Computational Statistics and Data Analysis 185 (2023) 107763

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Dynamic covariance estimation via predictive Wishart process with an application on brain connectivity estimation

Rui Meng [a,b,*,1], Fan Yang [c,1], Won Hwa Kim [d]

[a] Biological Sciences and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[b] Redwood Center for Theoretical Neuroscience, University of California, Berkeley, CA, USA
[c] University of Texas at Arlington, Arlington, TX, USA
[d] Pohang University of Science and Technology (POSTECH), Pohang, South Korea

ARTICLE INFO

Article history:
Received 25 April 2022
Received in revised form 7 April 2023
Accepted 8 April 2023
Available online 14 April 2023

Keywords:
Temporal dependence
Stochastic process
Covariance estimation
Bayesian inference
Variational inference
fMRI

ABSTRACT

Modeling the complex dependence in multivariate time series data is a fundamental problem in statistics and machine learning. Traditionally, the task has been approached with methods such as multivariate autoregressive models and multivariate generalized autoregressive conditional heteroskedasticity models, and Gaussian process based methods are recently becoming popular by leveraging the flexibility of non-parametric learning. However, few methods exist that directly model the dynamics of the covariance matrices except generalized Wishart process ($\mathcal{GWP}$), and even the generalized Wishart process is limited with applications on small dataset due to the extremely high computational capacity induced by multiple Gaussian processes. In this regard, a novel stochastic process named as Predictive Wishart Process ($\mathcal{PWP}$) is proposed, which provides a collection of positive semi-definite random matrices indexed by input variables. The $\mathcal{PWP}$ projects process realizations of $\mathcal{GWP}$ to a lower dimensional subspace to efficiently estimate every $\mathcal{GWP}$. The theoretical properties of it are examined, and both Bayesian inference and efficient variational expectation maximization are explored in relation to it. Moreover, the $\mathcal{PWP}$ is empirically tested on synthetically generated time-series data to validate competitive reconstructive performance and efficient predictive performance, and applied on a large-scale real functional magnetic resonance imaging (fMRI) dataset from Human Connectome Project (HCP) to demonstrate its practicality. A thorough statistical analysis with visualizations is conducted on the brain connectivity, and also a PWP-based multi-task learning framework is proposed to extract meaningful features from individual fMRIs.

1. Introduction

Accurate estimation of associations over a set of variables is a fundamental problem in statistics (and machine learning) with significant interest from diverse domains. Typically, the associations (e.g., covariance) are assumed to be static, and they are often estimated using structural equation models or graphical models (Biswal et al., 1995; Greicius, 2008; Biswal, 2012). However, when the given data are time-dependent, they often exhibit heteroscedasticity, i.e., the variances and correlations

* Corresponding author at: Biological Sciences and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
E-mail address: rmeng@lbl.gov (R. Meng).

[1] Rui Meng and Fan Yang are joint first authors.

https://doi.org/10.1016/j.csda.2023.107763

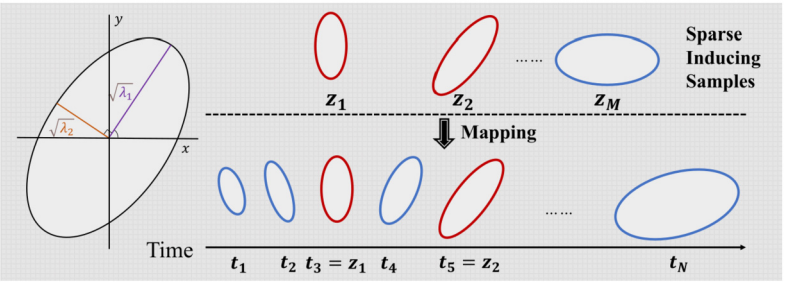


Fig. 1. A draw from a Predictive Wishart Process ($\mathcal{PWP}$). Each ellipse is a 2×2 covariance matrix index by observed time $\{t_i\}_{i=1}^N$ or inducing time $\{z_j\}_{j=1}^M$. The rotation indicates the correlation between the two variables, and the major and minor axes scale with the eigenvalues (i.e., λ_1, λ_2) of the matrix. A draw from a PWP consists of two steps: (i), we draw a collection of matrices indexed by inducing time; (ii), we map the collection of matrices to another collection of matrices indexed by observed time.

of variables of interest are time-varying (Dai et al., 2016; Seiler and Holmes, 2017; Zhu et al., 2019; Meng et al., 2021, 2022). Therefore, accounting for both temporal and spatial dependency in the covariance is critical in various motivating applications, e.g., capturing the time-varying volatility of a collection of risky assets in econometrics (Cappiello et al., 2006; Fox and West, 2011), and modeling spatial variations in correlations for customarily recorded multivariate measurements at a large collection of locations for geoscience (Gelfand et al., 2005; Fox and Dunson, 2015).

Such a problem routinely arise in brain connectivity analyses in Neuroimaging, which often requires estimating covariance from a knot of measurements (e.g., timeseries) across spatially parcellated Regions of Interest (ROIs) in the brain. Here, the covariance quantifies the level of associations between different ROIs as a functional connectivity (Smith, 2012). Conventional connectivity constructions assume that the functional associations are static in time over the entire scan period (Varoquaux et al., 2010; Chai et al., 2009). Nevertheless, several studies demonstrate that the functional connectivities *change over time* whose temporal variation may be significant (Hutchison et al., 2013; Hindriks et al., 2016). Therefore, deriving dynamic associations between ROIs is an important problem for both statistics and neuroscience, which investigates the time-varying co-activation patterns in the brain activities (Hutchison et al., 2013; Keilholz, 2014; Li et al., 2019).

Unfortunately, modeling such dynamic changes of covariance is quite challenging, because the given data are often in a large scale in length and typically only a single observation is recorded at each time stamp. In the statistical literature, modeling the dynamics of covariance has been tackled with Multivariate Generalized Autoregressive Conditional Heteroskedasticity (MGARCH) models (Engle, 2002), and alternative approaches were proposed such as Bayesian nonparametric models based on Wishart process (WP) (Fox and West, 2011; Wilson and Ghahramani, 2010). However, recent works including Generalized Wishart Process ($\mathcal{GWP}$) on Bayesian inference for WP are limited as they often require extremely high computational capacity due to the burden introduced from latent Gaussian processes, and hence makes it difficult to scale down for practical model inference.

