is supported by NSF DMS 1622490 and NIH R01 AI107034.

Modeling covariance (and correlation) matrices is a challenging problem due to the large dimensionality and positive-definiteness constraint. In this paper, we propose a novel Bayesian framework based on decomposing the covariance matrix into variance and correlation matrices. The highlight is that the correlations are represented as products of vectors on unit spheres. We propose a variety of distributions on spheres (e.g. the squared-Dirichlet distribution) to induce flexible prior distributions for covariance matrices that go beyond the commonly used inverse-Wishart prior. To handle the intractability of the resulting posterior, we introduce the adaptive $\Delta$-Spherical Hamiltonian Monte Carlo. We also extend our structured framework to dynamic cases and introduce unit-vector Gaussian process priors for modeling the evolution of correlation among multiple time series. Using an example of Normal-Inverse-Wishart problem, a simulated periodic process, and an analysis of local field potential data (collected from the hippocampus of rats performing a complex sequence memory task), we demonstrated the validity and effectiveness of our proposed framework for (dynamic) modeling covariance and correlation matrices.

**KEYWORDS**
Dynamic covariance modeling, Separation strategy, Geometric methods, $\Delta$-Spherical Hamiltonian Monte Carlo

## 1 | INTRODUCTION

Modeling covariance matrices—or more broadly, positive definite (PD) matrices—is one of the most fundamental problems in statistics. In general, the task is difficult because the number of parameters grows quadratically as the number of variables increases. The complexity of the challenge increases substantially if we allow dependencies to vary over time (or space) in order to account for the dynamic (non-stationary) nature of the underlying probability model. In this paper, we propose a novel solution to the problem by developing a flexible and yet computationally efficient inferential framework for both fixed and dynamic covariance matrices.

Within the Bayesian framework, it is common to use an Inverse-Wishart prior on the covariance matrix for computational convenience (Mardia et al., 1980; Anderson, 2003). This choice of prior however is very restrictive (e.g. common degrees of freedom for all components of variance) (Barnard et al., 2000; Tokuda et al., 2011). Alternatively, one may use decomposition strategies for more flexible modeling choices (see Barnard et al. (2000) for more details). For instance, Banfield and Raftery (1993), Yang and Berger (1994), Celeux and Govaert (1995), Leonard and Hsu (1992), Chiu et al. (1996), and Bensmail et al. (1997) proposed methods based on the spectral decomposition of the covariance matrix. Another strategy is to use the Cholesky decomposition of the covariance matrix or its inverse, e.g., Pourahmadi (1999, 2000); Liu (1993); Pinheiro and Bates (1996). In general, these methods fail to yield full flexibility (Pourahmadi, 1999, 2000, for a special type of 'joint mean-covariance model'), generality (Liu, 1993, for applications in 'monotone missing data'), or sacrifice for statistical interpretability (Pinheiro and Bates, 1996). The proposed method in this paper is based on the separation strategy (Barnard et al., 2000) and the Cholesky decomposition for separately modeling mean, variance and correlation in a fully flexible way.

To model the real-life spatio-temporal processes with complex dependence structures (e.g., brain signals during cognitive tasks), we extend our structured modeling framework for covariance and correlation matrices to the dynamic setting. To address the constraint for correlation processes (positive definite matrix at each time having unit diagonals and off-diagonal entries with magnitudes no greater than 1), we introduce unit-vector Gaussian process priors. There are other related works, e.g. generalized Wishart process (Wilson and Ghahramani, 2011), and latent factor process (Fox and Dunson, 2015), that explore the product of vector Gaussian processes. In general they do not grant full flexibility in modeling all of the mean, variance and correlation processes, and thus are outperformed by our proposed flexible framework (See more details in Section 4.2.2). Other approaches to model non-stationary processes use a representation in terms of a basis such as wavelets (Nason et al., 2000), the SLEX (Ombao et al., 2005), which are actually inspired by Fourier representations in Priestley (1965) and Dahlhaus (2000). These approaches are frequentist and do not easily provide a framework for inference (confidence intervals and hypothesis testing). The class of time-domain parametric models allows for the ARMA parameters to evolve over time (see, e.g. Rao, 1970) or via parametric latent signals (West et al., 1999; Prado et al., 2001). A restriction for this class of models is that some processes might not be adequately modeled by these parametric models.

This paper contains the following contributions:

- a general and flexible framework is proposed for modeling mean, variance and correlation processes separately;
- the sphere-product representation of a correlation matrix is introduced and distributions based on this representation, including the squared-Dirichlet distribution and normalized Gaussian distributions, are discussed;
- a new stochastic process is defined on the sphere and used to induce dynamic priors for covariance matrices;
- an efficient algorithm is introduced to infer correlation matrices and processes;
- posterior contraction phenomena are studied for both mean and covariance (correlation) processes.

There is a strong potential impact of this work in modeling the dynamic connectivity of brain processes. Brain connectivity plays a prominent role in neuroscience because it is associated with cognitive processes such as learning and memory and it is considered as a potential biomarker for mental and neurological diseases. There are a number of new methods that have been developed (Cribben et al., 2012; Fiecas and Ombao, 2016; Lindquist et al., 2014; Ting et al., 2015; Prado, 2013) but the main limitation of these methods (especially the ones that have a frequentist approach) is a lack of natural framework for inference. Moreover, parametric approaches (e.g. vector auto-regressive models) need to be tested for adequacy for modeling complex brain processes and often have high dimensional parameter spaces (especially with a large number of channels and high lag order). This work provides both a nonparametric Bayesian model and an efficient inferential method for modeling the complex dynamic dependence among multiple stochastic processes that is common in the study of brain connectivity.

The rest of the paper is organized as follows. In the next section, we present a geometric view of covariance matrices. Next, we will extend this view to allow covariance matrices to change over time. In Section 3, we will use this geometrical perspective to develop an effective and computationally efficient inferential method for modeling static and dynamic covariance matrices. Using simulated data, we will evaluate our method in Section 4. Then we apply our proposed method to local field potentials (LFP) data recorded from the hippocampus of rats performing a complex sequence memory task. In the final section, we conclude with discussions about future work.

## 2 | STRUCTURED BAYESIAN MODELING COVARIANCES (CORRELATIONS)

To derive flexible models for covariance and correlation matrices, we start with the Cholesky decomposition, form a sphere-product representation, and finally obtain the separation decomposition in Barnard et al. (2000) with correlations represented as products of vectors on spheres. The sphere-product representation is amenable for the inferential algorithm to handle the resulting intractability, and hence lays the foundation for full flexibility in choosing priors.

Any covariance matrix $\boldsymbol{\Sigma} = [\sigma_{ij}]$ is symmetric positive definite, and hence has a unique Cholesky decomposition

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^{\mathsf{T}}, \qquad \sigma_{ij} = \sum_{k=1}^{\min\{i,j\}} l_{ik} l_{jk}$$

where the Cholesky factor $\mathbf{L}$ is a lower triangular matrix. We denote the variance vector as $\boldsymbol{\sigma}^2 := [\sigma_1^2, \cdots, \sigma_D^2]^{\mathsf{T}}$, then each variance component, $\sigma_i^2$, can be written in terms of $\mathbf{L}$ as follows:

$$\sigma_i^2 := \sigma_{ii} = \sum_{k=1}^{i} l_{ik}^2 = \|\mathbf{l}_i\|^2, \quad \mathbf{l}_i := [l_{i1}, l_{i2}, \cdots, l_{ii}], \quad \mathbf{L} = \begin{bmatrix} \mathbf{l}_1 \\ \vdots \\ \mathbf{l}_D \end{bmatrix} \tag{1}$$

For $\boldsymbol{\Sigma}$ to be positive definite, it is equivalent to require all the leading principal minors $\{M_i\}$ to be positive,

$$M_i = \prod_{k=1}^{i} l_{kk}^2 > 0, \ i = 1, \cdots, D \iff l_{ii} \neq 0, \ i = 1, \cdots, D \tag{2}$$

Based on (1) and (2), $\mathbf{l}_i$, for $i \in \{1, \cdots, D\}$, is restricted to be on a sphere with radius $\sigma_i := \sqrt{\sigma_{ii}}$ excluding the equator, denoted as $\mathcal{S}_0^{i-1}(\sigma_i)$. Therefore the space of the Cholesky factor in terms of its rows can be written as a product of

spheres and we require

$$(\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_D) \in \mathcal{S}_0^0(\sigma_1) \times \mathcal{S}_0^1(\sigma_2) \cdots \times \mathcal{S}_0^{D-1}(\sigma_D) \tag{3}$$

Note that (3) is the sufficient and necessary condition for the matrix $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\mathsf{T}$ to be a covariance matrix. The norm requirements for correlations are immediately satisfied by the Cauchy-Schwarz inequality: $|\rho_{ij}| := \frac{|\sigma_{ij}|}{\sigma_i \sigma_j} = \frac{|\langle \mathbf{l}_i, \mathbf{l}_j \rangle|}{\|\mathbf{l}_i\| \|\mathbf{l}_j\|} \leq 1$.

For simplicity, we present probabilistic models involving covariance matrices in the following generic form:

$$\begin{aligned} \mathbf{y}|\boldsymbol{\Sigma}(\boldsymbol{\sigma}, \mathbf{L}) &\sim \ell(\mathbf{y}; \boldsymbol{\Sigma}(\boldsymbol{\sigma}, \mathbf{L})), \quad \boldsymbol{\Sigma}(\boldsymbol{\sigma}, \mathbf{L}) = \mathbf{L}\mathbf{L}^\mathsf{T} \\ \boldsymbol{\sigma} &\sim p(\boldsymbol{\sigma}) \\ \mathbf{L}|\boldsymbol{\sigma} &\sim p(\mathbf{L}; \boldsymbol{\sigma}), \quad \mathrm{vech}^\mathsf{T}(\mathbf{L}) \in \prod_{i=1}^{D} \mathcal{S}_0^{i-1}(\sigma_i) \end{aligned} \tag{4}$$

where $\boldsymbol{\sigma} := [\sigma_1, \cdots, \sigma_D]^\mathsf{T}$, and the half-vectorization in row order, $\mathrm{vech}^\mathsf{T}$, transforms the lower triangular matrix $\mathbf{L}$ into a vector $(\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_D) = (l_{11}, l_{21}, l_{22}, \cdots, l_{D1}, \cdots, l_{DD})$. The total dimension of $(\boldsymbol{\sigma}, \mathbf{L})$ is $\frac{D(D+1)}{2}$.[1]

Alternatively, if we separate variances from covariance, then we have a unique Cholesky decomposition for the following correlation matrix $\mathbf{P} = [\rho_{ij}]$

$$\mathbf{P} := \mathrm{diag}(\boldsymbol{\sigma}^2)^{-\frac{1}{2}} \boldsymbol{\Sigma}\, \mathrm{diag}(\boldsymbol{\sigma}^2)^{-\frac{1}{2}} = \mathbf{L}^*(\mathbf{L}^*)^\mathsf{T}, \qquad \rho_{ij} = \sum_{k=1}^{\min\{i,j\}} l_{ik}^* l_{jk}^*$$

where the Cholesky factor $\mathbf{L}^* = \mathrm{diag}(\boldsymbol{\sigma}^{-1})\mathbf{L}$ can be obtained by normalizing each row of $\mathbf{L}$. Thus we require

$$(\mathbf{l}_1^*, \mathbf{l}_2^*, \cdots, \mathbf{l}_D^*) \in \mathcal{S}_0^0 \times \mathcal{S}_0^1 \cdots \times \mathcal{S}_0^{D-1} \tag{5}$$

where $\mathcal{S}_0^{i-1} := \mathcal{S}_0^{i-1}(1)$. Similarly, (5) is the sufficient and necessary condition for $\mathbf{P} = \mathbf{L}^*(\mathbf{L}^*)^\mathsf{T}$ to be a correlation matrix. Then we have the following alternatively structured model for covariance $\boldsymbol{\Sigma}$ that involves correlation $\mathbf{P}$ explicitly

$$\begin{aligned} \mathbf{y}|\boldsymbol{\Sigma}(\boldsymbol{\sigma}, \mathbf{L}^*) &\sim \ell(\mathbf{y}; \boldsymbol{\Sigma}(\boldsymbol{\sigma}, \mathbf{L}^*)), \quad \boldsymbol{\Sigma}(\boldsymbol{\sigma}, \mathbf{L}^*) = \mathrm{diag}(\boldsymbol{\sigma})\mathbf{P}\,\mathrm{diag}(\boldsymbol{\sigma}), \quad \mathbf{P} = \mathbf{L}^*(\mathbf{L}^*)^\mathsf{T} \\ \boldsymbol{\sigma} &\sim p(\boldsymbol{\sigma}) \\ \mathbf{L}^* &\sim p(\mathbf{L}^*), \quad \mathrm{vech}^\mathsf{T}(\mathbf{L}^*) \in \prod_{i=1}^{D} \mathcal{S}_0^{i-1} \end{aligned} \tag{6}$$

Note, this direct decomposition $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma})\mathbf{P}\,\mathrm{diag}(\boldsymbol{\sigma})$ as a *separation strategy* is motivated by statistical thinking in terms of standard deviations and correlations (Barnard et al., 2000). This setting is especially relevant if the statistical quantity of interest is correlation matrix $\mathbf{P}$ itself, and we can then skip inference of the standard deviation $\boldsymbol{\sigma}$ by fixing it to a data-derived point estimate.

In what follows, we will show that the above framework includes the Inverse-Wishart prior as a special case, but it can be easily generalized to a broader range of priors for additional flexibility. Such flexibility enables us to better express prior knowledge, control the model complexity and speed up computation (Section 4.2.3) in modeling real-life phenomena. This is crucial in modeling spatio-temporal processes with complex structures.

[1] For each $i \in \{1, \cdots, D\}$, given $\sigma_i$, there are only $(i-1)$ free parameters on $\mathcal{S}_0^{i-1}(\sigma_i)$, so there are totally $\frac{D(D-1)}{2} + D$ free parameters.

## 2.1 | Connection to the Inverse-Wishart Prior

There are some interesting connections between the spherical product representations (3) (5) and the early development of the Wishart distribution (Wishart, 1928). The original Wishart distribution was derived by orthogonalizing multivariate Gaussian random variables leading to a lower triangular matrix whose elements $\{t^*_{ij}|i \geq j\}$ (analogous to $l_{ij}$ or $l^*_{ij}$) were called *rectangular coordinates*. This way, the probability density has a geometric interpretation as a product of volumes and approximate densities on a series of spherical shells with radius $\{t^*_{ii}\}$ (See more details in Sverdrup, 1947; Anderson, 2003). Now we demonstrate that the proposed schemes (4) (6) include the commonly used inverse-Wishart prior as a special case in modeling covariances.

Suppose $\boldsymbol{\Sigma}$ is a random sample from the inverse-Wishart distribution $\mathcal{W}_D^{-1}(\boldsymbol{\Psi}, \nu)$, where $\boldsymbol{\Psi} > 0$ is the scale matrix and $\nu \geq D$ is the degree of freedom. Therefore, $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}_D(\boldsymbol{\Psi}^{-1}, \nu)$. Denote $\mathbf{C}$ as the Cholesky factor of $\boldsymbol{\Psi}^{-1}$, i.e. $\boldsymbol{\Psi}^{-1} = \mathbf{C}\mathbf{C}^\mathsf{T}$. Then $\boldsymbol{\Sigma}^{-1}$ has the following Bartlett decomposition (Anderson, 2003; Smith and Hocking, 1972)

$$\boldsymbol{\Sigma}^{-1} = \mathbf{T}\mathbf{T}^\mathsf{T}, \quad \mathbf{T} := \mathbf{C}\mathbf{T}^*, \quad t^*_{ij} \sim \begin{cases} \chi_{D-i+1}, & i = j \\ \mathcal{N}(0,1), & i > j \\ \delta_0, & i < j \end{cases} \tag{7}$$

where the lower triangular matrix $\mathbf{T}$, the *Bartlett factor*, has the following density (Theorem 7.2.1 of Anderson, 2003)

$$p(\mathbf{T}) = \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{D(\nu-2)/2}\Gamma_D(\nu/2)} \prod_{i=1}^{D} |t_{ii}|^{\nu-i} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\mathbf{T}\mathbf{T}^\mathsf{T})\right)$$

with multivariate gamma function defined as $\Gamma_D(x) := \pi^{D(D-1)/4} \prod_{i=1}^{D} \Gamma[x + (1-i)/2]$.

Now taking the inverse of the first equation in (7) yields the following *reversed Cholesky decomposition*[2]

$$\boldsymbol{\Sigma} = \mathbf{U}\mathbf{U}^\mathsf{T}, \qquad \sigma_{ij} = \sum_{k=\max\{i,j\}}^{D} u_{ik}u_{jk}, \quad \mathrm{vech}(\mathbf{U}^\mathsf{T}) \in \prod_{i=1}^{D} \mathcal{S}_0^{D-i}(\sigma_i)$$

where $\mathbf{U} := \mathbf{T}^{-\mathsf{T}}$ is an upper triangular matrix. The following theorem describes the density of the reversed Cholesky factor $\mathbf{U}$ of $\boldsymbol{\Sigma}$, which enables us to treat the inverse-Wishart distribution as a special instance of strategy (4) or (6).

**Theorem 1** *Assume $\boldsymbol{\Sigma} \sim \mathcal{W}_D^{-1}(\boldsymbol{\Psi}, \nu)$. Then its reversed Cholesky factor $\mathbf{U}$ has the following density*

$$p(\mathbf{U}) = \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{D(\nu-2)/2}\Gamma_D(\nu/2)} |\mathbf{U}|^{-(\nu+D+1)} \prod_{i=1}^{D} u_{ii}^{i} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\mathbf{U}^{-\mathsf{T}}\mathbf{U}^{-1})\right)$$

**Proof** See Appendix A.

**Remark 1** *If we normalize each row of $\mathbf{U}$ and write*

$$\mathbf{U} = \mathrm{diag}(\boldsymbol{\sigma})\mathbf{U}^*, \quad \sigma_i = \sqrt{\sigma_{ii}} = \|\mathbf{u}_i\|, \quad u^*_{ij} = u_{ij}/\sigma_i,$$

[2] This can be achieved through the *exchange matrix* (a.k.a. reversal matrix, backward identity, or standard involutory permutation) $\mathbf{E}$ with 1's on the anti-diagonal and 0's elsewhere. Note that $\mathbf{E}$ is both involutory and orthogonal, i.e. $\mathbf{E} = \mathbf{E}^{-1} = \mathbf{E}^\mathsf{T}$. Let $\mathbf{E}\boldsymbol{\Sigma}\mathbf{E} = \mathbf{L}\mathbf{L}^\mathsf{T}$ be the usual Cholesky decomposition. Then $\boldsymbol{\Sigma} = (\mathbf{E}\mathbf{L}\mathbf{E})(\mathbf{E}\mathbf{L}\mathbf{E})^\mathsf{T} = \mathbf{U}\mathbf{U}^\mathsf{T}$ and define $\mathbf{U} := \mathbf{E}\mathbf{L}\mathbf{E}^\mathsf{T}$.