To tackle the problem above, we develop Predictive Wishart Process ($\mathcal{PWP}$), which is a novel *parsimonious stochastic process* which approximates the traditional $\mathcal{GWP}$. We thoroughly study the stochastic properties of the $\mathcal{PWP}$ and provide full Bayesian posterior inference, which has been dismissed in previous literature. This framework is *scalable* to generate time-varying covariance $\Sigma(x)$ for a given index x from large-scale data (see Fig. 1) under rigorous mathematical properties. The complexity of generating time-varying covariance matrices is *linear* with respect to the number of covariance matrices (N) as opposed to $\mathcal{GWP}$ whose complexity of generating latent variables in each GP is *cubic* in N. Due to the parsimony of the predictive process, both Bayesian and variational inferences of the dynamic covariance structure with $\mathcal{PWP}$ become efficient.

The main **contributions** of our work are summarized as:

(i) We introduce a novel matrix variate stochastic process and theoretically demonstrate its desirable properties;
(ii) We propose Markov chain Monte Carlo (MCMC) and variational expectation maximization inference associated with a hierarchical Gaussian model and illustrate both computational benefits and comparable predictive performance of $\mathcal{PWP}$;
(iii) We provide a multi-task learning framework using $\mathcal{PWP}$ to jointly model multiple large-scale signals, and empirically prove the efficiency and practicality of $\mathcal{PWP}$ by tackling a real large-scale problem where conventional methods fail.

Extensive experiments are carried on synthetic experiments (with ground truth) as well as on a large-scale real Neuroimaging study (i.e., Human Connectome Project (HCP)) with resting-state functional MRI (fMRI) (WU-Minn, 2017) for reconstruction and prediction of dynamic covariances. Utilizing $\mathcal{PWP}$ leads to improvement in characterizing behavioral scores with dynamic covariance; our pioneering exploration on modeling dynamic connectivity should be worth pursuing further.

2. Related works

There exists a large body of literature on modeling time-varying covariance matrix, and classical strategies for estimating the covariance rely on standard regression methods with the Cholesky decomposition of the covariance or precision matrices (Pourahmadi, 1999; Zhang and Leng, 2012). Alternatively, nonparametric approaches have been proposed in Yin et al. (2010); Fox and Dunson (2015).

For modeling multivariate time series, heteroscedastic modeling has a long history, where the main approaches including multivariate GARCH based models (Engle, 2002; Engle and Kroner, 1995; Engle and Sheppard, 2001), dynamic conditional correlation models (Lindquist et al., 2014; Lee and Kim, 2021), sliding-window based approach (Monti et al., 2014), multivariate stochastic volatility models (Chib et al., 2006; Kastner et al., 2017) and Wishart process (Gouriéroux et al., 2009; Wilson and Ghahramani, 2010). Specifically, Lindquist et al. (2014) focus on the dynamic conditional correlation model (DCC) and show that DCC outperforms the exponential weighted moving average (EWMA) approach and sliding-window based approach. Lee and Kim (2021) extend the DCC approach to a copula-based DCC to release the Gaussian distribution of the data. Monti et al. (2014) propose the smooth incremental Graphical Lasso estimation algorithm which considers both sparsity and temporal homogeneity in the covariance estimation. Warnick et al. (2018) model the dynamic functional network connectivity using a hidden Markov model.

Our approach is a Bayesian nonparametric model based on Wishart process, allowing a feasible modeling of spatial and temporal correlation of data. There exist two Wishart process based methods: Wishart autoregressive processes (Gouriéroux et al., 2009) that construct positive definite volatility matrices with latent autoregressive (AR) models, and generalized Wishart process ($\mathcal{GWP}$) (Wilson and Ghahramani, 2010) that utilize Gaussian process to model latent process instead of AR models. Due to the limited expressiveness of AR models, Wishart autoregressive process cannot handle the long temporal dependence. On the other hand, $\mathcal{GWP}$ led to a diverse class of covariance dynamics, but it is not scalable to large datasets due to the expensive computation induced from corresponding latent Gaussian processes. Our approach attains the best of both worlds by utilizing a predictive process to model the dependence within those latent functions.

3. Preliminary

In this section, we briefly review a predictive process ($\mathcal{PP}$) (Banerjee et al., 2008; Finley et al., 2009), as it sets the foundation of our proposed $\mathcal{PWP}$ construction. We begin with distributions over functions $u(x)$ using Gaussian process ($\mathcal{GP}$) as

$$u(x) \sim \mathcal{GP}(m(x), C(x, x')), \tag{1}$$

with a mean function $m(x)$ and a covariance function $C(x, x')$ of choice specified with hyper-parameters τ, and we will refer to it as the parent process.

In the remainder of this paper, we consider a zero-mean Gaussian process, i.e., $m(x) \equiv 0$. Given a collection of inducing inputs $\boldsymbol{z} = (z_1, \ldots, z_M)$, the collection of function values $\boldsymbol{u}$ has a joint Gaussian distribution as

$$\boldsymbol{u} = (u(z_1), \ldots, u(z_M))^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}^*), \tag{2}$$

where $\boldsymbol{C}^*$ is the covariance matrix introduced by the covariance function $C(x, x')$ on inducing points $\boldsymbol{z}$.

A predictive process, i.e. $\mathcal{PP}$, is derived from its parent process (1) on a completely specified lower dimensional subspace. Specifically, given (1), the predictive process is defined as $\tilde{u}(x) \sim \mathcal{PP}(0, C(x, x')) = \mathcal{GP}(0, \tilde{C}(x, x'))$ that is equivalent to a new specified Gaussian process defined by the covariance function

$$\tilde{C}(x, x') = \boldsymbol{c}^T(x)\boldsymbol{C}^{*-1}\boldsymbol{c}(x'), \tag{3}$$

where $\boldsymbol{c}(x) = (C(x, z_1), \ldots, C(x, z_M))^T$. Here, two major properties of $\mathcal{PP}$ are given (Banerjee et al., 2008):

$$\tilde{u}(x) = \boldsymbol{c}^T(x)\boldsymbol{C}^{*-1}\boldsymbol{u}, \tag{4}$$

$$\tilde{C}(x, x) \leq C(x, x). \tag{5}$$

Note that (4) shows that the predictive process can be treated as a linear projection on the subspace spanned by $\boldsymbol{u}$, and (5) reveals that the predictive process will underestimate the variance of its parent process. A modified predictive process proposed in Finley et al. (2009) can correct the bias of variances by replacing (3) with

$$\tilde{C}(x, x') = \begin{cases} C(x, x') & x = x' \\ \boldsymbol{c}^T(x)C^{*-1}\boldsymbol{c}(x') & x \neq x'. \end{cases} \tag{6}$$

In this paper, we construct our $\mathcal{PWP}$ based on the native $\mathcal{PP}$ rather than the modified version to design a concrete predictive process.

4. The predictive Wishart process

In this section, we first introduce the concept and construction of our proposed $\mathcal{PWP}$, then in the following we discuss its theoretical properties.

4.1. Construction of predictive Wishart process

Suppose that we have $\mathcal{V} \times \mathcal{D}$ independent predictive process functions with an unit variance in its parent process, i.e. $C(x, x) = 1$ for $x \in \mathcal{X}$, as

$$\tilde{u}_{vd}(x) \overset{ind}{\sim} \mathcal{PP}(0, C_d(x, x')), \tag{7}$$

where $v = 1, \ldots, \mathcal{V}$ represents the index of the degrees of freedom $\mathcal{V}$, and $d = 1, \ldots, \mathcal{D}$ is the index of the dimension of the multivariate features. We assume $\mathcal{V} \geq \mathcal{D}$ to ensure our construction is well defined. Here, the objective is to design a collection of positive semi-definite (p.s.d.) random matrices $\Sigma(x)$ (e.g., covariance matrices), indexed by any arbitrary input variable $x \in \mathcal{X}$ (e.g., time). Let $\tilde{\boldsymbol{u}}_v(x) = (\tilde{u}_{v1}(x), \ldots, \tilde{u}_{v\mathcal{D}}(x))^T$, and let $S \in \mathcal{S}^D$ represent a positive definite matrix with its unique lower Cholesky decomposition matrix L such that $LL^T = S$. We also denote $\tilde{U}(x) = (\tilde{\boldsymbol{u}}_1(x), \ldots, \tilde{\boldsymbol{u}}_{\mathcal{V}}(x))$.

Predictive Wishart Process ($\mathcal{PWP}$) is defined as a collection of p.s.d. random matrices $\{\Sigma(x)\}$ indexed by $x \in \mathcal{X}$, by modeling the process as

$$\Sigma(x) = L\tilde{U}(x)\tilde{U}(x)^T L^T = \sum_{v=1}^{\mathcal{V}} L\tilde{\boldsymbol{u}}_v(x)\tilde{\boldsymbol{u}}_v^T(x)L^T, \tag{8}$$

with all latent processes following independent predictive processes. We denote this process as $\mathcal{PWP}(L, \mathcal{V}, \tau)$ that depends on a lower triangular matrix L and a degree of freedom $\mathcal{V}$. The lower triangular matrix L models the marginal variance-covariance at any fixed timestamp and the degrees of freedom $\mathcal{V}$ describes the flexibility of temporal dependence and the hyper-parameters τ characterize latent processes.

If each predictive process of $\tilde{u}_{vd}(x)$ is replaced by its parent process (1), and then this process is formulated as *Generalized Wishart Process* ($\mathcal{GWP}$) (Wilson and Ghahramani, 2010) which is a generalization of the original Wishart process defined by Bru (1991). The *Predictive Inverse Wishart Process* ($\mathcal{PIWP}$), consequently, can be indirectly defined as $\Omega(x) = \Sigma^{-1}(x)$, given $\Sigma(x) \sim \mathcal{PWP}(L, \mathcal{V}, \tau)$. We note that at any index x, the distribution of $\Omega(x)$ is an inverse Wishart distribution.

4.2. Properties of predictive Wishart process

We first show that the proposed $\mathcal{PWP}$ at any input x follows a *well-defined* Wishart distribution $\mathcal{W}_{\mathcal{D}}$ in the theorem below.

Theorem 1. *For any input variable x, the distribution of $\Sigma(x) \sim \mathcal{PWP}(L, \mathcal{V}, \tau)$ at x is the Wishart distribution such that $\Sigma(x) \sim \mathcal{W}_{\mathcal{D}}(\mathcal{V}, S^*)$, where $S^* = LBL^T$ and B is the diagonal matrix with elements $b_d = \tilde{C}_d(x, x)$ for $d = 1, \ldots, \mathcal{D}$.*

Remarks 1. Theorem 1 shows the marginal distribution of $\mathcal{PWP}$ prior at any input x is a well-defined Wishart distribution, and the distribution of $\Sigma(x)$ in $\mathcal{PWP}$ is different from $\mathcal{GWP}$.

Notice that when the predictive process priors are replaced by modified predictive process priors (Banerjee et al., 2008; Finley et al., 2009), the distribution of $\Sigma(x)$ at any input variable x is the Wishart distribution such that $\Sigma(x) \sim \mathcal{W}_D(\mathcal{V}, S)$.

For simplicity, in the remainder of paper, we assume that all latent functions $\tilde{u}_{vd}$ share the same covariance function C. We derive expressions for the covariance between elements of $\Sigma(x)$ and $\Sigma(x')$ for any pair of inputs x and x' in Theorem 2, assuming L is diagonal and $\{\tilde{u}_{vd}\}$ have an identical predictive process prior. Proofs of Theorem 1 and 2 will be given in the Appendix A.

Theorem 2. *Assume that L is a diagonal matrix and $\{\tilde{u}_{vd}\}$ have an independent identical predictive process priors. For any pair of inputs variables x and x', the covariance between $\Sigma_{ij}(x)$ and $\Sigma_{kl}(x')$ is given as*

$$\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) = \begin{cases} 2\mathcal{V}l_i^4\tilde{C}^2(x, x'), & i = j = k = l; \\ \mathcal{V}l_i^2 l_j^2\tilde{C}^2(x, x'), & i = k \neq j = l; \\ 0, & otherwise. \end{cases} \tag{9}$$

Remarks 2. Theorem 2 discusses the temporal cross-relation of dynamic covariance matrices. The covariance turns out to be proportional to the $\tilde{C}^2(x, x')$ and hence shows that the selection of C undoubtedly plays an important role of controlling the autocorrelations. The covariance relation can be generalized to any lower triangular L.

Table 1
A summary of inference approaches for $\mathcal{PWP}$s. Here, $\boldsymbol{w}$, τ, L refer to the inducing variables, input-dependent hyper-parameters and input-independent hyper-parameters, respectively.