*then the following joint prior of* $(\boldsymbol{\sigma}, \mathbf{U}^*)$ *is inseparable in general:*

$$p(\boldsymbol{\sigma}, \mathbf{U}^*) \propto \prod_{i=1}^{D} |\sigma_i u_{ii}^*|^{i-(v+D+1)} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(\boldsymbol{\sigma}^{-1})(\mathbf{U}^*)^{-\mathsf{T}}(\mathbf{U}^*)^{-1}\,\mathrm{diag}(\boldsymbol{\sigma}^{-1}))\right\} \tag{8}$$

*With this result, we can conditionally model variance and correlation factor as* $p(\boldsymbol{\sigma}|\mathbf{U}^*)$ *and* $p(\mathbf{U}^*|\boldsymbol{\sigma})$ *respectively, similarly as in our proposed scheme* (4) *or* (6)*. This result facilitates the construction of a broader class of more flexible prior distributions for covariance matrix in the following section 2.2. It is also used to verify the validity of our proposed method* (6) *(see more details in Section 4.1.1). A similar result exists for the Wishart prior distribution regarding the Cholesky factor.*

## 2.2 | More Flexible Priors

Within the above framework, the only constraint on $\mathbf{U}$ or $\mathbf{L}$ is that it resides on the product of spheres with increasing dimensions. Using this fact, we can develop a broader class of priors on covariance matrices and thus be able to model processes with more complicated dependence in covariance structures. Since $\boldsymbol{\sigma}$ and $\mathbf{L}^*$ have independent priors in (6), in what follows we focus on the scheme (6), and for simplicity, we denote the normalized Cholesky factor as $\mathbf{L}$. Also, following Barnard et al. (2000), we assume a log-Normal prior on $\sigma$:

$$\log(\boldsymbol{\sigma}) \sim \mathcal{N}(\boldsymbol{\xi}, \boldsymbol{\Lambda})$$

We now discuss priors on $\mathbf{L}$ that properly reflect the prior knowledge regarding the covariance structure among variables. If two variables, $y_i$ and $y_j$ (assuming $i < j$) are known to be uncorrelated a priori, then can choose a prior that encourages $\mathbf{l}_i$ and $\mathbf{l}_j$ to be perpendicular to each other, e.g. $l_{jk} \approx 0$ for $k = 1, \cdots, i$. In contrast, if we believe *a priori* that there is a strong correlation between the two variables, we can specify that $\mathbf{l}_i$ and $\mathbf{l}_j$ be linearly dependent, e.g., by setting $l_{jk}|\mathbf{l}_i \approx l_{ik},\ k = 1, \cdots, i$ or $l_{jk}|\mathbf{l}_i \approx -l_{ik},\ k = 1, \cdots, i$. When there is no prior information, we might assume that components are uncorrelated and specify priors for $\mathbf{l}_i$ that concentrate on the (two) poles of $\mathcal{S}_0^{i-1}$,

$$p(\mathbf{l}_i) \propto |l_{ii}|, \quad i = 2, \cdots, D \tag{9}$$

Putting more prior probability on the diagonal elements of $\mathbf{L}$ renders fewer non-zero off-diagonal elements, which in turn leads to a larger number of perpendicular variables; that is, such a prior favors zeros in the correlation matrix $\mathbf{P}$. More generally, one can map a probability distribution defined on the simplex onto the sphere. We consider the following *squared-Dirichlet* distribution:

**Definition 1 (Squared-Dirichlet distribution)** *A random vector* $\mathbf{l}_i \in \mathcal{S}^{i-1}$ *is said to have a* squared-Dirichlet *distribution with parameter* $\boldsymbol{\alpha}_i := (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{ii})$ *if*

$$\mathbf{l}_i^2 := (l_{i1}^2, l_{i2}^2, \cdots, l_{ii}^2) \sim \mathrm{Dir}(\boldsymbol{\alpha}_i)$$

*Then we denote* $\mathbf{l}_i \sim \mathrm{Dir}^2(\boldsymbol{\alpha}_i)$ *and* $\mathbf{l}_i$ *has the following density*

$$p(\mathbf{l}_i) = p(\mathbf{l}_i^2)|2\mathbf{l}_i| \propto (\mathbf{l}_i^2)^{\boldsymbol{\alpha}_i - 1}|\mathbf{l}_i| = |\mathbf{l}_i|^{2\boldsymbol{\alpha}_i - 1} := \prod_{k=1}^{i} |l_{ik}|^{2\alpha_{ik}-1} \tag{10}$$

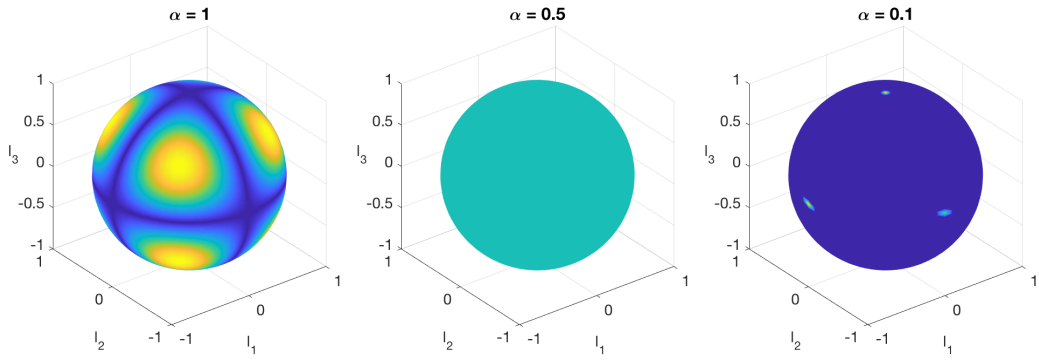


**FIGURE 1** Symmetric squared-Dirichlet distributions $\text{Dir}^2(\boldsymbol{\alpha})$ defined on the 2-sphere with different settings for concentration parameter $\boldsymbol{\alpha} = \alpha\mathbf{1}$. The uniform distribution on the simplex, $\text{Dir}(\mathbf{1})$, becomes non-uniform on the sphere due to the stretch of geometry (left); the symmetric Dirichlet distribution $\text{Dir}(\frac{1}{2}\mathbf{1})$ becomes uniform on the sphere (middle); with $\alpha$ closer to 0, the induced distribution becomes more concentrated on the polar points (right).

**Remark 2** *This definition includes a large class of flexible prior distributions on the unit sphere that specify different concentrations of probability density through the parameter $\boldsymbol{\alpha}_i$. For example, the above prior* (9) *corresponds to* $\boldsymbol{\alpha}_i = (\frac{1}{2}, \cdots, \frac{1}{2}, 1)$.

To induce a prior distribution for the correlation matrix $\mathbf{P} = \mathbf{L}\mathbf{L}^\mathsf{T}$, one can specify priors on row vectors of $\mathbf{L}$, $\mathbf{l}_i \sim \text{Dir}^2(\boldsymbol{\alpha}_i)$ for $i = 2, \cdots, D$. To encourage small correlation, we choose the concentration parameter $\boldsymbol{\alpha}_i$ so that the probability density concentrates around the (two) poles of $\mathcal{S}_0^{i-1}$, e.g. $0 < \alpha_{ik} \ll \alpha_{ii}$ for $k < i$. Figure 1 illustrates the density heat maps of some symmetric squared-Dirichlet distributions $\text{Dir}^2(\alpha\mathbf{1})$ on the 2-sphere. It is interesting that the squared-Dirichlet distribution induces two important uniform prior distributions over correlation matrices from Barnard et al. (2000) in an effort to provide flexible priors for covariance matrices, as stated in the following theorem.

**Theorem 2 (Uniform distributions)** *Let* $\mathbf{P} = \mathbf{L}\mathbf{L}^\mathsf{T}$. *Suppose* $\mathbf{l}_i \sim \text{Dir}^2(\boldsymbol{\alpha}_i)$, *for* $i = 2, \cdots, D$, *are independent, where* $\mathbf{l}_i$ *is the* $i$*-th row of* $\mathbf{L}$. *Then we have*

1. *If* $\boldsymbol{\alpha}_i = (\frac{1}{2}\mathbf{1}_{i-1}^\mathsf{T}, \alpha_{ii})$, $\alpha_{ii} = \frac{(i-2)D-1}{2}$, *then* $\mathbf{P}$ *follows a* marginally *uniform distribution, that is,* $\rho_{ij} \sim \text{Unif}(-1, 1)$, $\quad i \neq j$.
2. *If* $\boldsymbol{\alpha}_i = (\frac{1}{2}\mathbf{1}_{i-1}^\mathsf{T}, \alpha_{ii})$, $\alpha_{ii} = \frac{D-i}{2} + 1$, *then* $\mathbf{P}$ *follows a* jointly *uniform distribution, that is,* $p(\mathbf{P}) \propto 1$.

**Proof** See Appendix A.

Another natural spherical prior can be obtained by normalizing a multivariate Gaussian random vector. This is later generalized to a vector Gaussian process constrained to a sphere that serves as a suitable prior for modeling correlation processes. Now we consider the following *normalized multivariate Gaussian* distribution:

**Definition 2 (Normalized Gaussian distribution)** *A random vector* $\mathbf{l}_i \in \mathcal{S}^{i-1}$ *is said to have a* normalized multivariate Gaussian *distribution with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$ *if*

$$\mathbf{l}_i := \frac{\mathbf{z}}{\|\mathbf{z}\|_2}, \quad \mathbf{z} \sim \mathcal{N}_i(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

*Then we denote* $\mathbf{l}_i \sim n\text{-}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\mathbf{l}_i$ *has the following (conditional) density*

$$p(\mathbf{l}_i \mid \|\mathbf{z}\|_2 = \sigma_i) = p_{\mathcal{N}}(\mathbf{z}) \left| \frac{d\mathbf{z}}{d\mathbf{l}_i} \right| = p_{\mathcal{N}}(\mathbf{l}_i \sigma_i) \sigma_i^i = \frac{\sigma_i^i}{(2\pi)^{\frac{i}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\mathbf{l}_i \sigma_i - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{l}_i \sigma_i - \boldsymbol{\mu}) \right\}$$

**Remark 3** *This conditional density essentially defines the following* Fisher-Bingham *distribution (a.k.a.* generalized Kent *distribution, Kent, 1982; Mardia and Jupp, 2009). If* $\boldsymbol{\Sigma} = \mathbf{I}$*, then the above distribution reduces to the* von Mises-Fisher *distribution (Fisher, 1953; Mardia and Jupp, 2009) as a special case. If in addition* $\boldsymbol{\mu} = 0$*, then the above density becomes a constant; that is, the corresponding distribution is uniform on the sphere* $\mathcal{S}_0^{i-1}$.

**Definition 3 (Fisher-Bingham / Kent distribution)** *The probability density function of the Kent distribution for the random vector* $\mathbf{l}_i \in \mathcal{S}^{i-1}$ *is given by*

$$p(\mathbf{l}_i) \propto \exp\left\{ \kappa \boldsymbol{\gamma}_1^{\mathsf{T}} \mathbf{l}_i + \sum_{k=2}^{i} \beta_k (\boldsymbol{\gamma}_k^{\mathsf{T}} \mathbf{l}_i)^2 \right\}$$

*where* $\sum_{k=2}^{i} \beta_k = 0$ *and* $0 \leq 2|\beta_k| < \kappa$ *and the vectors* $\{\boldsymbol{\gamma}_k\}_{k=1}^{i}$ *are orthonormal.*

**Remark 4** *The parameters* $\kappa$ *and* $\boldsymbol{\gamma}_1$ *are called the* concentration *and the* mean direction *parameter, respectively. The greater the value of* $\kappa$*, the higher the concentration of the distribution around the mean direction* $\boldsymbol{\gamma}_1$*. The choice of* $\boldsymbol{\gamma}_1$ *could impact our priors when modeling correlations (see Section 4.1.2 for more details). Parameters* $\{\beta_k\}_{k=2}^{i}$ *determine the ellipticity of the contours of equal probability. The vectors* $\{\boldsymbol{\gamma}_k\}_{k=2}^{i}$ *determine the orientation of the equal probability contours on the sphere.*

**Remark 5** *If* $\beta_k = 0$ *for* $k = 2, \cdots, i$*, then this distribution reduces to* von Mises-Fisher *distribution (Fisher, 1953; Mardia and Jupp, 2009), denoted as* $\text{vMF}(\kappa, \boldsymbol{\gamma}_1)$*. If* $\kappa = 0$*, then it defines an antipodally symmetric distribution, named* Bingham *distribution (Bingham, 1974), denoted as* $\text{Bing}(\mathbf{A})$*, with* $\mathbf{l}_i^{\mathsf{T}} \mathbf{A} \mathbf{l}_i = \sum_{k=2}^{i} \beta_k (\boldsymbol{\gamma}_k^{\mathsf{T}} \mathbf{l}_i)^2$.

As before, to induce smaller correlations, one can put higher prior probabilities for $\mathbf{l}_i$ on the poles of $\mathcal{S}^{i-1}$. For example, we might consider $\mathbf{l}_i \sim \text{vMF}(\kappa, \mathbf{n}_i)$, or $\mathbf{l}_i \sim \text{Bing}(\zeta\, \text{diag}(\mathbf{n}_i))$, where $\mathbf{n}_i := (0, \cdots, 0, 1)^{\mathsf{T}}$ is denoted as the north pole. We will explore the effect of $\kappa$ and $\zeta$ on spherical priors for correlations in Section 4.1.2.

## 2.3 | Dynamically Modeling the Covariance

In this section, we extend the static model (6) to allow for time-varying covariance in non-stationary processes. To this end, we consider the following dynamic model:

$$\begin{aligned} \mathbf{y}_t \mid \boldsymbol{\Sigma}_t(\boldsymbol{\sigma}_t, \mathbf{L}_t) &\sim \ell(\mathbf{y}_t; \boldsymbol{\Sigma}_t(\boldsymbol{\sigma}_t, \mathbf{L}_t)), \quad \boldsymbol{\Sigma}_t(\boldsymbol{\sigma}_t, \mathbf{L}_t) = \text{diag}(\boldsymbol{\sigma}_t) \mathbf{P}_t \,\text{diag}(\boldsymbol{\sigma}_t),\ \mathbf{P}_t = \mathbf{L}_t \mathbf{L}_t^{\mathsf{T}} \\ \boldsymbol{\sigma}_t &\sim p(\boldsymbol{\sigma}_t) \\ \mathbf{L}_t &\sim p(\mathbf{L}_t),\ \text{vech}^{\mathsf{T}}(\mathbf{L}_t) \in \prod_{i=1}^{D} \mathcal{S}_0^{i-1} \end{aligned} \tag{11}$$

One can model the components of $\boldsymbol{\sigma}_t$ as independent dynamic processes using, e.g. an ARMA, GARCH, or log-Gaussian process specification. For $\mathbf{L}_t$, we use vector processes. Since each row of $\mathbf{L}_t$ has to be on a sphere of certain dimension, we require the unit norm constraint for the dynamic process over time. We refer to any multivariate process satisfying the following requirement as *unit-vector process (uvP).*

**Definition 4 (Unit-vector process)** *A vector stochastic process $\mathbf{l}_i(x)$ is called* unit-vector process, *if*

$$\|\mathbf{l}_i(x)\| \equiv 1, \quad \forall x \in \mathcal{X}$$

*where $\|\cdot\|$ could be any vector norm, e.g. 2-norm.*

Note, the norm constraint is imposed in the strong sense, not under the expectation. A unit-vector process can be obtained by normalizing an existing multivariate process, e.g. the *vector Gaussian process (vGP)*, as defined below.

**Definition 5 (Vector Gaussian process)** *A $D$-dimensional* vector Gaussian process $\mathbf{Z}(x) := (Z_1(x), \cdots, Z_D(x))$, *with vector mean function $\boldsymbol{\mu}(x) = (\mu_1(x), \cdots, \mu_D(x))$, covariance function $C$ and ($D$-dimensional) cross covariance $\mathbf{V}_{D\times D}$,*

$$\mathbf{Z}(x) \sim \mathcal{GP}_D(\boldsymbol{\mu}, C, \mathbf{V}_{D\times D})$$

*is a collection of $D$-dimensional random vectors, indexed by $x \in \mathcal{X}$, such that for any finite set of indices $\{x_1, \cdots, x_N\}$, the random matrix $\widetilde{\mathbf{Z}}_{N\times D} := (\mathbf{Z}(x_1), \cdots, \mathbf{Z}(x_N))^{\mathsf{T}}$ has the following matrix normal distribution*

$$\widetilde{\mathbf{Z}}_{N\times D} \sim \mathcal{MN}_{N\times D}(\mathbf{M}_{N\times D}, \mathbf{K}_{N\times N}, \mathbf{V}_{D\times D})$$

*where $\mathbf{M}_{N\times D} := (\mathbf{m}_1, \cdots, \mathbf{m}_D)$, and $\mathbf{m}_k = (\mu_k(x_1), \cdots, \mu_k(x_N))^{\mathsf{T}}$, and $\mathbf{K}$ is the kernel matrix with elements $K_{ij} = C(x_i, x_j)$.*

**Remark 6** *Note for each $k = 1, \cdots D$, we have the following marginal GP*

$$Z_k(x) \sim \mathcal{GP}(\boldsymbol{\mu}_k, C)$$

*In the above definition, we require a common kernel $C$ for all the marginal GPs, whose dependence is characterized by the cross covariance $\mathbf{V}_{D\times D}$. On the other hand, for any fixed $x^* \in \mathcal{X}$, we have*

$$\mathbf{Z}(x^*) \sim \mathcal{N}_D(\boldsymbol{\mu}(x^*), \mathbf{V}_{D\times D})$$

For simplicity, we often consider $\mu \equiv 0$ and $\mathbf{V}_{D\times D} = \mathbf{I}_D$. That is, $Z_k(x) \overset{iid}{\sim} \mathcal{GP}(0, C)$ for $k = 1, \cdots, D$. Normalizing vGP $\mathbf{Z}(\cdot)$ yields a *unit-vector Gaussian process (uvGP)* $\mathbf{Z}^*(\cdot) := \mathbf{Z}(\cdot)/\|\mathbf{Z}(\cdot)\|$, denoted as $\mathbf{Z}^*(\cdot) \sim \mathcal{GP}_D^{\mathcal{S}}(\boldsymbol{\mu}, C, \mathbf{V})$. Note for any fixed $x^* \in \mathcal{X}$, $\mathbf{Z}^*(x^*) \sim \text{n-}\mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$. Setting $\mu \equiv 0$, $\mathbf{V} = \mathbf{I}$ and conditioned on the length $\ell_n$ of each row of $\widetilde{\mathbf{Z}}$, we have

$$p(\widetilde{\mathbf{Z}}^* \mid \{\|z_{n\cdot}\| = \ell_n\}) = p_{\mathcal{MN}}(\widetilde{\mathbf{Z}}) \left|\frac{d\text{vec}(\widetilde{\mathbf{Z}})}{d\text{vec}(\widetilde{\mathbf{Z}}^*)}\right| = \frac{\prod_{n=1}^N \ell_n^D}{(2\pi)^{\frac{ND}{2}} |\mathbf{K}|^{\frac{D}{2}}} \exp\left\{-\frac{1}{2}\text{tr}\left[(\widetilde{\mathbf{Z}}^*)^{\mathsf{T}} \text{diag}(\{\ell_n\})\mathbf{K}^{-1} \text{diag}(\{\ell_n\})\widetilde{\mathbf{Z}}^*\right]\right\}$$

For each marginal GP, we select the following exponential function as the common kernel

$$C(x, x') = \gamma \exp(-0.5\|x - x'\|^s/\rho^s)$$

where $s$ controls the smoothness, the scale parameter $\gamma$ is given an inverse-Gamma prior, and the correlation length parameter $\rho$ is given a log-normal prior. Figure 2 shows a realization of vector GP $\mathbf{Z}_t$, its normalization (forming rows of) $\mathbf{L}_t$ and the induced correlation process $\mathbf{P}_t$ respectively under the model (11).

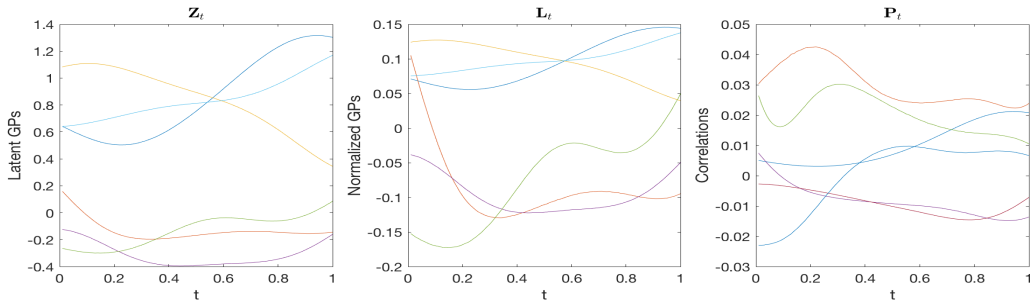


**FIGURE 2** A realization of vector GP $\mathbf{Z}_t$ (left), its normalization (forming rows of) $\mathbf{L}_t$ (middle) and the induced correlation process $\mathbf{P}_t$ (right).

In what follows, we focus on multivariate time series; therefore, we use the one dimensional time index $t \in \mathcal{X} = \mathbb{R}^+$. The overall structured dynamic model can be summarized as follows:

$$\begin{aligned}
\mathbf{y}_t &\sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), & \boldsymbol{\Sigma}_t &= \mathrm{diag}(\boldsymbol{\sigma}_t)\mathbf{L}_t\mathbf{L}_t^{\mathsf{T}}\,\mathrm{diag}(\boldsymbol{\sigma}_t) \\
\boldsymbol{\mu}_t &\sim \mathcal{GP}_D(0, C_\mu, \mathbf{I}), & C_\mu(t,t') &= \gamma_\mu \exp(-0.5\|t-t'\|^s/\rho_\mu^s) \\
\log\boldsymbol{\sigma}_t &\sim \mathcal{GP}_D(0, C_\sigma, \mathbf{I}), & C_\sigma(t,t') &= \gamma_\sigma \exp(-0.5\|t-t'\|^s/\rho_\sigma^s) \\
\mathbf{l}_i(t) &\sim \mathcal{GP}_i^{\mathcal{S}}(\mathbf{n}_i, C_L, \mathbf{I}), & C_L(t,t') &= \gamma_L \exp(-0.5\|t-t'\|^s/\rho_L^s) \\
\gamma_* &\sim \Gamma^{-1}(a_*, b_*), & \log\rho_* &\sim \mathcal{N}(m_*, V_*),\ *=\mu,\sigma,\text{ or } L
\end{aligned} \tag{12}$$

where a constant mean function $\mathbf{n}_i$ is used in the uvGP prior for $\mathbf{l}_i(t)$, with mean matrix $\mathbf{M} = \mathbf{1}_N \otimes \mathbf{n}_i^{\mathsf{T}}$ for the realization $\tilde{\mathbf{l}}_i$. With this structure, one can naturally model the evolution of variances and correlations separately in order to obtain more flexibility. If the focus is on modeling the correlation among multiple time series, then one can substitute $\boldsymbol{\sigma}_t$ with a point estimate $\widehat{\boldsymbol{\sigma}}$ from one trial and assume a steady variance vector. Alternatively, if sufficient trials are present, one can obtain an empirical estimate, $\widehat{\boldsymbol{\sigma}}_t$, from multiple trials at each time point. In the next section, we will discuss a computationally efficient sampling algorithm to obtain the resulting posterior distribution on the covariance (or correlation) matrices given the observed data.

## 3 | POSTERIOR INFERENCE

We now focus on obtaining the posterior probability of $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t, \mathbf{L}_t, \boldsymbol{\gamma} := (\gamma_\mu, \gamma_\sigma, \gamma_L)$ and $\boldsymbol{\rho} := (\rho_\mu, \rho_\sigma, \rho_L)$ in the model (12). Denote the realization of processes $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t, \mathbf{L}_t$ at discrete time points $\{t_n\}_{n=1}^N$ as $\widetilde{\boldsymbol{\mu}}_{N\times D}, \widetilde{\boldsymbol{\sigma}}_{N\times D}, \widetilde{\mathbf{L}}_{N\times D\times D}$ respectively. Transform the parameters $\widetilde{\boldsymbol{\tau}} := \log(\widetilde{\boldsymbol{\sigma}}), \boldsymbol{\eta} := \log(\boldsymbol{\rho})$ for the convenience of calculation. Denote $\widetilde{\mathbf{Y}}_{M\times N\times D} := \{\mathbf{Y}_1, \cdots, \mathbf{Y}_M\}$ for $M$ trials, $(\mathbf{Y}_m)_{N\times D} := [\mathbf{y}_{m1}, \cdots, \mathbf{y}_{mN}]^{\mathsf{T}}$ and $\mathbf{y}^*_{mn} := (\mathbf{y}_{mn} - \boldsymbol{\mu}_n) \circ e^{-\boldsymbol{\tau}_n}$ where $\circ$ is the Hadamard product (a.k.a. Schur product), i.e. the entry-wise product. Let $\mathbf{K}_*(\gamma_*, \eta_*) = \gamma_* \mathbf{K}_{0*}(\eta_*)$ and $\tilde{\mathbf{l}}_i^* := \tilde{\mathbf{l}}_i - \mathbf{1}_N \otimes \mathbf{n}_i^{\mathsf{T}}$.

### 3.1 | Metropolis-within-Gibbs

We use a Metropolis-within-Gibbs algorithm and alternate updating the model parameters $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\tau}}, \widetilde{\mathbf{L}}, \boldsymbol{\gamma}, \boldsymbol{\eta}$. We now list the parameters and their respective updates one by one.

($\gamma$). Note the prior for $\gamma$ is conditionally conjugate given $* = \mu, \tau, \text{ or } L$,

$$\gamma_* | \tilde{*}, \eta_* \sim \Gamma^{-1}(a'_*, b'_*), \quad a'_* = a_* + \frac{ND}{2}((D+1)/2 - D^{-1})^{[*=L]}, \quad b'_* = b_* + \frac{1}{2}\text{tr}(\tilde{*}^\mathsf{T}\mathbf{K}_{0*}(\eta_*)^{-1}\tilde{*})$$

where [condition] is 1 with the condition satisfied and 0 otherwise.

($\boldsymbol{\eta}$). Given $* = \mu, \tau, \text{ or } L$, we could sample $\eta_*$ using the slice sampler (Neal, 2003), which only requires log-posterior density and works well for scalar parameters,

$$\log p(\eta_* | \tilde{*}, \gamma_*) = -\frac{D((D+1)/2 - D^{-1})^{[*=L]}}{2} \log |\mathbf{K}_{0*}(\eta_*)| - \frac{1}{2}\text{tr}(\tilde{*}^\mathsf{T}\mathbf{K}_{0*}(\eta_*)^{-1}\tilde{*})/\gamma_* - \frac{1}{2}(\eta_* - m_*)^2/V_*$$

($\widetilde{\boldsymbol{\mu}}$). By the definition of vGP, we have $\widetilde{\boldsymbol{\mu}} | \gamma_\mu, \eta_\mu \sim \mathcal{MN}_{N\times D}(0, \mathbf{K}_\mu, \mathbf{I}_D)$; therefore, $\text{vec}(\widetilde{\boldsymbol{\mu}}) | \gamma_\mu, \eta_\mu \sim \mathcal{N}_{ND}(0, \mathbf{I}_D \otimes \mathbf{K}_\mu)$. On the other hand, one can write

$$\begin{aligned}\sum_{m=1}^{M}\sum_{n=1}^{N} {\mathbf{y}^*_{mn}}^\mathsf{T}\mathbf{P}_n^{-1}\mathbf{y}^*_{mn} &= \sum_{m=1}^{M} \text{vec}((\mathbf{Y}_m - \widetilde{\boldsymbol{\mu}})^\mathsf{T})^\mathsf{T} \,\text{diag}(\{\widetilde{\boldsymbol{\Sigma}}_n^{-1}\})\text{vec}((\mathbf{Y}_m - \widetilde{\boldsymbol{\mu}})^\mathsf{T}) \\ &= \sum_{m=1}^{M} (\text{vec}(\mathbf{Y}_m) - \text{vec}(\widetilde{\boldsymbol{\mu}}))^\mathsf{T} \widetilde{\boldsymbol{\Sigma}}_K^{-1} (\text{vec}(\mathbf{Y}_m) - \text{vec}(\widetilde{\boldsymbol{\mu}}))\end{aligned}$$

where $\widetilde{\boldsymbol{\Sigma}}_K^{-1} := K_{(D,N)}\,\text{diag}(\{\widetilde{\boldsymbol{\Sigma}}_n\})^{-1}K_{(N,D)}$, and $K_{(N,D)}$ is the *commutation matrix* of size $ND \times ND$ such that for any $N \times D$ matrix $\mathbf{A}$, $K_{(N,D)}\text{vec}(\mathbf{A}) = \text{vec}(\mathbf{A}^\mathsf{T})$ (Tracy and Dwyer, 1969; Magnus and Neudecker, 1979). Therefore, the prior on $\text{vec}(\widetilde{\boldsymbol{\mu}})$ is conditionally conjugate, and we have

$$\text{vec}(\widetilde{\boldsymbol{\mu}}) | \widetilde{\mathbf{Y}}, \widetilde{\boldsymbol{\Sigma}}, \gamma_\mu, \eta_\mu \sim \mathcal{N}_{ND}(\boldsymbol{\mu}', \boldsymbol{\Sigma}'), \quad \boldsymbol{\mu}' = \boldsymbol{\Sigma}'\widetilde{\boldsymbol{\Sigma}}_K^{-1}\sum_{m=1}^{M}\text{vec}(\mathbf{Y}_m), \quad \boldsymbol{\Sigma}' = \left(\mathbf{I}_D \otimes \mathbf{K}_\mu^{-1} + M\widetilde{\boldsymbol{\Sigma}}_K^{-1}\right)^{-1}$$

($\widetilde{\boldsymbol{\tau}}$). Using a similar argument by matrix Normal prior for $\widetilde{\boldsymbol{\tau}}$, we have $\text{vec}(\widetilde{\boldsymbol{\tau}}) | \gamma_\tau, \eta_\tau \sim \mathcal{N}_{ND}(0, \mathbf{I}_D \otimes \mathbf{K}_\tau)$. Therefore, we could use the elliptic slice sampler (ESS, Murray et al., 2010), which only requires the log-likelihood

$$\log p(\widetilde{\boldsymbol{\tau}}; \widetilde{\mathbf{Y}}, \widetilde{\boldsymbol{\mu}}) = -\sum_{n=1}^{N}\left[M\mathbf{1}_D^\mathsf{T}\boldsymbol{\tau}_n + \sum_{m=1}^{M}\frac{1}{2}{\mathbf{y}^*_{mn}}^\mathsf{T}\mathbf{P}_n^{-1}\mathbf{y}^*_{mn}\right] = -M\mathbf{1}_{ND}^\mathsf{T}\text{vec}(\widetilde{\boldsymbol{\tau}}) - \sum_{m=1}^{M}\frac{1}{2}\text{vec}(\mathbf{Y}^*_m)^\mathsf{T}\widetilde{\mathbf{P}}_K^{-1}\text{vec}(\mathbf{Y}^*_m)$$

where $\widetilde{\mathbf{P}}_K^{-1} := K_{(D,N)}\,\text{diag}(\{\widetilde{\mathbf{P}}_n\})^{-1}K_{(N,D)}$ and $\mathbf{Y}^*_m := (\mathbf{Y}_m - \widetilde{\boldsymbol{\mu}}) \circ \exp(-\widetilde{\boldsymbol{\tau}})$.

($\widetilde{\mathbf{L}}$). For each $n \in \{1, \cdots N\}$, we have $\text{vech}^\mathsf{T}(\mathbf{L}_n) \in \prod_{i=1}^{D}\mathcal{S}_0^{i-1}$. We could sample from its posterior distribution using the Δ-*Spherical Hamiltonian Monte Carlo (*Δ*-SphHMC)* described below. The log-posterior density of $\widetilde{\mathbf{L}}$ is

$$\log p(\widetilde{\mathbf{L}} | \widetilde{\mathbf{Y}}, \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\tau}}, \gamma_L, \eta_L) = -\sum_{n=1}^{N}\left[M\log|\mathbf{L}_n| + \sum_{m=1}^{M}\frac{1}{2}{\mathbf{y}^*_{mn}}^\mathsf{T}\mathbf{P}_n^{-1}\mathbf{y}^*_{mn}\right] - \frac{1}{2}\sum_{i=2}^{D}\text{tr}(\tilde{\mathbf{l}}_i^{*\mathsf{T}}\mathbf{K}_L^{-1}\tilde{\mathbf{l}}_i^*)$$

The derivative of log-likelihood with respect to $\mathbf{L}_n$ and the derivative of log-prior with respect to $\tilde{\mathbf{l}}_i$ can be calculated as

$$\frac{\partial}{\partial \mathbf{L}_n}\log p(\widetilde{\mathbf{L}}; \widetilde{\mathbf{Y}}, \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\tau}}) = -M\frac{\mathbf{I}_D}{\mathbf{L}_n} + \sum_{m=1}^{M}\text{tril}(\mathbf{P}_n^{-1}\mathbf{y}^*_{mn}{\mathbf{y}^*_{mn}}^\mathsf{T}\mathbf{L}_n^{-\mathsf{T}}), \quad \frac{\partial}{\partial \tilde{\mathbf{l}}_i}\log p(\widetilde{\mathbf{L}} | \gamma_L, \eta_L) = -\mathbf{K}_L^{-1}\tilde{\mathbf{l}}_i^*$$

## 3.2 | Spherical HMC

We need an efficient algorithm to handle the intractability in the posterior distribution of $\widetilde{\mathbf{L}}$ introduced by various flexible priors. *Spherical Hamiltonian Monte Carlo* (SphHMC, Lan et al., 2014; Lan and Shahbaba, 2016) is a Hamiltonian Monte Carlo (HMC, Duane et al., 1987; Neal, 2011) algorithm on spheres that can be viewed as a special case of geodesic Monte Carlo (Byrne and Girolami, 2013), or manifold Monte Carlo methods (Girolami and Calderhead, 2011; Lan et al., 2015). The algorithm was originally proposed to handle norm constraints in sampling so it is natural to apply it to sample each row of the Cholesky factor of a correlation matrix with unit 2-norm constraint. The general notation $\mathbf{q}$ is instantiated as $\mathbf{l}_i$ in this section.

Assume a probability distribution with density function $f(\mathbf{q})$ is defined on a $(D-1)$ dimensional sphere with radius $r$, denoted as $\mathcal{S}^{D-1}(r) := \{\mathbf{q} \in \mathbb{R}^D | \|\mathbf{q}\|_2 = r\}$. Due to the norm constraint, there are $(D-1)$ free parameters $\mathbf{q}_{-D} := (q_1, \cdots, q_{D-1})$, which can be viewed as the Cartesian coordinates for the manifold $\mathcal{S}_+^{D-1}(r)$ (Lan and Shahbaba, 2016). The last component is determined up to a sign: $q_D = \pm\sqrt{r^2 - \|\mathbf{q}_{-D}\|_2^2}$. To induce Hamiltonian dynamics on the sphere, we define the potential energy for position $\mathbf{q}$ as $U(\mathbf{q}) := -\log f(\mathbf{q})$. Endowing the canonical spherical metric $\mathbf{G}(\mathbf{q}_{-D}) = \mathbf{I}_{D-1} + \frac{\mathbf{q}_{-D}\mathbf{q}_{-D}^\mathsf{T}}{q_D^2}$ on the Riemannian manifold $\mathcal{S}^{D-1}(r)$, we introduce the auxiliary velocity vector $\mathbf{v}|\mathbf{q} \sim \mathcal{N}(0, \mathbf{G}(\mathbf{q})^{-1})$ and define the associated kinetic energy as $K(\mathbf{v};\mathbf{q}) := -\log f_{\mathcal{N}}(\mathbf{v}|\mathbf{q}) = -\frac{1}{2}\log|\mathbf{G}(\mathbf{q}_{-D})| + \frac{1}{2}\mathbf{v}_{-D}^\mathsf{T}\mathbf{G}(\mathbf{q}_{-D})\mathbf{v}_{-D}$ (Girolami and Calderhead, 2011). Therefore the total energy is defined as

$$E(\mathbf{q},\mathbf{v}) := U(\mathbf{q}) + K(\mathbf{v};\mathbf{q}) = \tilde{U}(\mathbf{q}) + K_0(\mathbf{v};\mathbf{q}) \tag{13}$$

where we denote $\tilde{U}(\mathbf{q}) := U(\mathbf{q}) - \frac{1}{2}\log|\mathbf{G}(\mathbf{q}_{-D})| = -\log f(\mathbf{q}) + \log|q_D|$, and $K_0(\mathbf{v};\mathbf{q}) := \frac{1}{2}\mathbf{v}_{-D}^\mathsf{T}\mathbf{G}(\mathbf{q}_{-D})\mathbf{v}_{-D} = \frac{1}{2}\mathbf{v}^\mathsf{T}\mathbf{v}$ (Lan and Shahbaba, 2016). Therefore the Lagrangian dynamics with above total energy (13) is (Lan et al., 2015)

$$\begin{aligned}\dot{\mathbf{q}}_{D-D} &= \mathbf{v}_{-D}\\ \dot{\mathbf{v}}_{-D} &= -\mathbf{v}_{-D}^\mathsf{T}\boldsymbol{\Gamma}(\mathbf{q}_{-D})\mathbf{v}_{-D} - \mathbf{G}(\mathbf{q}_{-D})^{-1}\nabla_{\mathbf{q}_{-D}}\tilde{U}(\mathbf{q})\end{aligned} \tag{14}$$

where $\boldsymbol{\Gamma}(\mathbf{q}_{-D}) = r^{-2}\mathbf{G}(\mathbf{q}_{-D}) \otimes \mathbf{q}_{-D}$ is the Christoffel symbols of second kind (see details in Lan and Shahbaba, 2016, for $r = 1$). A splitting technique is applied to yield the following geometric integrator (Lan et al., 2014; Lan and Shahbaba, 2016), which also includes the last coordinates $q_D, v_D$:

$$\begin{aligned}\mathbf{v}^- &= \mathbf{v} - \frac{h}{2}\mathcal{P}(\mathbf{q})\mathbf{g}(\mathbf{q})\\ \begin{bmatrix}\mathbf{q}'\\ \mathbf{v}^+\end{bmatrix} &= \begin{bmatrix}r & 0\\ 0 & \|\mathbf{v}^-\|_2\end{bmatrix}\begin{bmatrix}\cos(\|\mathbf{v}^-\|_2 r^{-1}h) + \sin(\|\mathbf{v}^-\|_2 r^{-1}h)\\ -\sin(\|\mathbf{v}^-\|_2 r^{-1}h) + \cos(\|\mathbf{v}^-\|_2 r^{-1}h)\end{bmatrix}\begin{bmatrix}r^{-1} & 0\\ 0 & \|\mathbf{v}^-\|_2^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{q}\\ \mathbf{v}^-\end{bmatrix}\\ \mathbf{v}' &= \mathbf{v}^+ - \frac{h}{2}\mathcal{P}(\mathbf{q}')\mathbf{g}(\mathbf{q}')\end{aligned} \tag{15}$$

where $\mathbf{g}(\mathbf{q}) := \nabla_{\mathbf{q}}\tilde{U}(\mathbf{q})$, $\mathcal{P}(\mathbf{q}) := \mathbf{I}_D - r^{-2}\mathbf{q}\mathbf{q}^\mathsf{T}$. (15) defines a mapping $\mathcal{T}_h : (\mathbf{q},\mathbf{v}) \mapsto (\mathbf{q}',\mathbf{v}')$. Denote $\|\mathbf{u}\|_{\mathcal{P}(\mathbf{q})}^2 := \mathbf{u}^\mathsf{T}\mathcal{P}(\mathbf{q})\mathbf{u}$. After applying such integrator $T$ times, a proposal $(\mathbf{q}_T,\mathbf{v}_T) = \mathcal{T}_h^T(\mathbf{q}_0,\mathbf{v}_0)$ is accepted with the following probability

$$\begin{aligned}a_{sphHMC} &= 1 \wedge \exp(-\Delta E)\\ \Delta E &= \tilde{U}(\mathbf{q}_T) - \tilde{U}(\mathbf{q}_0) - \frac{h^2}{8}\left[\|\mathbf{g}(\mathbf{q}_T)\|_{\mathcal{P}(\mathbf{q})}^2 - \|\mathbf{g}(\mathbf{q}_0)\|_{\mathcal{P}(\mathbf{q})}^2\right] - \frac{h}{2}\left[\langle\mathbf{v}_0,\mathbf{g}(\mathbf{q}_0)\rangle + \langle\mathbf{v}_T,\mathbf{g}(\mathbf{q}_T)\rangle\right] - h\sum_{\tau=1}^{T-1}\langle\mathbf{v}_\tau,\mathbf{g}(\mathbf{q}_\tau)\rangle\end{aligned} \tag{16}$$

We can prove the following limiting result (Beskos et al., 2011).