Inference [+]	Parameters		
	$\boldsymbol{w}$	τ	L
PWP-MCMC	MCMC	MCMC	MCMC
PWP-VEM	VI	(optimized)	(optimized)

[+] PWP-MCMC: Bayesian inference with Markov chain Monte Carlo (MCMC) on all parameters, PWP-VEM: Variational expectation maximization (VEM) with variational inference (VI) on latent variables $\boldsymbol{w}$ and optimization on remaining parameters.

Remarks 3. Although the priors of $\Sigma(x)$ from $\mathcal{PWP}$ and $\mathcal{GWP}$ both belong to Wishart distribution, they have different scale matrices, $S = LL^T$ for $\mathcal{GWP}$ and $S^* = LBL^T$ for $\mathcal{PWP}$. Because $\tilde{C}_d(x,x) = b_d$ and $C_d(x,x) = 1$, ignoring the subscript d, this similarity between $\mathcal{PWP}$ and $\mathcal{GWP}$ depends on how well $\tilde{C}$ approximates C. Notice that $\tilde{C}$ is the Nyström approximation of C in (3) (Zhang et al., 2008), and the error $\|\tilde{C} - C\|_F$ under the Frobenious norm has an upper bound which is a polynomial function of the square root of the quantization error $\sum_{i=1}^{N} \|x_i - z_{c(i)}\|$ with c coding each input x_i with the closest inducing input z_j. Therefore, the difference of prior of $\Sigma(x)$ from $\mathcal{PWP}$ and $\mathcal{GWP}$ is determined on the displacement of inducing inputs and quantitatively influenced by the quantization error. We suggest the K-mean sampling method for the displacement of inducing inputs, and the sampling approach is used to minimize the quantization error.

5. Hierarchical Gaussian model with $\mathcal{PWP}$

Given a $\mathcal{D}\times N$ dataset $\boldsymbol{Y} = (\boldsymbol{y}(x_1), \cdots, \boldsymbol{y}(x_N))$ with $\mathcal{D}$-dimensional multivariate features indexed by the input variables $x_1, \cdots, x_N$. We consider a conditional Gaussian model with time-varying covariance modeled by $\mathcal{PWP}$ as

$$\begin{aligned} &\boldsymbol{y}_i|\Sigma_i \sim \mathcal{N}(\boldsymbol{0}, \Sigma_i), \\ &\Sigma(x) \sim \mathcal{PWP}(L, \mathcal{V}, \tau), \end{aligned} \tag{10}$$

where $\boldsymbol{y}_i = \boldsymbol{y}(x_i)$ and $\Sigma_i = \Sigma(x_i)$. We propose two inference approaches: 1) Bayesian and 2) Variational inferences. Specifically, Bayesian inference is a Markov Chain Monte Carlo method (MCMC), which accurately provides the samples of posterior distributions. As MCMC can be computationally expensive because it would take long time to converge, we also propose a variational inference which is well suited for large datasets. Moreover, in practice, learning the uncertainty of model parameters L and τ is not of interest and thus we treat them as hyper-parameters to relieve computational burden. Two inference methods are briefly summarized in Table 1 and will be described in details in the following sections respectively.

5.1. Bayesian inference approach

This section discusses a Bayesian inference with $\mathcal{PWP}$. In the context of (10), the objective is to infer the posterior $p(\Sigma(x)|\boldsymbol{y})$ using Gibbs sampling (Geman and Geman, 1984), which is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations in cycles from the conditional distribution of one parameter with the remaining parameters fixed to their current values.

For the sampling, we rewrite (10) as a hierarchical model:

$$\boldsymbol{y}_i|L, \tilde{U}_i \sim \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{0}, L\tilde{U}_i\tilde{U}_i^T L^T), \tag{11}$$

$$\tilde{\boldsymbol{u}}_{vd} = \boldsymbol{c}^T \boldsymbol{C}^{*-1} \boldsymbol{w}_{vd}, \tag{12}$$

$$\boldsymbol{w}_{vd}|\tau \sim \mathcal{N}(\boldsymbol{w}_{vd}|\boldsymbol{0}, C^*), \tag{13}$$

where $\tilde{U}_i = \tilde{U}(x_i)$, $\tilde{\boldsymbol{u}}_{vd} = (\tilde{u}_{vd}(x_1), \ldots, \tilde{u}_{vd}(x_N))^T$, $\boldsymbol{w}_{vd} = (u_{vd}(z_1), \ldots, u_{vd}(z_M))^T$. Here $u_{vd}(x)$ refers to the function of the parent process with respect to $\tilde{u}_{vd}(x)$. On the other hand, C^* refers to covariance between $\{z_i\}_{i=1}^M$ and $\boldsymbol{c}$ is cross covariance between $\{x_i\}_{i=1}^N$ and $\{z_i\}_{i=1}^M$.

As for the prior specification, we set prior of hyper-parameters of GPs $\tau \sim \pi(\tau)$ and the prior of the lower triangular matrix $L \sim \pi(L)$. The prior of τ is chosen based on the choice of covariance function C. In the experiments, we consider two types of covariance functions, one for periodic covariance function and the other for square exponential function. We put a flat normal distribution as a prior of the log of lengthscale parameters. And for $\pi(L)$, we put independent standard Gaussian priors for the entries on or below the diagonal of L. We then design a Gibbs sampling procedure as

$$p(\boldsymbol{w}|\boldsymbol{Y}, \tau, L) \propto p(\boldsymbol{Y}|\boldsymbol{w}, L, \tau)p(\boldsymbol{w}|\tau), \tag{14}$$

$$p(\tau|\boldsymbol{Y}, \boldsymbol{w}, L) \propto p(\boldsymbol{Y}|\boldsymbol{w}, L, \tau)p(\boldsymbol{w}|\tau)\pi(\tau), \tag{15}$$

$$p(L|\boldsymbol{Y}, \boldsymbol{w}, \tau) \propto p(\boldsymbol{Y}|\boldsymbol{w}, L, \tau)\pi(L), \tag{16}$$

where $\boldsymbol{w}$ represent the vector of functions evaluated from the inducing points, τ denote the input-dependent hyper-parameters in $\mathcal{PWP}$ and they are also the hyper-parameters in the covariance function C, and L denote the input-independent hyper-parameters in $\mathcal{PWP}$. Furthermore, we present the details of parameter initialization, posterior sampling and inducing point selection regarding the MCMC implementation for the Bayesian inference approach.

5.1.1. Parameter initialization

According to Theorem 1 that $\Sigma(x) \sim \mathcal{W}(\mathcal{V}, S^*)$, the prior expectation of covariance matrix $\Sigma(x)$ equals $\mathcal{V}S^*$. In the initialization step, we assume that $\Sigma(x_1), \ldots, \Sigma(x_N)$ are independent, then the covariance matrix $\Sigma(x)$ has an unbiased estimate $\hat{\Sigma}(x) = \frac{1}{N-1}\sum_{i=1}^{N} \boldsymbol{y}_i \boldsymbol{y}_i^T$. Consequently, $\hat{L}$ can be estimated by the Cholesky decomposition of $\frac{\hat{\Sigma}}{\mathcal{V}}$ by assuming that S^* is close to S. Then following (10), we estimate the $\boldsymbol{w}$ and τ by maximizing the log likelihood of $\boldsymbol{Y}$ given $\hat{L}$.

5.1.2. Details on posterior sampling

We first sample the $\tilde{\boldsymbol{u}}_{vd}$ from its posterior distribution via indirect sampling of $\boldsymbol{w}_{vd}$ using (14). Given the property of predictive process (4) and $\boldsymbol{w}_{vd}$, $\tilde{\boldsymbol{u}}_{vd}$ is generated via $\tilde{\boldsymbol{u}}_{vd} = \boldsymbol{c}C^{*-1}\boldsymbol{w}_{vd}$. As for the sampling of $\boldsymbol{w}$, we employ the Elliptical Slice sampling and this sampling procedure requires to computing the posterior of $\boldsymbol{w}$, taking $\mathcal{O}(M^2N)$ time complexity where N is the number of observations and M is the number of inducing points. Therefore the sampling complexity is linear to the number of observations N. In contrast to $\mathcal{GWP}$ in which sampling the latent function from the posterior would take $\mathcal{O}(N^3)$ time complexity, $\mathcal{PWP}$ is much more efficient, especially when the number of inducing points M is significantly smaller N, i.e., $M \ll N$. Then, we sample τ using (15) and sample L using (16). Since the posterior of τ and L do not have a closed-form expression, we leverage Metropolis Hastings for sampling.

5.1.3. Inducing points selection

For selecting the inducing points, we take equal-spaced points $\{z_i\}_{i=1}^{M}$ over the whole input space $\mathcal{X}$ to ensure the better prediction performance over the whole input space. These z are fed in (2) that leads to the definition the $\mathcal{PWP}$. While Bayesian inference yields the true posterior for better estimation of covariance, it is often intractable due to slow convergence with exhaustive sampling. We therefore propose an efficient variational inference in the following.

5.2. Variational expectation maximization

Variational inference provides an alternative efficient inference approach at the price of precision of the posterior approximation. It is a Bayesian technique of approximating the posterior which has emerged as an important tool (Jordan et al., 1999; Blei et al., 2017). We consider the same hierarchical model from ((11), (12) and (13)), and L and τ are treated as hyper-parameters as opposed to Bayesian inference. This is because learning the posterior distribution of those hyper-parameters is not of interest in practice, and it would save computation in training.

Given above specifications, the evidence lower bound (ELBO), a lower bound of the log marginal likelihood is derived with Shannon entropy H as

$$\log p(\boldsymbol{Y}) \geq \mathrm{E}_{q(\boldsymbol{w})}[\log p(\boldsymbol{Y}, \boldsymbol{w})] + H(q(\boldsymbol{w})) = \mathrm{ELBO}, \tag{17}$$

where $q(w)$ is a variational distribution of w.

We assume $q(\boldsymbol{w})$ belongs to normal distribution. Instead of directly maximizing the ELBO (17) with respect to $q(\boldsymbol{w})$ and (L, τ) via stochastic gradient descend, we iteratively and conditionally update $q(\boldsymbol{w})$ and (L, τ) until they converge. It is called variational expectation maximization (VEM) inference (Bernardo et al., 2003). Specifically, given (L, τ), maximizing the ELBO (17) is equivalent to minimizing the Kullback-Leibler divergence between the variational distribution $q(\boldsymbol{w})$ and the posterior distribution $p(\boldsymbol{w}|\boldsymbol{y})$. Due to the Gaussian assumption in $q(\boldsymbol{w})$, we approximately update $q(\boldsymbol{w})$ via the Laplace approximation (Bishop, 2006) $q^*(\boldsymbol{w})$. On the other hand, given a $q(\boldsymbol{w})$, (L, τ) are updated by

$$\begin{aligned} L^*, \tau^* &= \arg\max_{L,\tau} \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{w})}[\log \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{0}, L\tilde{U}_i\tilde{U}_i^T L^T)] + R \\ &= \arg\max_{L,\tau} \sum_{i=1}^{N} [\log \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{0}, L\langle\tilde{U}_i\rangle\langle\tilde{U}_i\rangle^T L^T)] + R \\ &= \arg\max_{L,\tau} \sum_{i=1}^{N} \mathcal{L}_i + R, \end{aligned} \tag{18}$$

where both regularization term, the KL divergence between $q(\boldsymbol{w})$ and $p(\boldsymbol{w})$, $R = \mathrm{KL}(q(\boldsymbol{w})\|p(\boldsymbol{w}))$ and latent variables $\tilde{U}_i$ depend on τ, and $\langle\cdot\rangle = \mathrm{E}_{q(\boldsymbol{w})}[\cdot]$. We iteratively update $q(w)$ and (L, τ) until they converge.

Algorithm 1: Variational expectation maximization algorithm for multitask learning.

Input : Observations $\boldsymbol{Y}$, Hyper-parameters of covariance functions $\boldsymbol{\tau}$;
Output: Variational distribution $q(\boldsymbol{w})$, Task-specified features $\{L_i\}_{i=1}^N$;
do
  Fix all task-specified features $\{L_i\}_{i=1}^N$ and update the variational distribution $q(\boldsymbol{w})$ by the Laplace approximation on $p(\boldsymbol{w}|\boldsymbol{Y}, \{L_i\}_{i=1}^N)$;
  for $i \leftarrow 1$ **to** N **do**
    Fix the variational distribution $q(\boldsymbol{w})$ and update the L_i by maximizing the term in ELBO that is only related to L_i:

$$L_i^* = \arg\max_{L_i} \sum_{j=1}^{N_i} [\log \mathcal{N}(\boldsymbol{y}_{i,j}|\boldsymbol{0}, L_i \langle \tilde{U}_{ij} \rangle \langle \tilde{U}_{ij} \rangle^T L_i^T)], \tag{23}$$

    where $\langle \cdot \rangle = E_{q(\boldsymbol{w})}[\cdot]$;
  end
while *Both* $q(\boldsymbol{w})$ *and* $\{L_i\}_{i=1}^N$ *converge*;