**Theorem 3** *Let $h \to 0$ we have the following energy conservation*

$$E(\mathbf{q}(T), \mathbf{v}(T)) - E(\mathbf{q}(0), \mathbf{v}(0)) = \tilde{U}(\mathbf{q}(T)) - \tilde{U}(\mathbf{q}(0)) - \int_0^T \langle \mathbf{v}(t), \mathbf{g}(\mathbf{q}(t)) \rangle dt = 0$$

**Proof** See Appendix B.

## 3.3 | Adaptive Spherical HMC

There are two tuning parameters in HMC and its variants: the step size $h$ and the number of integration (leapfrog) steps $T$. Hand tuning heavily relies on domain expertise and could be inefficient. Here, we adopt the 'No-U-Turn' idea from Hoffman and Gelman (2014) and introduce a novel adaptive algorithm that obviates manual tuning of these parameters.

First, for any given step size $h$, we adopt a rule for setting the number of leapfrog steps based on the same philosophy as 'No-U-Turn' (Hoffman and Gelman, 2014). The idea is to avoid waste of computation occurred (e.g. when the sampler backtracks on its trajectory) without breaking the detailed balance condition for the MCMC transition kernel. $\mathcal{S}^{D-1}(r)$ is a compact manifold where any two points $\mathbf{q}(0), \mathbf{q}(t) \in \mathcal{S}^{D-1}(r)$ have bounded geodesic distance $\pi r$. We adopt the stopping rule for the leapfrog when the sampler exits the orthant of the initial state, that is, the trajectory measured in geodesic distance is at least $\frac{\pi}{2} r$, which is equivalent to $\langle \mathbf{q}(0), \mathbf{q}(t) \rangle < 0$. On the other hand, this condition may not be satisfied within reasonable number of iterations because the geometric integrator (15) does not follow a geodesic (great circle) in general (only the middle part does), therefore we set some threshold $T_{\max}$ for the number of tests, and adopt the following 'Two-Orthants' (as the starting and end points occupy two orthants) rule for the number of leapfrogs:

$$T_{2orth} = \min_{\tau \in \{0, \cdots, T_{\max}\}} \{\tau : \langle \mathbf{q}_0, \mathbf{q}_\tau \rangle < 0\} \tag{17}$$

Alternatively, one can stop the leapfrog steps in a *stochastic* way based on the geodesic distance travelled:

$$T_{stoch} = \min_{\tau} \{\tau : Z_\tau = 0\}, \quad Z_\tau \sim \text{Bern}(p_\tau), \quad p_\tau = \frac{r^{-2} \langle \mathbf{q}_0, \mathbf{q}_\tau \rangle + 1}{2} \tag{18}$$

These stopping criteria are already time reversible, so the recursive binary tree as in 'No-U-Turn' algorithm (Hoffman and Gelman, 2014) is no longer needed.

Lastly, we adopt the *dual averaging* scheme (Nesterov, 2009) for the adaptation of step size $h$. See Hoffman and Gelman (2014) for more details. We summarize our *Adaptive Spherical Hamiltonian Monte Carlo (adp-SphHMC)* in Algorithm 1.

To sample $\mathbf{L}$ (or $\mathbf{L}_t$), we could use a *multi-Spherical HMC* for the product of spheres, e.g. $\prod_{i=1}^{D} \mathcal{S}_0^{i-1}$, can run in parallel. More specifically, on each factor sphere $\mathcal{S}_0^{i-1}$, we update the parameter vector $\mathbf{l}_i$ according to (15). However, we calculate the acceptance probability (16) based on the sum of total energy of all components; that is, updating and accepting $\text{vech}^{\mathsf{T}}(\mathbf{L})$ (or $\text{vech}^{\mathsf{T}}(\mathbf{L}_t)$) all together. For this specific application of sampling Cholesky factor of correlation (covariance) matrix, we refer to the resulting algorithm as Δ-*Spherical HMC (Δ-SphHMC)*.

The computational complexity of the GP prior is $O(N^3)$, and that of the likelihood evaluation is $O(MD^2)$. Finally MCMC updates of $\widetilde{\boldsymbol{\mu}}_{N\times D}, \widetilde{\boldsymbol{\sigma}}_{N\times D}, \widetilde{\mathbf{L}}_{N\times D\times D}$ have complexity $O(ND)$, $O(ND)$ and $O(ND^2)$, respectively. To scale up applications to larger dimension $D$, one may use prior knowledge to group the data. In this way, one does not have to maintain the full covariance/correlation matrices/processes. Subgroups correspond to blocks in covariance matrices which may

**Algorithm 1** Adaptive Spherical HMC (adp-SphHMC)

Given $\mathbf{q}_0$, $a_0$, $N$, $N^{adapt}$.

Set $h_0 = 1$ or using Algorithm 4 of Hoffman and Gelman (2014), $\mu = \log(10h_0)$, $\overline{h}_0 = 1$, $\overline{A}_0 = 0$, $\gamma = 0.05$, $n_0 = 10$, $\kappa = 0.75$.

**for** $n = 1$ to $N$ **do**

  Sample a new velocity $\mathbf{v}_{n-1} \sim \mathcal{N}(0, \mathbf{I}_D)$, and set $\mathbf{v}_{n-1} = \mathcal{P}(\mathbf{q}_{n-1})\mathbf{v}_{n-1}$.

  Set $\mathbf{q}^{(0)} = \mathbf{q}_{n-1}$, $\mathbf{v}^{(0)} = \mathbf{v}_{n-1}$.

  **for** $\tau = 0$ to $T-1$ ($T = T_{2orth}$ or $T_{stoch}$) **do**

    Run leapfrog step (15) to update $(\mathbf{q}^{(\tau+1)}, \mathbf{v}^{(\tau+1)}) \leftarrow \mathcal{T}_{h_{n-1}}(\mathbf{q}^{(\tau)}, \mathbf{v}^{(\tau)})$.

    **if** Stopping criterion (17) (or (18)) is satisfied **then**

      Break

    **end if**

  **end for**

  Accept the proposal $(\mathbf{q}^{(T)}, \mathbf{v}^{(T)})$ with probability $a_n^{sphHMC}$ in (16) and set $\mathbf{q}_n = \mathbf{q}^{(T)}$; otherwise set $\mathbf{q}_n = \mathbf{q}_{n-1}$.

  **if** $n \leq N^{adapt}$ **then**

    Set $\overline{A}_n = \left(1 - \frac{1}{n+n_0}\right)\overline{A}_{n-1} + \frac{1}{n+n_0}(a_0 - a_n)$.

    Set $\log h_n = \mu - \frac{\sqrt{n}}{\gamma}\overline{A}_n$, and $\log \overline{h}_n = n^{-\kappa}\log h_n + (1 - n^{-\kappa})\log \overline{h}_{n-1}$.

  **else**

    Set $h_n = \overline{h}_{N^{adpat}}$.

  **end if**

**end for**

be arranged in some 'band' along the diagonal. Assuming no correlation among sub-groups, one can take advantage of such '$w$-band' structure in modeling covariance matrices/processes. More specifically, we can assume $\mathbf{L}_t$ is $w$-band lower triangular matrix for each time $t$, i.e. if $l_{ij} = 0$ for $i < j$ or $i - j \geq w$, then the resulting covariance/correlation will be $(2w-1)$-banded. In this way the complexity of likelihood evaluation and updating $\widetilde{\mathbf{L}}$ will be reduced to $O(MwD)$ and $O(NwD)$ resepctively. Therefore the computational cost would scale linearly with the dimension $D$. This technique will be investigated in Section 4.2.3.

## 4 | NUMERICAL EXPERIMENTS

In this section, we use several examples to illustrate our structured modeling for covariance and application of Δ-SphHMC. As a first example, we consider the Normal-inverse-Wishart problem. Since there is conjugacy and we know the true posterior, we use this to verify our method and study the flexibility of priors in Section 2.2. Then we test our dynamical modeling method in Section 2.3 on a simulated periodic model. This simulation also demonstrates the advantage of our model in full flexibility compared to a state-of-the-art nonparametric covariance regression model based on latent factor process (Fox and Dunson, 2015). Finally, we apply our methodology to local potential field data collected from a memory sequence experiment to model the evolving pattern of multiple time series.

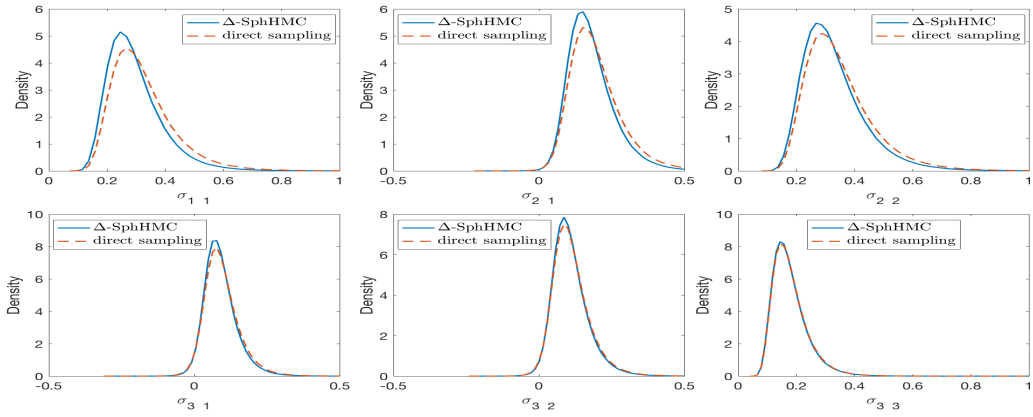


**FIGURE 3** Marginal posterior densities of $\sigma_{ij}$ in the normal-inverse-Wishart problem. Solid blue lines are estimates by Δ-SphHMC and dashed red lines are estimates by direct sampling. All densities are estimated with $10^6$ samples.

## 4.1 | Normal-inverse-Wishart Problem

Now we consider the following example involving inverse-Wishart prior

$$\begin{aligned} \mathbf{y}_n|\boldsymbol{\Sigma} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}), \quad n = 1, \cdots, N \\ \boldsymbol{\Sigma} &\sim \mathcal{W}_D^{-1}(\boldsymbol{\Psi}, \nu) \end{aligned} \tag{19}$$

It is known that the posterior of $\boldsymbol{\Sigma}|\mathbf{Y}$ is still inverse-Wishart distribution:

$$\boldsymbol{\Sigma}|\mathbf{Y} \sim \mathcal{W}_D^{-1}(\boldsymbol{\Psi} + (\mathbf{Y} - \boldsymbol{\mu}_0)(\mathbf{Y} - \boldsymbol{\mu}_0)^{\mathsf{T}}, \nu + N), \qquad \mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]^{\mathsf{T}} \tag{20}$$

In all examples of this section, we consider dimension $D = 3$ and generate data $\mathbf{Y}$ with the following setting

$$\boldsymbol{\mu}_0 = 0, \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \frac{1}{11}(\mathbf{I} + \mathbf{1}\mathbf{1}^{\mathsf{T}})$$

where $\rho_{ij} \equiv \rho = 0.50$ for $i \neq j$. We generate $N = 20$ data points so that the prior is not overwhelmed by data.

### 4.1.1 | Verification of Validity

Specifying conditional priors based on (8) in the structured model (6), we want to check the validity of our proposed method by comparing the posterior estimates using Δ-SphHMC agains the truth (20).

We sample $\boldsymbol{\tau} := \log(\boldsymbol{\sigma})$ using standard HMC and $\mathbf{U}^*$ using Δ-SphHMC. They are updated in Metropolis-Within-Gibbs scheme. $10^6$ samples are collected after burning the first 10% and subsampling every 1 of 10. For each sample of $\boldsymbol{\tau}$ and $\text{vech}(\mathbf{U}^*)$, we calculate $\boldsymbol{\Sigma} = \text{diag}(e^{\boldsymbol{\tau}})\mathbf{U}^*(\mathbf{U}^*)^{\mathsf{T}}\,\text{diag}(e^{\boldsymbol{\tau}})$. Marginal densities of entries in $\boldsymbol{\Sigma}$ are estimated with these samples and plotted against the results by direct sampling in Figure 3. Despite of sampling variance, these estimates closely match the results by direct sampling, indicating the validity of our proposed method.

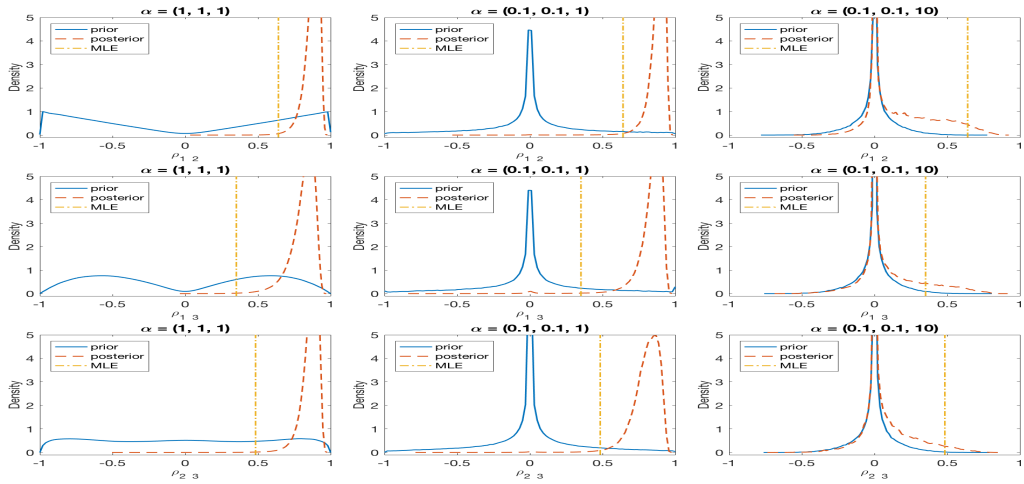


**FIGURE 4** Marginal posterior, prior (induced from squared-Dirichlet distribution) densities of correlations and MLEs with different settings for concentration parameter $\boldsymbol{\alpha}$, estimated with $10^6$ samples.

### 4.1.2 | Examining Flexibility of Priors

We have studied several spherical priors for the Cholesky factor of correlation matrix proposed in Section 2.2. Now we examine the flexibility of these priors in providing prior information for correlation with various parameter settings.

With the same data generated according to (19), we now consider the squared-Dirichlet prior (10) for $\mathbf{L}$ in the structured model (6) with the following setting

$$\begin{aligned} \tau_i = \log(\sigma_i) \sim N(0, 0.1^2), \quad i = 1, \cdots, D \\ \mathbf{l}_i \sim \text{Dir}^2(\boldsymbol{\alpha}_i), \quad \boldsymbol{\alpha}_i = (\alpha \mathbf{1}_{i-1}, \alpha_0), \quad i = 2, \cdots, D \end{aligned} \tag{21}$$

where we consider three cases i) $\alpha = 1,\ \alpha_0 = 1$; ii) $\alpha = 0.1,\ \alpha_0 = 1$; iii) $\alpha = 0.1,\ \alpha_0 = 10$.

We run standard HMC for $\boldsymbol{\tau}$ and $\Delta$-SphHMC for $\mathbf{L}$ to generate $10^6$ samples as in Section 4.1.1. Then we convert the posterior samples of $\mathbf{L}$ to the correlation matrix using $\mathbf{P} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$. We also generate $10^6$ prior samples for $\mathbf{L}$ from (21) and convert them to samples of $\mathbf{P}$. For each entry of $\mathbf{P}$, we estimate the marginal posterior (prior) density based on these posterior (prior) samples. The posteriors, priors and maximal likelihood estimates (MLEs) of correlations $\rho_{ij}$ with different $\boldsymbol{\alpha}$'s are plotted in Figure 4 respectively. In general, the posteriors are compromise between priors and the likelihoods (MLEs). With more and more weight put around the poles (last component) of each factor sphere, the priors become increasingly dominant till the posteriors almost fall on priors when $\boldsymbol{\alpha} = (0.1, 0.1, 10)$. In this extreme case, the squared-Dirichlet distributions induce priors in favor of trivial (zero) correlations. We have similar conclusion on Bingham prior and von Mises-Fisher prior but results are reported in the supplementary file.

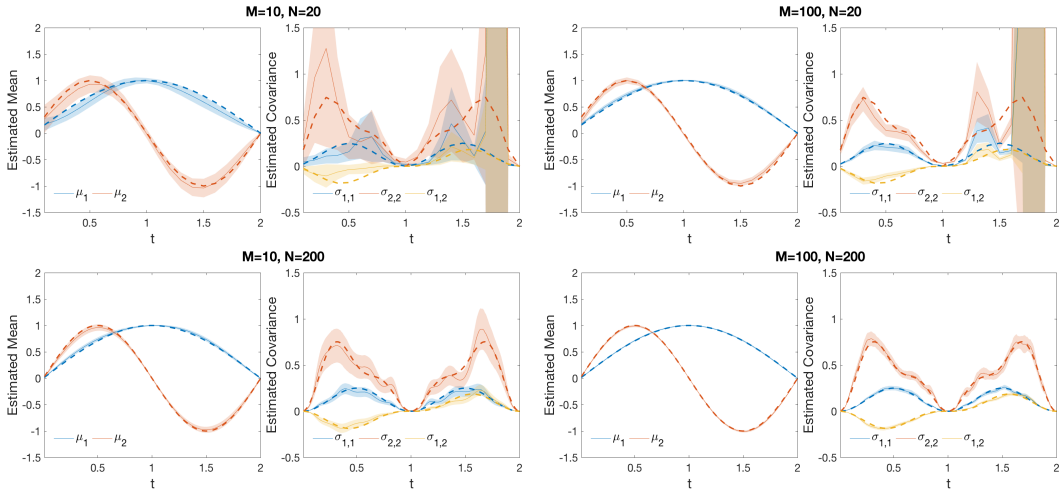


**FIGURE 5** Estimation of the underlying mean functions $\mu_t$ (left in each of 4 subpannels) and covariance functions $\Sigma_t$ (right in each of 4 subpannels) of 2-dimensional periodic processes. $M$ is the number of trials, and $N$ is the number of discretization points. Dashed lines are true values, solid lines are estimates and shaded regions are 95% credible bands.

## 4.2 | Simulated Periodic Processes

In this section, we investigate the performance of our dynamic model (12) on the following simulated example

$$\begin{aligned} &y(t) \sim \mathcal{N}_D(\mu(t), \Sigma(t)), \quad \Sigma(t) = L(t)L(t)^{\mathsf{T}} \circ S, \quad t \in [0,2] \\ &\mu_i(t) = \sin(it\pi/D), \quad L_{ij}(t) = (-1)^i \sin(it\pi/D)(-1)^j \cos(jt\pi/D), \quad j \le i = 1, \cdots, D, \\ &S_{ij} = (|i-j|+1)^{-1}, \quad i,j = 1, \cdots, D \end{aligned} \tag{22}$$

Based on the model (22), we generated $M$ trials (process realizations) of data $y$ at $N$ evenly spaced points for $t$ in $[0,2]$, and therefore the whole data set $\{y(t)\}$ is an $M \times N \times D$ array. We first consider $D = 2$ to investigate the posterior contraction phenomena and the model flexibility; then we consider $D = 100$ over a shorter period $[0,1]$ to show the scalability using the '$w$-band' structure.

### 4.2.1 | Posterior Contraction

Posterior contraction describes the phenomenon that the posterior concentrates on smaller and smaller neighborhood of the true parameter (function) given more and more data (van der Vaart and van Zanten, 2008). We investigate such phenomena in both mean functions and covariance functions in our model (12) using the following settings $i)M = 10, N = 20$; $ii)M = 100, N = 20$; $iii)M = 10, N = 200$; $iv)M = 100, N = 200$.

To fit the data using the model (12), we set $s = 2$, $a = (1,1,1)$, $b = (0.1, 10^{-3}, 0.2)$, $m = (0,0,0)$ for all settings, $V = (1, 0.5, 1)$ for $N = 20$ and $V = (1,1,0.3)$ for $N = 200$. We also add an additional nugget of $10^{-5} I_n$ to all the covariance kernel of GPs to ensure non-degeneracy. Following the procedure in Section 3.1, we run MCMC for $1.5 \times 10^5$ iterations, burn in the first $5 \times 10^4$ and subsample 1 for every 10 to obtain $10^4$ posterior samples in the end. Based on those posterior samples, we estimate the underlying mean functions and covariance functions and plot the estimates in Figure 5.

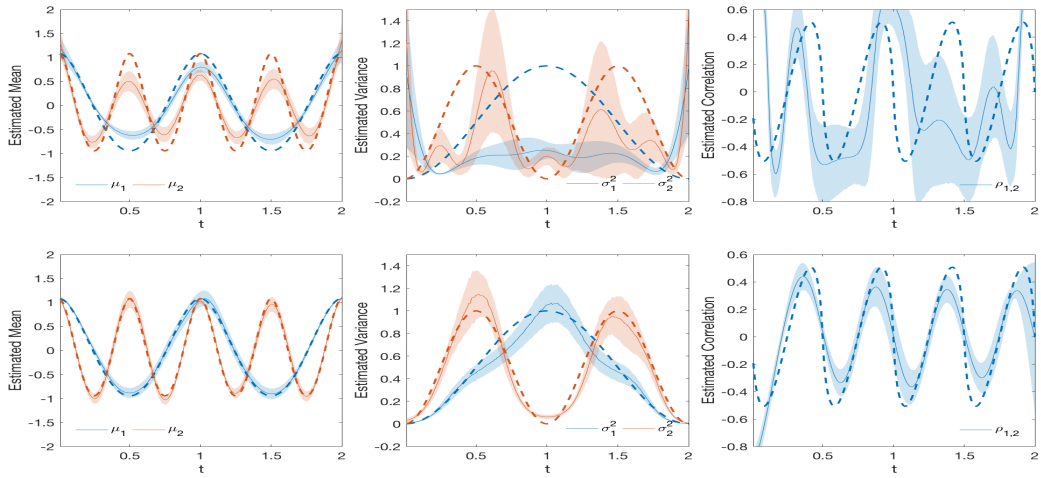


**FIGURE 6** Estimation of the underlying mean functions $\mu_t$ (left column), variance functions $\sigma_t$ (middle column) and correlation function $\rho_t$ (right column) of 2-dimensional periodic processes, using latent factor process model (upper row) and our flexible model (lower row), based on $M = 10$ trials of data over $N = 200$ evenly spaced points. Dashed lines are true values, solid lines are estimates and shaded regions are 95% credible bands.

The first row of Figure 5 shows the estimated mean and covariance functions for $N = 20$ discretization points and the second row shows the results for $N = 200$ points. Both mean and covariance functions have narrower credible bands for more discretization points $N$. On the other hand, the first column shows the estimates with $M = 10$ trials while the second the column shows the results for $M = 100$ trials. The posteriors for both mean and covariance functions contract further with more trials (realizations of the process) $M$. Both $M$ and $N$ have effect on the amount of data information thereafter on the posterior contraction but the contraction rate may depend on them differently. In general the posterior of mean function contracts to the truth faster than the posterior of covariance function. With $M = 100$ trials and $N = 200$ discretization points, both mean and covariance functions are recovered very well by the model (12).

### 4.2.2 | Full Flexibility

Since our method (12) models mean, variance and correlation processes separately, it grants full flexibility in dynamically modeling them when they behave differently. This contrasts latent factor based models for which mean and covariance processes are tied together. One of the state-of-the-art models of this type is Bayesian nonparametric covariance regression (Fox and Dunson, 2015):

$$y(x) \sim \mathcal{N}_D(\mu(x), \Sigma(x)), \qquad \mu(x) = \Lambda(x)\psi(x), \quad \Sigma(x) = \Lambda(x)\Lambda(x)^{\mathsf{T}} + \Sigma_0 \tag{23}$$

We tweak the simulated example (22) for $D = 2$ to let mean and correlation processes have higher frequency than variance processes, as shown in the dashed lines in Figure 6. We generate $M = 10$ trials of data over $N = 200$ evenly spaced points. In this case, the true mean processes $\mu(x)$ and true covariance processes $\Sigma(x)$ behave differently but are modeled with a common loading matrix $\Lambda(x)$ in model (23). This imposes difficulty on (23) to have a latent factor process $\psi(x)$ that could properly accommodate the heterogeneity in mean and covariance processes. Figure 6 shows that due

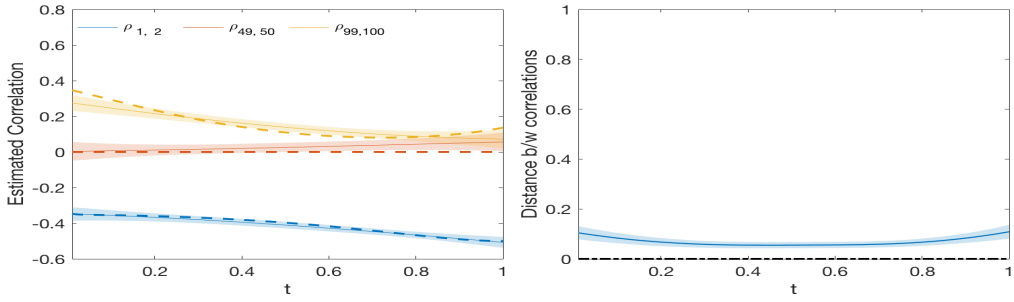


**FIGURE 7** Posterior estimation of the underlying correlation functions $P_t$ (left) and its 2-norm distance to the truth (right) of 100-dimensional periodic processes with 2-band structure, based on $M = 100$ trials of data over $N = 100$ discretization points. Dashed lines are true values, solid lines are estimates and shaded regions are 95% credible bands.

to the different behavior of mean, variance and correlation processes, latent factor based model (23) (upper row) fails to generate satisfactory fit for all of them. Our fully flexible model (12) (bottom row), on the contrary, successfully produces more accurate characterization of all these processes. Note that this artificial example is used to demonstrate the flexibility of our dynamic model (12). For cases that are not as extreme, (23) may performance equally well. See more discussion in Section 5 and more details in the supplementary file.

### 4.2.3 | Scalability

Now we use the same simulation model (22) for $D = 100$ dimensions to test the scalability of our dynamic model (12). However instead of the full covariance, we only consider a diagonal covariance matrix plus 4 non-zero off-diagonal entries $\sigma_{1,2}$ ($\sigma_{2,1}$) and $\sigma_{99,100}$ ($\sigma_{100,99}$). We focus on the correlation process in this example thus set $\mu_t \equiv 0$ and $\sigma_t \equiv 1$, for $t \in [0, 1]$. More specifically when generating data $\{y_t\}$ with (22), if $i \notin \{2, 100\}$ we set $i$-th rows $L_i = S_i = e_i$ with $e_i$ being the $i$-th row of identity matrix.

To apply our dynamical model (12) in this setting, we let $\mathbf{L}_t$ have '$w$-band' structure with $w = 2$ at each time $t$. Setting $s = 2$, $a = 1$, $b = 0.1$, $m = 0$ and $V = 10^{-3}$, $N = 100$ and $M = 100$, we repeat the MCMC runs for $7.5 \times 10^4$ iterations, burn in the first $2.5 \times 10^4$ and subsample 1 for every 10 to obtain $5 \times 10^3$ posterior samples in the end. Based on those samples, we estimate the underlying correlation functions and only plot $\rho_{1,2}$, $\rho_{49,50}$ and $\rho_{99,100}$ in Figure 7. With the '$w$-band' structure, we have less entries in the covariance matrix and focus on the 'in-group' correlation. Our dynamical model (12) is sensitive enough to discern the informative non-zero components from the non-informative ones in these correlation functions. Unit-vector GP priors provide flexibility for the model to capture the changing pattern of informative correlations. The left panel of Figure 7 shows that the model (12) correctly identify the non-zero components $\rho_{1,2}$ and $\rho_{99,100}$ and characterize their evolution. The right panel shows that the 2-norm distance between the estimated and true correlation matrices, $\|\widehat{P}(t) - P(t)\|_2$, is small, indicating that our dynamic model (12) performs well with higher dimension in estimating complex dependence structure among multiple stochastic processes.

## 4.3 | Analysis of Local Field Potentials

We will now use the proposed model (12) to analyze a local field potential (LFP) data. The goal of this analysis is to elucidate how memory and cognition arise from functional interactions among brain regions, by modeling how their

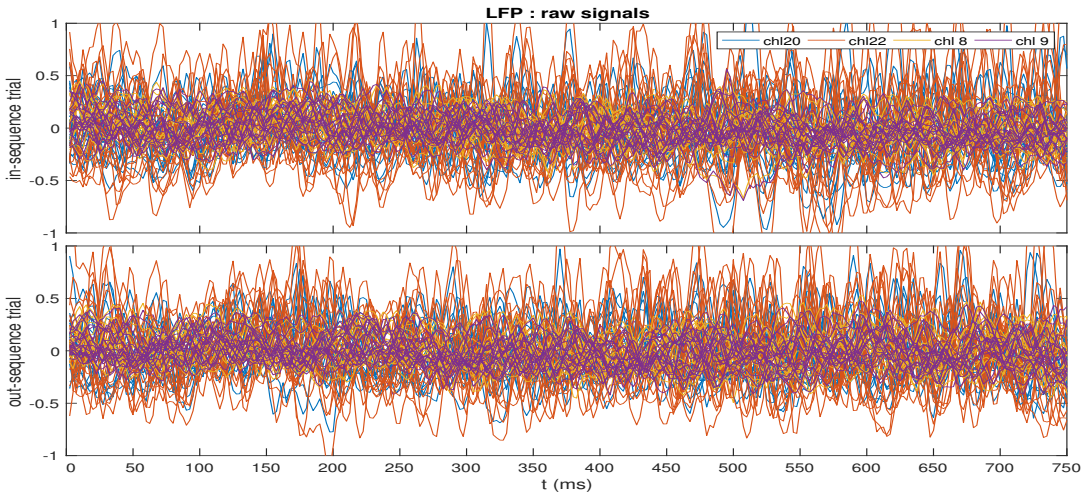


**FIGURE 8** LFP signals on "in sequence" and "out of sequence" trials. It is difficult to identify differences between the two conditions based on a mere visual inspection of the LPFs.

dynamic connectivity varies during performance of complex memory tasks. Here we focus on multi-site LFP data recorded from 24 electrodes spanning the dorsal CA1 subregion of the hippocampus as rats performed a hippocampus-dependent sequence memory task (Allen et al., 2016). The task involves repeated presentations of a sequence of odors (e.g., ABCDE) at a single port and requires rats to correctly determine whether each odor is presented 'in sequence' (InSeq; e.g., ABCDE; by holding their nosepoke response until the signal at 1.2s) or 'out of sequence' (OutSeq; e.g., ABDDE; by withdrawing their nose before the signal). In previous work using the same dataset, Holbrook et al. (2016) used a direct MCMC algorithm named positive definite Hamiltonian Monte Carlo (PDHMC) to study the spectral density matrix of LFP from 4 selected channels. However, they did not examine how their correlations varied across time and recording site. These limitations are addressed in this paper.

For practical reasons, we focused our analyses on the time window from 0ms to 750ms (with 0 corresponding to when the rat's nose enters the odor port). Critically, this includes a time period during which the behavior of the animal is held constant (0-500ms) so differences in LFP reflect the cognitive processes associated with task performance, and, to serve as a comparison, a time period near 750ms during which the behavioral state of the animal is known to be different (i.e., by 750ms the animal has already withdrawn from the port on the majority of OutSeq trials, but is still in the port on InSeq trials). We also focused our analyses on two sets of adjacent electrodes (electrodes 20 and 22, and electrodes 8 and 9), which allows for comparisons between probes that are near each other ($<$1mm; i.e., 20:22 and 8:9) or more distant from each other ($>$2mm; i.e., 20:8, 20:9, 22:8, and 22:9). Figure 8 shows $M = 20$ trials of these LFP signals from $D = 4$ channels under both InSeq and OutSeq conditions. Our main objective was to quantify how correlations among these LFP channels varied across trial types (InSeq vs OutSeq) and over time (within the first 750ms of trials). To do so, we discretize the time window of $0.75$ seconds into $N = 300$ equally spaced small intervals. Under each experiment condition (InSeq or OutSeq), we treat all the signals as a 4 dimensional time series and fit them using our proposed dynamic covariance (correlation) model (12) in order to discover the evolution of their relationship. Note, we model the mean processes, variance processes and correlation processes separately, which is more natural to statisticians.

We set $s = 2$, $a = (1, 1, 1)$, $b = (1, 0.1, 0.2)$, $m = (0, 0, 0)$, $V = (1, 1.2, 2)$, and the general results are not very sensitive to the choice of these fine-tuning parameters. We also scale the discretized time points into $(0, 1]$ and add an additional

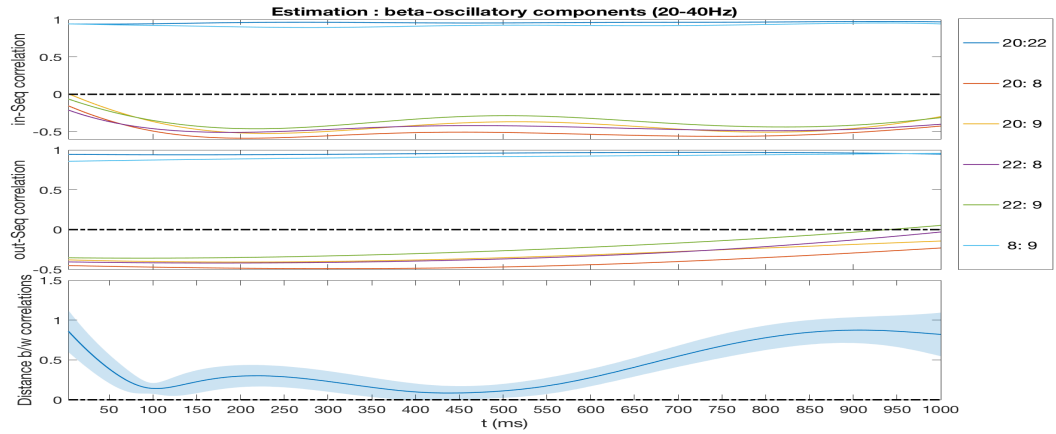


**FIGURE 9** Estimated correlation processes of LFPs (beta) under in-sequence condition (top), out-of-sequence condition (middle) and the (Frobenius) distance between two correlation matrices (bottom).

nugget of $10^{-5}I_n$ to the covariance kernel of GPs. We follow the same procedure in Section 3.1 to collect $7.5 \times 10^4$ samples, burn in the first $2.5 \times 10^4$ and subsample 1 for every 10. The resulting $10^4$ samples yield estimates of correlation processes as shown in Figure 9 for beta-filtered traces (20-40Hz) but similar patterns were also observed for theta-filtered traces (4-12Hz; see the supplement). The bottom panel of Figure 9 shows the dissimilarity between correlation processes under different conditions measured by the Frobenius norm of the difference between the two correlation matrices.

Our approach revealed many important patterns in the data. First, it showed that electrodes near each other (20:22 and 8:9) displayed remarkably high correlations in their LFP activity on InSeq and OutSeq trials, whereas correlations were considerably lower among more distant electrodes (20:8, 20:9, 22:8, and 22:9). Second, it revealed that the correlations between InSeq and OutSeq matrices evolved during the presentation of individual trials. These results are consistent with other analyses on learning (see, e.g., Fiecas and Ombao, 2016). As expected, InSeq and OutSeq activity was very similar at the beginning of the time window (e.g., before 350ms), which is before the animal has any information about the InSeq or OutSeq status of the presented odor, but maximally different at the end of the time window, which is after it has made its response on OutSeq trials. Most important, however, is the discovery of InSeq vs OutSeq differences before 500ms, which reveal changes in neural activity associated with the complex cognitive process of identifying if events occurred in their expected order. These findings highlight the sensitivity of our novel approach, as such differences have not been detected with traditional analyses.

## 5 | CONCLUSION

In this paper, we propose a novel Bayesian framework that grants full flexibility in modeling covariance and correlation matrices. The method is motivated by the statistical interpretation of covariances in terms of variances and correlations. It extends the separation strategy proposed by Barnard et al. (2000) and uses the Cholesky decomposition to maintain the positive definiteness of the correlation matrix. By defining priors on spheres for rows of the Cholesky factor of the correlation matrix, this framework introduces a series of flexible priors for covariances that go beyond the commonly used but restrictive inverse-Wishart distribution. These priors defined on spheres, including the squared-Dirichlet

distribution, and normalized Gaussian distributions (the Von Mises-Fisher distribution and the Bingham distribution), have freedom in controlling correlation size when modeling the covariance and correlation matrices. Adaptive $\Delta$-Spherical HMC is introduced to handle the intractability of the resulting posterior. What is more, we extend this structured scheme to dynamical models to capture complex dependence among multiple stochastic processes. The unit-vector Gaussian process is then introduced as a flexible prior process to model the evolution of correlation. The resulting framework was shown to be effective in Bayesian modeling covariance and correlation matrices using the Normal-inverse-Wishart problem, a simulated periodic process, and an analysis of LFP data.

In future work, we will explore the low-rank structure of covariance and correlation matrices to further scale our method to problems of greater dimensionality. For example, we can adopt the similar decomposition as in the latent factor process model (Fox and Dunson, 2015), and assume $\text{vech}^{\mathsf{T}}(\mathbf{L}_t) \in (\mathcal{S}^k)^D$ for some $k \ll D$.