5.3. Prediction of covariance at new timestamp

For both Bayesian and variational EM inferences, given a new time stamp x^*, we extract posterior samples $\{\boldsymbol{w}^{(s)}, \tau^{(s)}, L^{(s)}\}_{s=1}^S$ from MCMC or variational distributions, then we sample the corresponding $\tilde{u}_{vd}^* = \tilde{u}_{vd}(x^*)$ using

$$\tilde{u}_{vd}^{*(s)} = \boldsymbol{c}^{*T} C^{*-1} \boldsymbol{w}_{vd}^{(s)}, \tag{19}$$

where $\boldsymbol{c}^*$ denotes the vector of covariance functions evaluated between the new time stamp x^* and inducing inputs $\{z_i\}_{i=1}^M$, i.e. $\boldsymbol{c}(x^*)$, and $\boldsymbol{w}_{vd}^{(s)}$ represents the s^{th} posterior sample. Consequently, according to the construction (8), we obtain the posterior predictive samples of $\Sigma^* = \Sigma(x^*)$ by

$$\Sigma^{*(s)} = \sum_{v=1}^{\mathcal{V}} L^{(s)} \tilde{u}_{vi}^{*(s)} \tilde{u}_{vj}^{*(s)} L^{(s)T}. \tag{20}$$

At last, we estimate Σ^* using the posterior predictive mean of the samples $\{\Sigma^{*(s)}\}_{s=1}^S$.

6. Multi-task learning with $\mathcal{PWP}$

In this section, we consider a scenario of feature selection for multiple tasks, where each task is assigned with unique features. Assume that we have N tasks in which the i^{th} task consists of a multivariate time series with length N_i, i.e. $\boldsymbol{Y}_i = \{\boldsymbol{y}_{i,j}\}_{j=1}^{N_i}$. The corresponding time stamps are denoted as $\boldsymbol{x}_i = \{x_{i,j}\}_{j=1}^{N_i}$ and each observation $\boldsymbol{y}_{i,j} \in \mathbb{R}^M$ is assigned to the time stamp $x_{i,j}$. A hierarchical model is formulated as

$$\begin{aligned} &\boldsymbol{y}_{i,j}|\Sigma_{i,j} \sim \mathcal{N}(0, \Sigma_{i,j}), \\ &\Sigma_i(x) \sim \mathcal{PWP}(L_i, \mathcal{V}, \tau), \end{aligned} \tag{21}$$

where $\Sigma_{i,j} = \Sigma_i(x_{i,j})$. We assume that the model of $\Sigma_i(\cdot)$ shares the same degree of freedom $\mathcal{V}$ and the same hyper-parameters in GPs τ, but has individual effect modeled by the task-specified lower triangular matrix L_i for the ith task. This specification suggests that covariances across tasks share the same latent temporal process prior, and covariances within each task share a task-specified correlation structure modeled by the lower triangular matrix L_i. Thus, we take the L_i as a feature for task i which directly refers to task-specific feature.

To find out task-specific features, we estimate $L_i(\tau)$ for each task i under different settings of τ where τ can be treated as different scale and $L_i(\tau)$ is the feature at the scale τ. Because in the multi-task learning context, extracting task-specified feature is of interest and thus we treat $L_i(\tau)$ as model parameters. Specifically, we consider a square exponential covariance function in $\mathcal{PWP}$ where τ is the length scale parameter, and we define a $\mathcal{PWP}$ Multi-scale Descriptor ($\mathcal{PWPMD}$) as

$$\mathcal{PWPMD}_\tau(i) = \{L_i^*; L_i^*, q^*(\boldsymbol{w}) = \arg\max_{L_i, q(\boldsymbol{w})} (\text{ELBO}|\tau)\}. \tag{22}$$

Here, under each setting of τ, L_i^* becomes a feature for the i^{th} task. It has the same size of the feature of each task regardless of the number of observations N_i, and can be used for downstream prediction tasks. To infer the multi-scale descriptor, we propose a variational EM algorithm and describe it in Algorithm 1.

Table 2
Parameter posterior credible intervals 50 (2.5, 97.5), RMSE of the reconstruction for Σs, NLML with mean (standard deviation) and corresponding average inference time for 100 iterations.

	True	$\mathcal{GWP}$	$\mathcal{PWP}(\mathcal{B})_{20}$	$\mathcal{PWP}(\mathcal{B})_{50}$	$\mathcal{PWP}(\mathcal{B})_{100}$	$\mathcal{PWP}(\mathcal{VI})_{20}$	$\mathcal{PWP}(\mathcal{VI})_{50}$	$\mathcal{PWP}(\mathcal{VI})_{100}$	$\mathcal{DCC}$
L_{00}	1	1.12(0.98,1.28)	1.36(0.97.1.68)	1.17(1.06,1.43)	**0.99**(0.84, 1.08)	1.63	1.55	1.57	-
L_{01}	0	**-0.02**(-0.04,0.07)	-0.06(-0.19, -0.01)	0.04(-0.03, 0.15)	**0.02**(-0.02,0.08)	0.33	0.31	0.31	-
L_{11}	1	1.04(0.92,1.11)	1.05(0.87, 1.21)	1.12(1.02,1.26)	**1.02**(0.78,1.46)	1.16	1.10	1.10	-
RMSE (Σ_{00})	-	1.15	1.23	1.10	**0.55**	0.84	0.81	0.95	2.71
RMSE (Σ_{01})	-	**0.48**	0.95	0.74	0.90	0.67	0.65	0.70	1.67
RMSE (Σ_{11})	-	0.53	**0.46**	0.61	0.49	0.95	0.85	0.88	1.50
NLML	-	1098.94(2.88)	1105.88(6.69)	1096.82(3.37)	1105.27(4.19)	1082.05	1083.90	**1081.53**	-
Time (sec)	-	50.24	**25.39**	35.97	43.81	-	-	-	-

Subscript indicates the number of inducing points used in each model. (B) refers to Bayesian inference and (VI) refers to Variation inference. For all Bayesian inference, we have informative initialization on all latent variables based on the true values. We also provide the ground true parameters L.