It is novel, to the best of our knowledge, to systematically define the vector Gaussian process through the matrix normal distribution. In the current work, we only consider unit cross covariance matrix $\mathbf{V} = \mathbf{I}$; there is possibility for theoretic exploration in non-unit cross-covariances to better capture the relationships between component processes. Another future direction is to generalize our dynamic modeling of multivariate time series to dynamical regression on covariates $x \in \mathcal{X}$. It would be interesting to investigate how they influence the response along the process.

The limitation of this current analysis on LFP data is that it is conducted on a single rat. The proposed model will be generalized to account for variation among rats. In the future, we will apply this sensitive approach to other datasets, including simultaneous LFP recordings from multiple brain regions in rats as well as fMRI data collected from human subjects performing the same task.

## REFERENCES

Allen, T. A., Salz, D. M., McKenzie, S. and Fortin, N. J. (2016) Nonspatial sequence coding in ca1 neurons. *Journal of Neuroscience*, **36**, 1547–1563.

Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis.* Wiley Series in Probability and Statistics. Hoboken, N. J.: Wiley Interscience.

Banfield, J. D. and Raftery, A. E. (1993) Model-based gaussian and non-gaussian clustering. *Biometrics*, 803–821.

Barnard, J., McCulloch, R. and Meng, X.-L. (2000) Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 1281–1311.

Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997) Inference in model-based cluster analysis. *statistics and Computing*, **7**, 1–10.

Beskos, A., Pinski, F. J., Sanz-Serna, J. M. and Stuart, A. M. (2011) Hybrid monte carlo on hilbert spaces. *Stochastic Processes and their Applications*, **121**, 2201–2230.

Bingham, C. (1974) An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, 1201–1225.

Byrne, S. and Girolami, M. (2013) Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, **40**, 825–845.

Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models. *Pattern recognition*, **28**, 781–793.

Chiu, T. Y., Leonard, T. and Tsui, K.-W. (1996) The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, **91**, 198–210.

Clausen, T. (1832) Über die Function $\sin\phi + (1/2^2)\sin 2\phi + (1/3^2)\sin 3\phi +$ etc. *Journal für die reine und angewandte Mathematik*, **8**, 298–300.

Cribben, I., Haraldsdottir, R., Atlas, L. Y., Wager, T. D. and Lindquist, M. A. (2012) Dynamic connectivity regression: Determining state-related changes in brain connectivity. *NeuroImage*, **61**, 907 – 920. URL: .

Dahlhaus, R. (2000) A likelihood approximation for locally stationary processes. *Ann. Statist.*, **28**, 1762–1794. URL: .

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

Fiecas, M. and Ombao, H. (2016) Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association*, **111**, 1440–1453. URL: .

Fisher, N. I., Lewis, T. and Embleton, B. J. (1987) *Statistical analysis of spherical data*. Cambridge university press.

Fisher, R. (1953) Dispersion on a sphere. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 217, 295–305. The Royal Society.

Fox, E. B. and Dunson, D. B. (2015) Bayesian nonparametric covariance regression. *Journal of Machine Learning Research*, **16**, 2501–2542.

Girolami, M. and Calderhead, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, **(with discussion) 73**, 123–214.

Hoffman, M. D. and Gelman, A. (2014) The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research*, **15**, 1593–1623.

Holbrook, A., Lan, S., Vandenberg-Rodes, A. and Shahbaba, B. (2016) Geodesic lagrangian monte carlo over the space of positive definite matrices: with application to bayesian spectral density estimation. *arXiv preprint arXiv:1612.08224*.

Kent, J. T. (1982) The fisher-bingham distribution on the sphere. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71–80.

Lan, S. and Shahbaba, B. (2016) *Algorithmic Advances in Riemannian Geometry and Applications*, chap. 2, 25–71. Advances in Computer Vision and Pattern Recognition. Springer International Publishing, 1 edn.

Lan, S., Stathopoulos, V., Shahbaba, B. and Girolami, M. (2015) Markov chain monte carlo from lagrangian dynamics. *Journal of Computational and Graphical Statistics*, **24**, 357–378.

Lan, S., Zhou, B. and Shahbaba, B. (2014) Spherical hamiltonian monte carlo for constrained target distributions. vol. 32, 629–637. Beijing: The 31st International Conference on Machine Learning.

Leonard, T. and Hsu, J. S. (1992) Bayesian inference for a covariance matrix. *The Annals of Statistics*, 1669–1696.

Lindquist, M. A., Xu, Y., Nebel, M. B. and Caffo, B. S. (2014) Evaluating dynamic bivariate correlations in resting-state fmri: A comparison study and a new approach. *NeuroImage*, **101**, 531 – 546. URL: .

Liu, C. (1993) Bartlett' s decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. *Journal of Multivariate Analysis*, **46**, 198–206.

Magnus, J. R. and Neudecker, H. (1979) The commutation matrix: some properties and applications. *The Annals of Statistics*, 381–394.

— (1980) The elimination matrix: some lemmas and applications. *SIAM Journal on Algebraic Discrete Methods*, **1**, 422–449.

Mardia, K. V. and Jupp, P. E. (2009) *Directional statistics*, vol. 494. John Wiley & Sons.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1980) Multivariate analysis (probability and mathematical statistics).

Minka, T. (1997; revised 12/00) *Old and New Matrix Algebra Useful for Statistics*. MIT Media Lab note. URL: .

Murray, I., Adams, R. P. and MacKay, D. J. (2010) Elliptical slice sampling. *JMLR: W&CP*, **9**, 541–548.

Nason, G. P., Von Sachs, R. and Kroisandt, G. (2000) Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 271–292. URL: .

Neal, R. M. (2003) Slice sampling. *Annals of Statistics*, **31**, 705–767.

— (2011) MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (eds. S. Brooks, A. Gelman, G. Jones and X. L. Meng), 113–162. Chapman and Hall/CRC.

Nesterov, Y. (2009) Primal-dual subgradient methods for convex problems. *Mathematical programming*, **120**, 221–259.

Ombao, H., von Sachs, R. and Guo, W. (2005) Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, **100**, 519–531. URL: .

Onstott, T. C. (1980) Application of the bingham distribution function in paleomagnetic studies. *Journal of Geophysical Research: Solid Earth*, **85**, 1500–1510.

Pinheiro, J. C. and Bates, D. M. (1996) Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, **6**, 289–296.

Pourahmadi, M. (1999) Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 677–690.

— (2000) Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 425–435.

Prado, R. (2013) Sequential estimation of mixtures of structured autoregressive models. *Computational Statistics & Data Analysis*, **58**, 58 – 70. URL: . The Third Special Issue on Statistical Signal Extraction and Filtering.

Prado, R., West, M. and Krystal, A. D. (2001) Multichannel electroencephalographic analyses via dynamic regression models with time-varying lag–lead structure. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **50**, 95–109. URL: .

Priestley, M. (1965) Evolutionary and non-stationary processes. *J. Roy. Statist. Soc. Ser. B*, **27**, 204–237.

Rao, T. S. (1970) The fitting of non-stationary time-series models with time-dependent parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, **32**, 312–322. URL: .

Smith, W. and Hocking, R. (1972) Algorithm as 53: Wishart variate generator. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **21**, 341–345.

Sverdrup, E. (1947) Derivation of the wishart distribution of the second order sample moments by straightforward integration of a multiple integral. *Scandinavian Actuarial Journal*, **1947**, 151–166.

Ting, C. M., Seghouane, A. K., Salleh, S. H. and Noor, A. M. (2015) Estimating effective connectivity from fmri data using factor-based subspace autoregressive models. *IEEE Signal Processing Letters*, **22**, 757–761.

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A. and Tuerlinckx, F. (2011) Visualizing distributions of covariance matrices. *Columbia Univ., New York, USA, Tech. Rep*, 18–18.

Tracy, D. S. and Dwyer, P. S. (1969) Multivariate maxima and minima with matrix derivatives. *Journal of the American Statistical Association*, **64**, 1576–1594.

van der Vaart, A. W. and van Zanten, J. H. (2008) Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, **36**, 1435–1463. URL: .

West, M., Prado, R. and Krystal, A. D. (1999) Evaluation and comparison of eeg traces: Latent structure in nonstationary time series. *Journal of the American Statistical Association*, **94**, 375–387. URL: .

Wilson, A. G. and Ghahramani, Z. (2011) Generalised wishart processes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*.

Wishart, J. (1928) The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 32–52.

Yang, R. and Berger, J. O. (1994) Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 1195–1211.

## A | CONNECTION TO KNOWN PRIORS

The following lemma is essential in proving that our proposed methods (4) (6) generalize existing methods in specifying priors, including the inverse-Wishart distribution, and two uniform distributions (Barnard et al., 2000) as well.

**Lemma 1** *Let $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{U}^\mathsf{T}$ be the reversed Cholesky decomposition of $\boldsymbol{\Sigma}$. The Jocobian of the transformation $\boldsymbol{\Sigma} \mapsto \mathbf{U}$ is*

$$\left|\frac{d_h\boldsymbol{\Sigma}}{d_h\mathbf{U}^\mathsf{T}}\right| := \left|\frac{\partial \text{vech}\boldsymbol{\Sigma}}{\partial \text{vech}\mathbf{U}^\mathsf{T}}\right| = 2^D \prod_{i=1}^{D} |u_{ii}|^i$$

*Let $\mathbf{P} = \mathbf{L}\mathbf{L}^\mathsf{T}$ be the Cholesky decomposition of $\mathbf{P}$. The Jacobian of the transformation $\mathbf{L} \mapsto \mathbf{P}$ is*

$$\left|\frac{d_h\mathbf{L}}{d_h\mathbf{P}}\right| := \left|\frac{\partial \text{vech}\mathbf{L}}{\partial \text{vech}\mathbf{P}}\right| = 2^{-D} \prod_{i=1}^{D} |l_{ii}|^{i-(D+1)}$$

**Proof** Note we have

$$d\boldsymbol{\Sigma} = d\mathbf{U}\mathbf{U}^\mathsf{T} + \mathbf{U}d\mathbf{U}^\mathsf{T}$$

Taking vec on both sides and applying its property

$$d\text{vec}\boldsymbol{\Sigma} = (\mathbf{U} \otimes \mathbf{I})d\text{vec}\mathbf{U} + (\mathbf{I} \otimes \mathbf{U})d\text{vec}\mathbf{U}^\mathsf{T}$$

Applying the elimination $\mathbf{L}_D$ on both sides

$$\begin{aligned} d\text{vech}\boldsymbol{\Sigma} &= \mathbf{L}_D[(\mathbf{U} \otimes \mathbf{I})\mathbf{K}_D d\text{vec}\mathbf{U}^\mathsf{T} + (\mathbf{I} \otimes \mathbf{U})d\text{vec}\mathbf{U}^\mathsf{T}] = \mathbf{L}_D(\mathbf{K}_D + \mathbf{I})(\mathbf{I} \otimes \mathbf{U})d\text{vec}\mathbf{U}^\mathsf{T} \\ &= 2\mathbf{L}_D\mathbf{N}_D(\mathbf{I} \otimes \mathbf{U})\mathbf{L}_D^\mathsf{T} d\text{vech}\mathbf{U}^\mathsf{T} = 2\mathbf{L}_D\mathbf{N}_D\mathbf{L}_D^\mathsf{T}\mathbf{D}_D^\mathsf{T}(\mathbf{I} \otimes \mathbf{U})\mathbf{L}_D^\mathsf{T} d\text{vech}\mathbf{U}^\mathsf{T} \end{aligned}$$

where $\mathbf{K}_D$ is the *commutation matrix* such that $\mathbf{K}_D\text{vec}\mathbf{A} = \text{vec}\mathbf{A}^\mathsf{T}$ for matrix $\mathbf{A}_{D\times D}$, $\mathbf{N}_D := (\mathbf{K}_D + \mathbf{I})/2$, and $\mathbf{D}_D$ is the *duplication matrix* which is regarded as the inverse of the elimination matrix $\mathbf{L}_D$. The last equation is by $\mathbf{D}_D\mathbf{L}_D\mathbf{N}_D = \mathbf{N}_D = \mathbf{N}_D^\mathsf{T}$ (Lemma2.1 and Lemma3.5 in Magnus and Neudecker, 1980). Thus according to (Lemma3.4 and Lemma4.1 in Magnus and Neudecker, 1980) we have

$$\left|\frac{d_h\boldsymbol{\Sigma}}{d_h\mathbf{U}^\mathsf{T}}\right| = \left|\frac{\partial \text{vech}\boldsymbol{\Sigma}}{\partial \text{vech}\mathbf{U}^\mathsf{T}}\right| = |2\mathbf{L}_D\mathbf{N}_D\mathbf{L}_D^\mathsf{T}\mathbf{D}_D^\mathsf{T}(\mathbf{I} \otimes \mathbf{U})\mathbf{L}_D^\mathsf{T}| = 2^{D(D+1)/2}|\mathbf{L}_D\mathbf{N}_D\mathbf{L}_D^\mathsf{T}||\mathbf{L}_D(\mathbf{I} \otimes \mathbf{U}^\mathsf{T})\mathbf{D}_D| = 2^D \prod_{i=1}^{D} |u_{ii}^i|$$

By similar argument, we have

$$\left|\frac{d_h\mathbf{P}}{d_h\mathbf{L}}\right| = \left|\frac{\partial \text{vech}\mathbf{P}}{\partial \text{vech}\mathbf{L}}\right| = |2\mathbf{L}_D\mathbf{N}_D\mathbf{L}_D^\mathsf{T}\mathbf{D}_D^\mathsf{T}(\mathbf{L} \otimes \mathbf{I})\mathbf{L}_D^\mathsf{T}| = 2^{D(D+1)/2}|\mathbf{L}_D\mathbf{N}_D\mathbf{L}_D^\mathsf{T}||\mathbf{L}_D(\mathbf{L}^\mathsf{T} \otimes \mathbf{I})\mathbf{D}_D| = 2^D \prod_{i=1}^{D} |l_{ii}|^{D+1-i}.$$

Thus it completes the proof. ∎

**Proof of Theorem 1** We know that the density of $\boldsymbol{\Sigma} \sim \mathcal{W}_D^{-1}(\boldsymbol{\Psi}, \nu)$ is

$$p_{\mathcal{W}^{-1}}(\boldsymbol{\Sigma}; \boldsymbol{\Psi}, \nu) = \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{D\nu/2}\Gamma_D(\nu/2)} |\boldsymbol{\Sigma}|^{-(\nu+D+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right)$$

By Lemma 1 we have

$$p(\mathbf{U}) = p(\boldsymbol{\Sigma}) \left|\frac{d_h \boldsymbol{\Sigma}}{d_h \mathbf{U}^{\mathsf{T}}}\right| = 2^D p_{\mathcal{W}^{-1}}(\mathbf{U}\mathbf{U}^{\mathsf{T}}; \boldsymbol{\Psi}, \nu) \prod_{i=1}^{D} |u_{ii}^i| = \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{D(\nu-2)/2}\Gamma_D(\nu/2)} |\mathbf{U}|^{-(\nu+D+1)} \prod_{i=1}^{D} u_{ii}^i \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\mathbf{U}^{-\mathsf{T}}\mathbf{U}^{-1})\right).$$

Then the proof is completed. ∎

**Proof of Theorem 2** To prove the first result, we use Lemma 1

$$p(\mathbf{P}) = p(\mathbf{L}) \left|\frac{d_h \mathbf{L}}{d_h \mathbf{P}}\right| \propto \prod_{i=2}^{D} |\mathbf{l}_i|^{2\boldsymbol{\alpha}_i - 1} \prod_{i=1}^{D} |l_{ii}|^{i-(D+1)} = \prod_{i=1}^{D} |l_{ii}|^{(i-3)(D+1)}$$

On the other hand, from Equation (8) in Barnard et al. (2000), we have the density of marginally uniform distribution:

$$p(\mathbf{P}) \propto |\mathbf{P}|^{\frac{D(D-1)}{2}-1} (\prod_i \mathbf{P}_{ii})^{-\frac{D+1}{2}} = (\prod_{j=1}^{D} l_{jj}^2)^{\frac{D(D-1)}{2}-1} (\prod_{j=1}^{D}\prod_{i=1}^{j} l_{ii}^2)^{-\frac{D+1}{2}} = \prod_{j=1}^{D} |l_{jj}|^{(j-3)(D+1)}$$

where $\mathbf{P}_{ii}$ is the $i$-th principal minor of $\mathbf{P}$. Similarly by Lemma 1 we can prove the second result

$$p(\mathbf{P}) = p(\mathbf{L}) \left|\frac{d_h \mathbf{L}}{d_h \mathbf{P}}\right| \propto \prod_{i=2}^{D} |\mathbf{l}_i|^{2\boldsymbol{\alpha}_i - 1} \prod_{i=1}^{D} |l_{ii}|^{i-(D+1)} \propto 1 .$$

Therefore we have finished the proof. ∎

# B | SPHERICAL HAMILTONIAN MONTE CARLO

## B.1 | Derivation of the geometric integrator for SphHMC

The Lagrangian dynamics (14) on the sphere $\mathcal{S}^{D-1}(r)$ with the first $(D-1)$ coordinates can be split (14) into the following two smaller dynamics:

$$\begin{cases} \dot{\mathbf{q}}_{-D} = \mathbf{v}_{-D} \\ \dot{\mathbf{v}}_{-D} = -\mathbf{v}_{-D}^{\mathsf{T}} \boldsymbol{\Gamma}(\mathbf{q}_{-D}) \mathbf{v}_{-D} \end{cases} \tag{24a}$$

$$\begin{cases} \dot{\mathbf{q}}_{-D} = 0 \\ \dot{\mathbf{v}}_{-D} = -\nabla_{\mathbf{q}_{-D}}^{-1} \tilde{U}(\mathbf{q}_{-D}) \end{cases} \tag{24b}$$

where (24a) is the equation of geodesic on manifold $\mathcal{S}^D$ which has analytical solution; and (24b) has analytical solution. Both define volume preserving maps.