7. Simulation study

In this section, we performed an experiment on the synthetic multivariate time-series data which were generated based on ground truth covariance matrices Σs to validate both covariance reconstruction and predictive performance of $\mathcal{PWP}$.

7.1. Experimental setup

Synthetic Data Generation. We generated multivariate time series data using the $\mathcal{GWP}$ model with a periodic covariance function for all $\{u_{vd}(x)\}$ such that $k(x,x') = \sigma^2 e^{-2\sin(\pi*(x-x')/p)^2}$, with a scale parameter σ and a period parameter p. Specifically, $N = 350$, $\mathcal{D} = 2$ and $\mathcal{V} = 3$, L was chosen as an identity matrix and hyper-parameters were set as $\sigma = 1, p = 100$, assuming that the period of the time series is 100. The first 300 data points were used for training and the following 50 samples were used for testing.
Baselines. Most recent methods such as $\mathcal{GWP}$ and zero-mean multivariate GARCH models, i.e., Dynamic Conditional Correlation ($\mathcal{DCC}$) (Orskaug, 2009), were chosen as the baseline methods.
Setup. For $\mathcal{PWP}$, different number of inducing points (i.e., $M = 20, 50$, and 100) with the same type of periodic covariance function were investigated. We implemented both Bayesian inference and variational EM inference for $\mathcal{PWP}$. We fixed the hyper-parameter $p = 100$ since that is difficult to learn.

For Bayesian inference, we initialized L at the values near the true values in $\mathcal{GWP}$, latent variables $\boldsymbol{w}$ at the estimates via the inverse of (12) with the true $\tilde{U}$. This yields informative initialization to identify the property of the global optima in $\mathcal{GWP}$ and $\mathcal{PWP}$ for inferences. During the Bayesian inference of $\mathcal{PWP}$, we used 5000 samples whose first 2500 samples were burned-in. For variational EM inference, L and $\boldsymbol{w}$ were randomly initialized.
Evaluation Metric. In Table 2, we displayed the root mean square error (RMSE) of parameters for L as the evaluation of inference. We displayed the RMSE between true variance-covariance matrices and corresponding reconstruction as the evaluation of covariance reconstruction. Moreover we also provided the negative log marginal likelihood (NLML) to evaluate the model fitting.

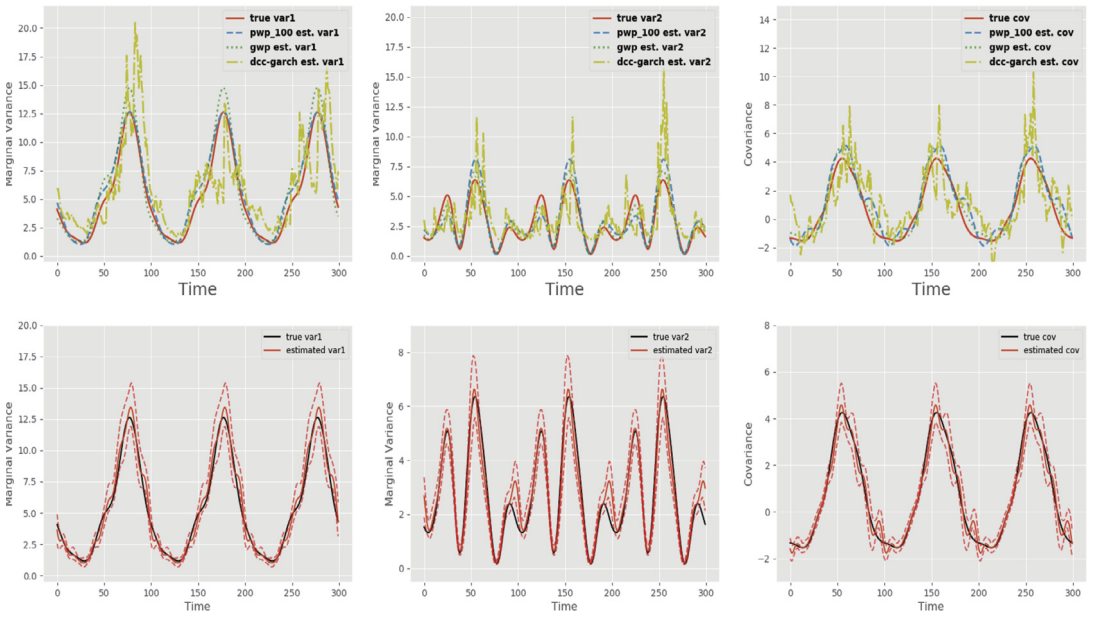
In Table 3, We showed the predictive performance of $\mathcal{PWP}$ with i-step ahead forecast, where observations until the last timestamp x in training data are considered to predict $\Sigma(x+i)$ and $i = 1, \ldots, 50$.

7.2. Results and discussions

Parameter estimation and model fitting results in Table 2 illustrate that $\mathcal{PWP}$ has a significantly better covariance matrix estimation performance than the $\mathcal{DCC}$ model due to the notably smaller RMSE. Comparing with the $\mathcal{GWP}$, with a suitable number of inducing points, $\mathcal{PWP}$ has a competitive result for both parameter estimation and covariance matrix estimation. As for the computational benefits, the computation time of $\mathcal{PWP}$ is significantly lowered compared with $\mathcal{GWP}$ in the same Bayesian setting.

As for the predictive performance, we conducted Bayesian inference for $\mathcal{PWP}$ as a fair comparison with the Bayesian inference in $\mathcal{GWP}$. We reported the RMSEs of predicted covariance matrices and true covariance matrices for $\mathcal{GWP}$, $\mathcal{PWP}$ with 20, 50 and 100 inducing points and $\mathcal{DCC}$ models in Table 3. The averaged RMSEs over all entries for the five models are 0.53, 0.52, 0.70, 0.70 and 2.53. It shows that $\mathcal{PWP}$ has a comparable performance compared with $\mathcal{GWP}$ and significantly outperforms the $\mathcal{DCC}$. Moreover, we visualized the ground truth for Σs and the reconstruction of Σs in $\mathcal{PWP}_{100}$ in Fig. 2 and showed the uncertainty quantification of covariance matrices in $\mathcal{PWP}$, illustrating that $\mathcal{PWP}$ achieves a great uncertainty quantification in the sense that the confident intervals cover almost the true values with a narrow band-width.