The mapping $\mathcal{I} : \mathbf{q}_{-D} \mapsto \mathbf{q} = (\mathbf{q}_{-D}, q_D)$ can be viewed as an imbedding of $\mathcal{S}_+^{D-1}$ into $\mathbb{R}^D$. Denote its Jacobian as

$d\mathcal{I}(\mathbf{q}) := \begin{bmatrix} \mathbf{I}_{D-1} \\ -\frac{\mathbf{q}_{-D}^{\mathsf{T}}}{q_D} \end{bmatrix}$. Then we have

$$d\mathcal{I}(\mathbf{q})^{\mathsf{T}} d\mathcal{I}(\mathbf{q}) = \mathbf{G}(\mathbf{q}_{-D}), \qquad d\mathcal{I}(\mathbf{q})\mathbf{G}(\mathbf{q}_{-D})^{-1} d\mathcal{I}(\mathbf{q})^{\mathsf{T}} = \mathcal{P}(\mathbf{q}) = \mathbf{I} - r^{-2}\mathbf{q}\mathbf{q}^{\mathsf{T}}$$
$$\nabla_{\mathbf{q}_{-D}} \tilde{U}(\mathbf{q}) = d\mathcal{I}(\mathbf{q})^{\mathsf{T}} \nabla_{\mathbf{q}} \tilde{U}(\mathbf{q}), \qquad \mathbf{v} = d\mathcal{I}(\mathbf{q})\mathbf{v}_{-D}, \qquad \mathbf{v}^{\mathsf{T}}\mathbf{v} = \mathbf{v}_{-D}^{\mathsf{T}} \mathbf{G}(\mathbf{q}_{-D})\mathbf{v}_{-D}$$

Then Equation (24a) has the following solution with full coordinates

$$\begin{aligned} \begin{bmatrix} \mathbf{q}(t) \\ \mathbf{v}(t) \end{bmatrix} &= \begin{bmatrix} \mathbf{I} & 0 \\ 0^{\mathsf{T}} & r^{-1}\|\mathbf{v}(0)\|_2 \end{bmatrix} \begin{bmatrix} \cos(r^{-1}\|\mathbf{v}(0)\|_2 t) & \sin(r^{-1}\|\mathbf{v}(0)\|_2 t) \\ -\sin(r^{-1}\|\mathbf{v}(0)\|_2 t) & \cos(r^{-1}\|\mathbf{v}(0)\|_2 t) \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ 0^{\mathsf{T}} & r\|\mathbf{v}(0)\|_2^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{q}(0) \\ \mathbf{v}(0) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{q}(0)\cos(r^{-1}\|\mathbf{v}(0)\|_2 t) + r\mathbf{v}(0)\|\mathbf{v}(0)\|_2^{-1}\sin(r^{-1}\|\mathbf{v}(0)\|_2 t) \\ -r^{-1}\mathbf{q}(0)\|\mathbf{v}(0)\|_2 \sin(r^{-1}\|\mathbf{v}(0)\|_2 t) + \mathbf{v}(0)\cos(r^{-1}\|\mathbf{v}(0)\|_2 t) \end{bmatrix} \end{aligned} \tag{25}$$

and Equation (24b) has the following solution in full coordinates

$$\begin{aligned} &\mathbf{q}(t) = \mathbf{q}(0) \\ &\mathbf{v}(t) = \mathbf{v}(0) - \frac{t}{2} di(\mathbf{q}(0)) \nabla_{\mathbf{q}_{-D}}^{-1} \tilde{U}(\mathbf{q}(0)) = \mathbf{v}(0) - \frac{t}{2}\mathcal{P}(\mathbf{q})\nabla_{\mathbf{q}} \tilde{U}(\mathbf{q}(0)) \end{aligned} \tag{26}$$

So numerically updating (26) for $h/2$, updating (25) for $h$ and updating (26) for another $h/2$ yield the integrator (15).

## B.2 | Reformulating Acceptance

At the end of the numerical simulation, a proposal $(\mathbf{q}_T, \mathbf{v}_T)$ is accepted according to the following probability

$$a_{sphHMC} = 1 \wedge \exp(-\Delta E), \qquad \Delta E = E(\mathbf{q}_T, \mathbf{v}_T) - E(\mathbf{q}_0, \mathbf{v}_0) \tag{27}$$

Such classic definition of acceptance probability can be reformulated by replacing $\Delta E$ in (27) with

$$\Delta E = \sum_{\tau=1}^{T} \Delta E_\tau \qquad \Delta E_\tau = E(\mathbf{q}_\tau, \mathbf{v}_\tau) - E(\mathbf{q}_{\tau-1}, \mathbf{v}_{\tau-1})$$

With (15) we can write

$$\begin{aligned}
\Delta E' =& E(\mathbf{q}', \mathbf{v}') - E(\mathbf{q}, \mathbf{v}) \\
=& \tilde{U}(\mathbf{q}') - \tilde{U}(\mathbf{q}) + \frac{1}{2}\mathbf{v}'^{\mathsf{T}}_{-D}\mathbf{G}(\mathbf{q}'_{-D})\mathbf{v}'_{-D} - \frac{1}{2}\mathbf{v}^{\mathsf{T}}_{-D}\mathbf{G}(\mathbf{q}_{-D})\mathbf{v}_{-D} \\
=& \Delta\tilde{U} - \frac{1}{2}\|\mathbf{v}\|_2^2 + \frac{1}{2}\left\|\mathbf{v}^+ - \frac{h}{2}\mathcal{P}(\mathbf{q}')\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}')\right\|_2^2 \\
=& \Delta\tilde{U} - \frac{1}{2}\|\mathbf{v}\|_2^2 + \frac{1}{2}\mathbf{v}^{+\mathsf{T}}\mathbf{v}^+ - \frac{h}{2}\mathbf{v}^{+\mathsf{T}}\mathcal{P}(\mathbf{q}')\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}') + \frac{h^2}{8}\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}')^{\mathsf{T}}\mathcal{P}(\mathbf{q}')\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}') \\
=& \Delta\tilde{U} - \frac{1}{2}\|\mathbf{v}\|_2^2 + \frac{1}{2}\|\mathbf{v}^-\|_2^2 - \frac{h}{2}\mathbf{v}^{+\mathsf{T}}\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}') + \frac{h^2}{8}\|\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}')\|^2_{\mathcal{P}(\mathbf{q}')} \\
=& \Delta\tilde{U} - \frac{1}{2}\|\mathbf{v}\|_2^2 - \frac{h}{2}\mathbf{v}^{+\mathsf{T}}\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}') + \frac{h^2}{8}\|\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}')\|^2_{\mathcal{P}(\mathbf{q}')} + \frac{1}{2}\|\mathbf{v}\|_2^2 - \frac{h}{2}\mathbf{v}^{\mathsf{T}}\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}) + \frac{h^2}{8}\|\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q})\|^2_{\mathcal{P}(\mathbf{q})} \\
=& \Delta\tilde{U} - \frac{h}{2}\left[\mathbf{v}'^{\mathsf{T}}\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}') + \mathbf{v}^{\mathsf{T}}\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q})\right] - \frac{h^2}{8}\left[\|\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}')\|^2_{\mathcal{P}(\mathbf{q}')} - \|\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q})\|^2_{\mathcal{P}(\mathbf{q})}\right]
\end{aligned}$$

where $\mathcal{P}(\mathbf{q}')\mathbf{v}^+ = \mathbf{v}^+$, $\mathcal{P}(\mathbf{q})\mathbf{v}^- = \mathbf{v}^-$, and $\|\mathbf{v}^+\|_2^2 = \|\mathbf{v}^-\|_2^2$. Accumulating the above terms over $\tau = 1, \cdots, T$ yields the reformulated acceptance probability (16).

We now prove the energy conservation theorem 3 (Beskos et al., 2011).

**Theorem 3** *Let $h \to 0$ we have the following energy conservation*

$$E(\mathbf{q}(T), \mathbf{v}(T)) - E(\mathbf{q}(0), \mathbf{v}(0)) = \tilde{U}(\mathbf{q}(T)) - \tilde{U}(\mathbf{q}(0)) - \int_0^T \langle \mathbf{v}(t), \mathbf{g}(\mathbf{q}(t))\rangle dt = 0$$

**Proof** With the second equation of Lagrangian dynamics (14) we have

$$\begin{aligned}
-\langle \mathbf{v}(t), \mathbf{g}(\mathbf{q}(t))\rangle &= \mathbf{v}(t)^{\mathsf{T}}\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}(t)) = \mathbf{v}_{-D}(t)^{\mathsf{T}}d\mathcal{I}(\mathbf{q})^{\mathsf{T}}\nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}(t)) = \mathbf{v}_{-D}(t)^{\mathsf{T}}\nabla_{\mathbf{q}_{-D}}\tilde{U}(\mathbf{q}(t)) \\
&= \mathbf{v}_{-D}(t)^{\mathsf{T}}\mathbf{G}(\mathbf{q}_{-D}(t))\left[\dot{\mathbf{v}}_{-D}(t) + \mathbf{v}^{\mathsf{T}}_{-D}(t)\boldsymbol{\Gamma}(\mathbf{q}_{-D}(t))\mathbf{v}_{-D}(t)\right] \\
&= \mathbf{v}_{-D}(t)^{\mathsf{T}}\mathbf{G}(\mathbf{q}_{-D}(t))\dot{\mathbf{v}}_{-D}(t) + \frac{1}{2}\mathbf{v}_{-D}(t)^{\mathsf{T}}d\mathbf{G}(\mathbf{q}_{-D}(t))\mathbf{v}_{-D}(t) \\
&= \frac{d}{dt}\frac{1}{2}\mathbf{v}_{-D}(t)^{\mathsf{T}}\mathbf{G}(\mathbf{q}_{-D}(t))\mathbf{v}_{-D}(t) = \frac{d}{dt}\frac{1}{2}\|\mathbf{v}(t)\|_2^2
\end{aligned}$$

Then we have the first equality hold because

$$-\int_0^T \langle \mathbf{v}(t), \mathbf{g}(\mathbf{q}(t))\rangle dt = \frac{1}{2}\|\mathbf{v}(T)\|_2^2 - \frac{1}{2}\|\mathbf{v}(0)\|_2^2$$

Lastly, from the first equation of Lagrangian dynamics (14)

$$\tilde{U}(\mathbf{q}(T)) - \tilde{U}(\mathbf{q}(0)) = \int_0^T \dot{\tilde{U}}(\mathbf{q}(t)) = \int_0^T \langle \dot{\mathbf{q}}(t), \nabla_{\mathbf{q}}\tilde{U}(\mathbf{q}(t))\rangle dt = \int_0^T \langle \mathbf{v}(t), \mathbf{g}(\mathbf{q}(t))\rangle dt$$

Therefore the second equality is proved.

# C | GRADIENT CALCULATION IN NORMAL-INVERSE-WISHART PROBLEM

We use the representation (6) and derive log-posterior (log-likelihood and log-prior) and the corresponding gradients for (19) using matrix calculus.

## C.1 | Gradients of log-likelihood

Denote $\mathbf{y}_n^* := (\mathbf{y}_n - \boldsymbol{\mu}_0)/\boldsymbol{\sigma}$. Then the log-likelihood becomes

$$\ell(\mathbf{y}^*; \boldsymbol{\sigma}, \mathbf{P}) = -N\mathbf{1}_D^\mathsf{T} \log \boldsymbol{\sigma} - \frac{N}{2} \log |\mathbf{P}| - \frac{1}{2} \sum_{n=1}^{N} \mathbf{y}_n^{*\mathsf{T}} \mathbf{P}^{-1} \mathbf{y}_n^*$$

$\left[\frac{\partial \ell}{\partial \boldsymbol{\tau}}\right]$. We calculate the gradient of log-likelihood with respect to $\boldsymbol{\sigma}$

$$\frac{\partial \ell}{\partial \sigma_k} = -N\sigma_k^{-1} + \sum_{n=1}^{N} \sum_i \frac{y_{ni}^*}{\sigma_i} \delta_{ik} (\mathbf{P}^{-1}\mathbf{y}_n^*)_i, \quad i.e. \ \frac{\partial \ell}{\partial \boldsymbol{\sigma}} = -N\boldsymbol{\sigma}^{-1} + \sum_{n=1}^{N} \mathrm{diag}(\mathbf{y}_n^*/\boldsymbol{\sigma})(\mathbf{P}^{-1}\mathbf{y}_n^*)$$

And with the transformation $\boldsymbol{\tau} = \log(\boldsymbol{\sigma})$ it becomes

$$\frac{\partial \ell}{\partial \boldsymbol{\tau}} = \frac{d\boldsymbol{\sigma}^\mathsf{T}}{d\boldsymbol{\tau}} \frac{\partial \ell}{\partial \boldsymbol{\sigma}} = \mathrm{diag}(\boldsymbol{\sigma}) \left[ -\frac{N}{\boldsymbol{\sigma}} + \sum_{n=1}^{N} \mathrm{diag}(\mathbf{y}_n^*/\boldsymbol{\sigma})(\mathbf{P}^{-1}\mathbf{y}_n^*) \right] = -N\mathbf{1}_D + \sum_{n=1}^{N} \mathrm{diag}(\mathbf{y}_n^*)(\mathbf{P}^{-1}\mathbf{y}_n^*)$$

$\left[\frac{\partial \ell}{\partial \mathbf{U}^*} \left(\frac{\partial \ell}{\partial \mathsf{L}}\right)\right]$. When $\mathbf{P} = \mathbf{U}^*(\mathbf{U}^*)^\mathsf{T}$, $\frac{1}{2}\log|\mathbf{P}| = \log|\mathbf{U}^*| = \mathbf{1}_D^\mathsf{T} \log|\mathrm{diag}(\mathbf{U}^*)|$ and thus we have

$$\frac{\partial \ell}{\partial \mathbf{U}^*} = -\frac{N\mathbf{I}_D}{\mathbf{U}^*} + \sum_{n=1}^{N} \frac{dg_n(\widetilde{\mathbf{U}})}{d\mathbf{U}^*}$$

where $\frac{\mathbf{I}_D}{\mathbf{U}^*} = \mathrm{diag}(\{(u_{ii}^*)^{-1}\})$ is a diagonal matrix formed by element-wise division, $\widetilde{\mathbf{U}} := (\mathbf{U}^*)^{-1}$ and $g_n(\widetilde{\mathbf{U}}) := -\frac{1}{2}\mathbf{y}_n^{*\mathsf{T}}\widetilde{\mathbf{U}}^\mathsf{T}\widetilde{\mathbf{U}}\mathbf{y}_n^*$.

Taking differential directly on $g_n(\mathbf{U}^*) := -\frac{1}{2}\mathbf{y}_n^{*\mathsf{T}}(\mathbf{U}^*)^{-\mathsf{T}}(\mathbf{U}^*)^{-1}\mathbf{y}_n^*$, and noting that differential and trace operators are exchangeable, we have

$$\begin{aligned} dg_n(\mathbf{U}^*) &= -\frac{1}{2}\mathrm{tr}(\mathbf{y}_n^{*\mathsf{T}} d(\mathbf{U}^*)^{-\mathsf{T}}(\mathbf{U}^*)^{-1}\mathbf{y}_n^* + \mathbf{y}_n^{*\mathsf{T}}(\mathbf{U}^*)^{-\mathsf{T}} d(\mathbf{U}^*)^{-1}\mathbf{y}_n^*) \\ &= \frac{1}{2}\left[\mathrm{tr}(\mathbf{y}_n^{*\mathsf{T}}(\mathbf{U}^*)^{-\mathsf{T}} d(\mathbf{U}^*)^\mathsf{T}\mathbf{P}^{-1}\mathbf{y}_n^*) + \mathrm{tr}(\mathbf{y}_n^{*\mathsf{T}}\mathbf{P}^{-1} d\mathbf{U}^*(\mathbf{U}^*)^{-1}\mathbf{y}_n^*)\right] \\ &= \mathrm{tr}(\mathbf{y}_n^{*\mathsf{T}}\mathbf{P}^{-1} d\mathbf{U}^*(\mathbf{U}^*)^{-1}\mathbf{y}_n^*) = \mathrm{tr}((\mathbf{U}^*)^{-1}\mathbf{y}_n^*\mathbf{y}_n^{*\mathsf{T}}\mathbf{P}^{-1} d\mathbf{U}^*) \end{aligned}$$

Conversion from differential to normal derivative form in the numerator layout (Minka, 1997; revised 12/00) yields

$$\frac{\partial g_n(\mathbf{U}^*)}{\partial (\mathbf{U}^*)^\mathsf{T}} = \mathrm{tril}((\mathbf{U}^*)^{-1}\mathbf{y}_n^*\mathbf{y}_n^{*\mathsf{T}}\mathbf{P}^{-1}), \quad i.e., \ \frac{\partial g_n(\mathbf{U}^*)}{\partial \mathbf{U}^*} = \mathrm{triu}(\mathbf{P}^{-1}\mathbf{y}_n^*\mathbf{y}_n^{*\mathsf{T}}(\mathbf{U}^*)^{-\mathsf{T}})$$

Finally, we have

$$\frac{\partial \ell}{\partial \mathbf{U}^*} = -\frac{N\mathbf{I}_D}{\mathbf{U}^*} + \mathrm{triu}(\mathbf{P}^{-1}\sum_{n=1}^{N} \mathbf{y}_n^*\mathbf{y}_n^{*\mathsf{T}}(\mathbf{U}^*)^{-\mathsf{T}})$$

When $\mathbf{P} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$, by similar argument as above, we have

$$\frac{\partial \ell}{\partial \mathbf{L}} = -\frac{N\mathbf{I}_D}{\mathbf{L}} + \mathrm{tril}(\mathbf{P}^{-1}\sum_{n=1}^{N}\mathbf{y}_n^*\mathbf{y}_n^{*\mathsf{T}}\mathbf{L}^{-\mathsf{T}})$$

## C.2 | Gradients of log-priors

The logarithm of conditional prior $p(\boldsymbol{\sigma}|\mathbf{U}^*)$ after transformation $\boldsymbol{\tau} = \log(\boldsymbol{\sigma})$ becomes

$$\begin{aligned}\log p(\boldsymbol{\tau}|\mathbf{U}^*) = \log p(\boldsymbol{\sigma}|\mathbf{U}^*) + \log\left|\frac{d\boldsymbol{\sigma}}{d\boldsymbol{\tau}}\right| &= \sum_{i=1}^{D}(i-(\nu+D+1))\tau_i - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})) + \sum_{i=1}^{D}\tau_i \\ &= \sum_{i=1}^{D}(i-(\nu+D))\tau_i - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}}))\end{aligned}$$

$\left[\frac{d}{d\boldsymbol{\tau}}\log p(\boldsymbol{\tau}|\mathbf{U}^*)\right]$. We calculate the derivative of $\log p(\boldsymbol{\tau}|\mathbf{U}^*)$ with respect to $\boldsymbol{\tau}$

$$\frac{d}{d\boldsymbol{\tau}}\log p(\boldsymbol{\tau}|\mathbf{U}^*) = \mathbf{i} - (\nu+D) + \frac{dg(\boldsymbol{\tau})}{d\boldsymbol{\tau}}$$

where $\mathbf{i} = [1,\cdots,D]^{\mathsf{T}}$, and $g(\boldsymbol{\tau}) = -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}}))$.

Noting that differential and trace operators are exchangeable, we have

$$\begin{aligned}dg(\boldsymbol{\tau}) &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}d\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}}) + \boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}d\,\mathrm{diag}(e^{-\boldsymbol{\tau}})) \\ &= \frac{1}{2}\left[\mathrm{tr}(\mathbf{P}^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\,\mathrm{diag}(d\boldsymbol{\tau})) + \mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\,\mathrm{diag}(d\boldsymbol{\tau}))\right] \\ &= \sum_{i=1}^{D}d\tau_i\sum_{j=1}^{D}\psi_{ij}e^{-\tau_j}\rho^{ji}e^{-\tau_i}\end{aligned}$$

Thus

$$\frac{dg(\boldsymbol{\tau})}{d\boldsymbol{\tau}} = \mathrm{diag}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1})\,\mathrm{diag}(e^{-\boldsymbol{\tau}}) = \mathrm{diag}(\mathbf{P}^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\boldsymbol{\Psi})\,\mathrm{diag}(e^{-\boldsymbol{\tau}}) = \mathrm{diag}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1})\circ e^{-\boldsymbol{\tau}}$$

where diag acting on a vector forms a diagonal matrix while the action a matrix means extracting the diagonal vector. $\circ$ is the Hadamard product (a.k.a. Schur product), i.e. the entrywise product.

$\left[\frac{d}{d\mathbf{U}^*}\log p(\mathbf{U}^*|\boldsymbol{\tau})\right]$. Now consider the derivative of $\log p(\mathbf{U}^*|\boldsymbol{\tau})$ with respect to the matrix $\mathbf{U}^*$. We have

$$\frac{d}{d\mathbf{U}^*}\log p(\mathbf{U}^*|\boldsymbol{\tau}) = \frac{\mathrm{diag}(\mathbf{i}-(\nu+D+1))}{\mathbf{U}^*} + \frac{dg(\mathbf{U}^*)}{d\mathbf{U}^*}$$

where $g(\mathbf{U}^*) = -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})(\mathbf{U}^*)^{-\mathsf{T}}(\mathbf{U}^*)^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}}))$, and $\frac{\mathrm{diag}(\mathbf{i})}{\mathbf{U}^*}$ is a diagonal matrix formed by element-wise division.

Again by the exchangeability between differential and trace, we have

$$\begin{aligned} dg(\mathbf{U}^*) &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})d(\mathbf{U}^*)^{-\mathsf{T}}(\mathbf{U}^*)^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}}) + \boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})(\mathbf{U}^*)^{-\mathsf{T}}d(\mathbf{U}^*)^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})) \\ &= \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})(\mathbf{U}^*)^{-\mathsf{T}}d(\mathbf{U}^*)^{\mathsf{T}}\mathbf{P}^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})) + \mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}d\mathbf{U}^*(\mathbf{U}^*)^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}}))\right] \\ &= \frac{1}{2}\left[\mathrm{tr}(\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}d\mathbf{U}^*(\mathbf{U}^*)^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\boldsymbol{\Psi}) + \mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}d\mathbf{U}^*(\mathbf{U}^*)^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}}))\right] \\ &= \mathrm{tr}((\mathbf{U}^*)^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}d\mathbf{U}^*) \end{aligned}$$

Therefore we have

$$\frac{dg(\mathbf{U}^*)}{d(\mathbf{U}^*)^{\mathsf{T}}} = \mathrm{tril}((\mathbf{U}^*)^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\mathbf{P}^{-1}),$$

that is,

$$\frac{dg(\mathbf{U}^*)}{d\mathbf{U}^*} = \mathrm{triu}(\mathbf{P}^{-1}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})\boldsymbol{\Psi}\,\mathrm{diag}(e^{-\boldsymbol{\tau}})(\mathbf{U}^*)^{-\mathsf{T}}),$$

$\left[\frac{d}{d\boldsymbol{\tau}}\log p(\boldsymbol{\tau}),\ \frac{d}{d\mathbf{L}}\log p(\mathbf{L})\right]$. Lastly, the log-priors for (21) and their gradients after transformation $\boldsymbol{\tau} := \log(\boldsymbol{\sigma})$ are calculated

$$\log p(\boldsymbol{\tau}) = -\frac{1}{2}\boldsymbol{\tau}^{\mathsf{T}}\boldsymbol{\tau}, \qquad \frac{d}{d\boldsymbol{\tau}}\log p(\boldsymbol{\tau}) = -\boldsymbol{\tau}$$

$$\log p(\mathbf{l}_i) = \log p(\mathbf{l}_i^2) + \mathbf{1}_i^{\mathsf{T}}\log|2\mathbf{l}_i| = (2(\boldsymbol{\alpha}_i - 1) + \mathbf{1}_i)^{\mathsf{T}}\log|\mathbf{l}_i|, \qquad \frac{d}{d\mathbf{l}_i}\log p(\mathbf{l}_i) = \frac{2(\boldsymbol{\alpha}_i - 1) + \mathbf{1}_i}{\mathbf{l}_i}$$

The bottom row can be written as

$$\log p(\mathbf{L}) = \sum_{i=1}^{D}(2\boldsymbol{\alpha}_i - 1)^{\mathsf{T}}\log|\mathbf{l}_i|, \qquad \frac{d}{d\mathbf{L}}\log p(\mathbf{L}) = \frac{2\boldsymbol{\alpha} - 1}{\mathbf{L}}$$

where $\frac{1}{\mathbf{L}}$ denotes a lower-triangular matrix with $l_{ij}^{-1}$ being its $(i, j)$ entry $(i \geq j)$.

## D | MORE NUMERICAL RESULTS

## E | FLEXIBILITY OF VON MISES-FISHER PRIOR AND BINGHAM PRIOR

Now let's consider the following von Mises-Fisher prior (Fisher et al., 1987; Fisher, 1953; Mardia and Jupp, 2009) for $\mathbf{l}_i$, the $i$-th row of the Cholesky factor $\mathbf{L}$ of correlation matrix $\mathbf{P}$ in the structured model (6).

**Definition 6 (Von Mises-Fisher distribution)** *The probability density function of the von Mises-Fisher distribution for the random vector $\mathbf{l}_i \in \mathcal{S}^{i-1}$ is given by*

$$p(\mathbf{l}_i) = C_i(\kappa)\exp(\kappa\boldsymbol{\mu}^{\mathsf{T}}\mathbf{l}_i)$$

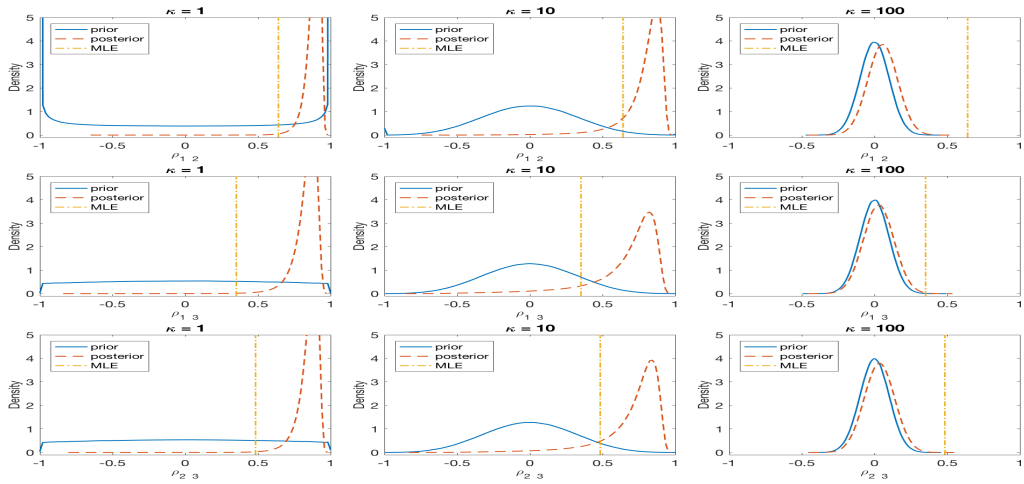


**FIGURE 10** Marginal posterior, prior (induced from von Mises-Fisher distribution) densities of correlations and MLEs with different settings for concentration parameter $\kappa$, estimated with $10^6$ samples.

*where* $\kappa \geq 0$, $\|\boldsymbol{\mu}\| = 1$ *and the normalization constant* $C_i(\kappa)$ *is equal to*

$$C_i(\kappa) = \frac{\kappa^{i/2-1}}{(2\pi)^{i/2} I_{i/2-1}(\kappa)}$$

*where* $I_\nu$ *denotes the modified* Bessel *function of the first kind at order* $\nu$. *Denote* $\mathbf{l}_i \sim \text{vMF}(\kappa, \boldsymbol{\mu})$.

Since we have no prior knowledge about the mean direction $\boldsymbol{\mu}$, we choose $\boldsymbol{\mu} = \mathbf{n}_i = (\mathbf{0}_{i-1}, 1)^\mathsf{T}$ that favors the polar direction, i.e.

$$\mathbf{l}_i \sim \text{vMF}(\kappa, \mathbf{n}_i), \qquad p(\mathbf{l}_i) \propto \exp(\kappa l_{ii}), \quad i = 2, \cdots, D$$

where we consider i) $\kappa = 1$; ii) $\kappa = 10$; iii) $\kappa = 100$. With the von Mises-Fisher prior, we have

$$\log p(\mathbf{L}) = \sum_{i=1}^{D} \kappa l_{ii} = \kappa \text{tr}(\mathbf{L}), \quad \frac{d}{d\mathbf{L}} \log p(\mathbf{L}) = \kappa \mathbf{I}$$

We repeat the experiment in Section 4.1.2 with the von Mises-Fisher prior for $\mathbf{l}_i$. The posteriors, priors and maximal likelihood estimates (MLE) of correlations with different $\kappa$'s are plotted in Figure 10 respectively. With larger concentration parameter $\kappa$, the posterior is pulled more towards 0.

Finally, we consider the following Bingham prior (Bingham, 1974; Onstott, 1980) for $\mathbf{l}_i$ in the structured model (6).

**Definition 7 (Bingham distribution)** *The probability density function of the Bingham distribution for the random vector* $\mathbf{l}_i \in \mathcal{S}^{i-1}$ *is given by*

$$p(\mathbf{l}_i) = {}_1F_1(\frac{1}{2}; \frac{n}{2}; \mathbf{Z})^{-1} \exp(\mathbf{l}_i^\mathsf{T} \mathbf{M}\mathbf{Z}\mathbf{M}^\mathsf{T} \mathbf{l}_i)$$

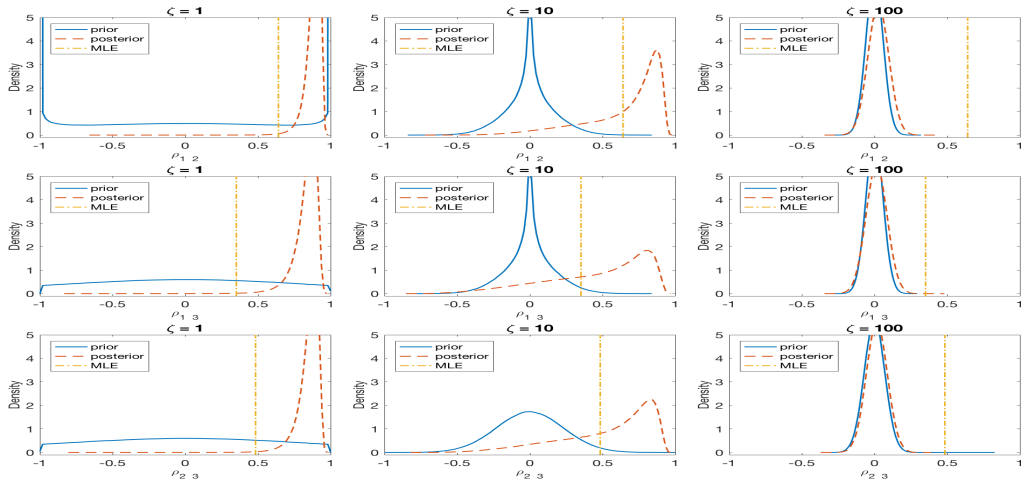


**FIGURE 11** Marginal posterior, prior (induced from Bingham distribution) densities of correlations and MLEs with different settings for concentration parameter $\zeta$, estimated with $10^6$ samples.

*where* $\mathbf{M}$ *is an orthogonal orientation matrix,* $\mathbf{Z}$ *is a diagonal concentration matrix, and* ${}_1F_1(\cdot;\ \cdot;\ \cdot)$ *is a* confluent hypergeometric function *of matrix argument. Denote* $\mathbf{l}_i \sim \text{Bing}(\mathbf{M}, \mathbf{Z})$.

Note, according to Bingham (1974), this distribution is defined for $\mathbf{Z}$ up to an arbitrary scalar matrix $\zeta_0\mathbf{I}$. Therefore, we consider $\mathbf{M} = \mathbf{I}$ and $\mathbf{Z} = \zeta\,\text{diag}(\mathbf{n}_i)$ that favors the polar direction, i.e.

$$\mathbf{l}_i \sim \text{Bing}(\mathbf{I}, \zeta\,\text{diag}(\mathbf{n}_i)), \qquad p(\mathbf{l}_i) \propto \exp(\zeta l_{ii}^2), \quad i = 2, \cdots, D$$

where we consider i) $\zeta = 1$; ii) $\zeta = 10$; iii) $\zeta = 100$. The log-prior and its gradient are calculated as follows

$$\log p(\mathbf{L}) = \sum_{i=1}^{D} \zeta l_{ii}^2 = \zeta \|\,\text{diag}(\mathbf{L})\|^2, \quad \frac{d}{d\mathbf{L}} \log p(\mathbf{L}) = 2\zeta\,\text{diag}(\mathbf{L})$$

We repeat the above experiment with the Bingham prior for $\mathbf{l}_i$. The posteriors, priors and maximal likelihood estimates (MLE) of correlations with different $\zeta$'s are plotted in Figure 11 respectively. With larger concentration parameter $\zeta$, the posteriors are pulled more towards the induced priors and concentrate on 0.

## E.1 | More Comparison to Latent Factor Process Model

The example of simulated periodic process in Section 4.2.2 is consider for $D = 2$ for simplicity and convenience of visualization. Here we consider higher dimension $D = 10$. The purpose here is not to show the scalability, but rather to investigate the robustness of our dynamic model (12) in terms of full flexibility.

We generate $M = 20$ trials of data over $N = 100$ evenly spaced points over $[0, 1]$. The true mean, variance and correlation functions are modified from the example (22) using the Clausen functions (Clausen, 1832). Seen from Figure 12, they behave more intricately with higher heterogeneity among those processes. This could impose further challenge for latent factor based models like (23) compared to $D = 2$. We repeat the experiments in Section 4.2.2 and compare

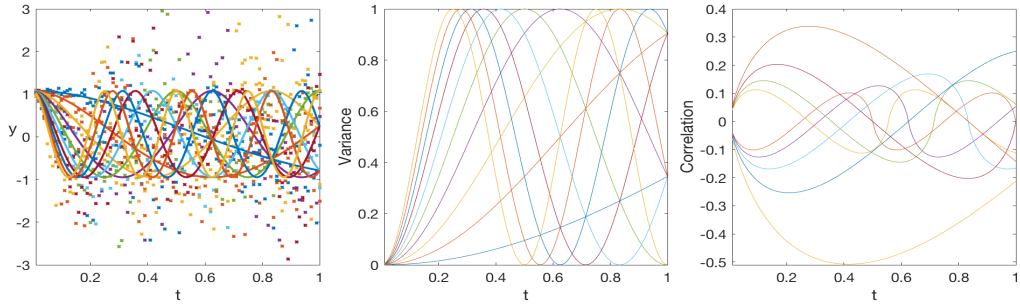


**FIGURE 12** Simulated data $y$ over the underlying mean functions $\mu_t$ (left), the variance functions $\Sigma_t$, and the correlation functions $\mathrm{P}_t$ (right) of 10-dimension periodic processes.

our dynamic model (12) with the latent factor process model (23) by Fox and Dunson (2015). To aid the visualization, we subtract the estimated process from their true values and plot the error functions in Figure 13. Even if we have tried our best to tune the parameters, e.g. $L$, the number of basis functions, and $k$, the size of latent factors, the latent factor process model (Fox and Dunson, 2015) is outperformed by our flexible dynamic model (12) in reducing estimation errors.

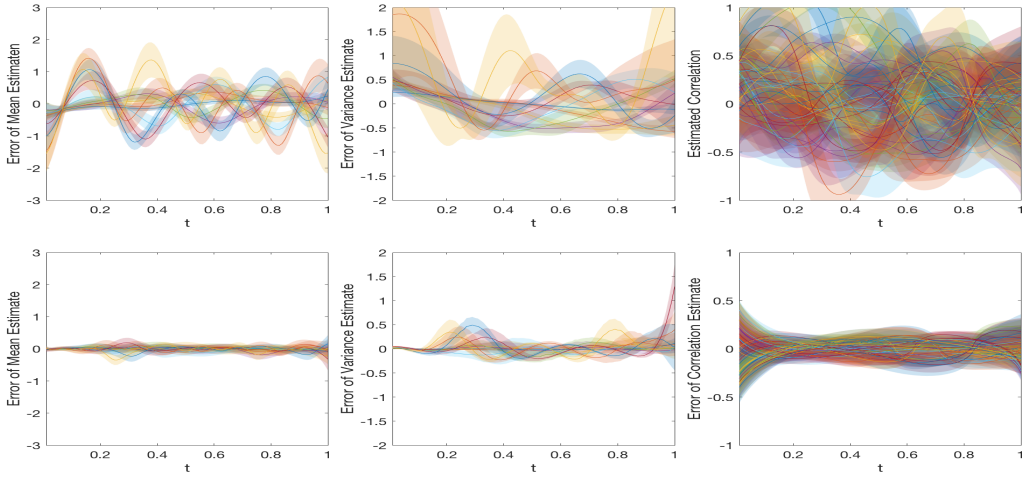


**FIGURE 13** Estimated error functions of the underlying mean $\mu_t$ (left column), variance $\sigma_t$ (middle column) and correlation $\rho_t$ (right column) of 10-dimensional periodic processes, using latent factor process model (upper row) and our flexible model (lower row), based on $M = 20$ trials of data over $N = 100$ evenly spaced points. Solid lines are estimated errors and shaded regions are 95% credible bands.

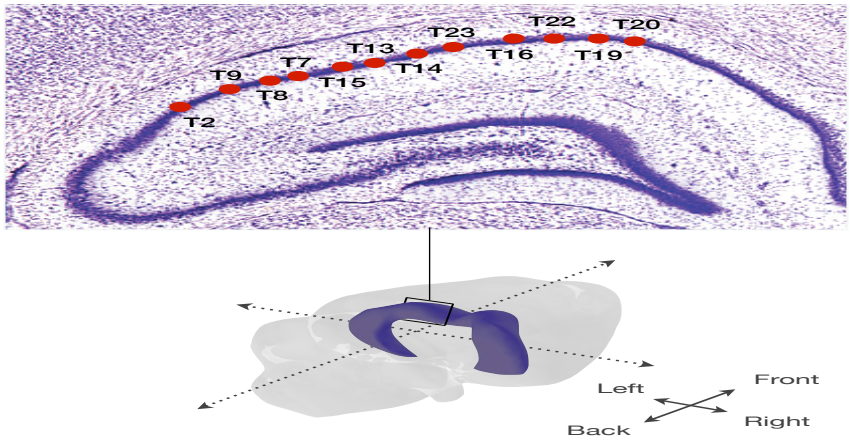


**FIGURE 14** Locations of recorded LFP signals in CA1 subregion of the hippocampus of the rat.

## E.2 | More Results on the Analysis of LFP data

In Section 4.3, we studied the LFP data collected from the hippocampus of rats performing a complex sequence memory task. Figure 14 shows 12 locations from CA1 subregion of the hippocampus of the rat where LFP signals are record. Figure 15 shows the theta-filtered traces (4-12Hz; left panel) and the estimated correlation processes under different experiment conditions (InSeq vs OutSeq; right panel). Here we observe the similar dynamic pattern of correlation matrices under two conditions that diverge after 600ms, indicating the neural activity associated with the cognitive process of identifying whether events occurred in their expected order.

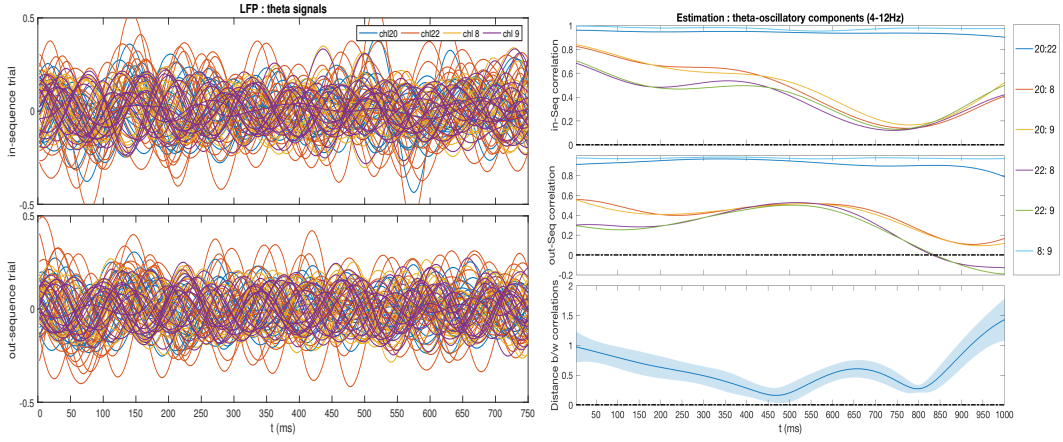


**FIGURE 15** Results of LFP theta signals: data (left), estimation of correlations (right).