With respect to the computational benefits, we find that as the number of inducing points decrease the computation time would be significantly shorter than that from $\mathcal{GWP}$. It matches the theoretical analysis of the computational complexity which is linear to the number of observations N in contrast to the $\mathcal{O}(N^3)$ in $\mathcal{GWP}$.



(a) Marginal Var. of y_1 (b) Marginal Var. of y_2 (c) Covariance

Fig. 2. Top: Reconstruction of Σs; Bottom: 95% confident intervals (shown in red dashed lines) in the reconstruction with $\mathcal{PWP}_{100}$, (a) the marginal variances at the first dimension (1st diagonal element of Σs), (b) the marginal variances at the second dimension (2nd diagonal element of Σs), (c) the covariances (symmetric off-diagonal element of Σ)s. Our proposed $\mathcal{PWP}$ delivers smoother estimations compared with $\mathcal{DCC}$ and also provides a comparable fitting performance compared to $\mathcal{GWP}$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 3
RMSE between predicted $\hat{\Sigma}^*$ and true Σ^* element-wisely for the next 50 timestamps. $\mathcal{PWP}$ has a comparable performance with $\mathcal{GWP}$ even with much less inducing points.

Model$^+$	Variance 1	Variance2	Covariance
$\mathcal{GWP}$	0.72	**0.49**	0.45
$\mathcal{PWP}_{20}$	**0.50**	0.84	**0.36**
$\mathcal{PWP}_{50}$	0.93	0.70	0.58
$\mathcal{PWP}_{100}$	0.75	0.95	0.55
$\mathcal{DCC}$	5.10	2.19	1.41

$^+$ Subscript indicates the number of inducing points used for $\mathcal{PWP}$. For $\mathcal{GWP}$ and $\mathcal{PWP}$, Bayesian inference and informative initialization on all latent variables based on the true values were used.

The results in Table 2 show that: For Bayesian inference, as the number of inducing inputs (M) increases, the parameter estimates of L become closer to the true values. However, the performance of covariance reconstruction and data fitting does not always improve as M increases in our setting. This may be caused by the efficiency of sampling the inducing variables $\boldsymbol{w}$. Even with an efficient elliptical slice sampling, as the size of $\boldsymbol{w}$ increases, the sampling step suffers from the slow mixing of sampling and cause undesirable fitting performance. It demonstrates that $\mathcal{PWP}$ becomes more expressive with more inducing points but fitting becomes more difficult, which emphasizes that the importance of the selection of inducing points.

On the other hand, $\mathcal{PWP}$ has a comparable prediction performance with $\mathcal{GWP}$ even with less inducing points. This may be because the learning of Gaussian processes in $\mathcal{GWP}$ is affected by over-fitting, while the learning of predictive processes in $\mathcal{PWP}$ resists this issue.

As for the variational EM inference of $\mathcal{PWP}$, it would provide biased estimates on L but we find that those estimates are consistently robust under different settings of the inducing points. Beside that, the variational EM inference provides comparable performance on both covariance reconstruction and model fitting.

7.3. PWP inference in high dimensional time series

In this section, we explored the model performance of $\mathcal{PWP}$ in high dimensional time series. In particular, we kept the original settings except for dimension size $\mathcal{D}$, degree of freedom $\mathcal{V}$ and hyperparameter σ^2, and investigated the

Table 4
RMSE of the reconstruction across all elements in Σs and RMSE of L.

Metric	$\mathcal{D}$	$\mathcal{PWP}(\mathcal{B})_{20}$	$\mathcal{PWP}(\mathcal{B})_{50}$	$\mathcal{PWP}(\mathcal{B})_{100}$	$\mathcal{PWP}(\mathcal{VI})_{20}$	$\mathcal{PWP}(\mathcal{VI})_{50}$	$\mathcal{PWP}(\mathcal{VI})_{100}$
RMSE(L)	2	0.15	0.10	0.02	0.38	0.37	0.42
	5	0.30	0.33	0.34	0.26	0.26	0.26
	10	0.48	0.32	0.37	0.30	0.31	0.32
	20	0.47	0.48	0.51	0.29	0.35	0.36
RMSE(Σ)	2	0.25	0.39	0.21	0.25	0.25	0.27
	5	0.21	0.25	0.19	0.13	0.13	0.13
	10	0.39	0.27	0.30	0.16	0.17	0.17
	20	0.67	0.73	0.75	0.15	0.20	0.21

Subscript indicates the number of inducing points used for $\mathcal{PWP}$.

model behavior. Specifically, we considered the dimension size $\mathcal{D} = 2, 5, 10, 20$ with degree of freedom $\mathcal{V} = 3, 6, 11, 21$ and hyperparameter $\sigma^2 = \frac{1}{3}, \frac{1}{5}, \frac{1}{11}, \frac{1}{21}$ respectively. In this case, the marginal distribution at each time sample would follow $\Sigma(x) \sim \mathcal{W}(\mathcal{V}, \frac{1}{\mathcal{V}} I)$, implying that the expectation of covariance matrix should be an identity matrix, i.e., $E[\Sigma(x)] = I$. Such a setting should make the model comparison fair since the generated data are under an unit scale. We then conducted both Bayesian inference and variational EM inference, and reported the root mean square error (RMSE) across each element in the lower triangular matrix L denoted as RMSE(L) and the RMSE across all covariance matrices and all elements Σ_{ij} denoted as RMSE(Σ). The evaluation metrics are given in Table 4.

The result shows that the Bayesian inference performs worse as the dimension size increases in terms of both parameter estimation RMSE(L) and the reconstruction RMSE(Σ). Also, as dimension size $\mathcal{D}$ increases, the larger number of inducing points does not significantly improve the parameter estimation performance. It may be because the inference is more difficult for high dimensional cases. On the other hand, in the case of $\mathcal{D} = 2$, variational inference performs worse than Bayesian inference in terms of parameter estimation RMSE(L), while as $\mathcal{D}$ increases variational inference performs better than Bayesian inference. It suggests that in our model, the variational inference would be preferred for high dimensional data, since optimization in variational inference would be more robust for sampling in MCMC for high dimensional data.

8. Analysis of dynamic brain connectivity

We performed two experiments on dynamic functional brain connectivity using real brain imaging data to confirm the practicality of $\mathcal{PWP}$. As $\mathcal{GWP}$ was not scalable for the real data, we compared $\mathcal{PWP}$ with $\mathcal{DCC}$-GARCH models for the individual analysis of dynamic functional connectivity. Then, we performed a multi-task learning task on multiple rs-fMRI timeseries via variational EM algorithm to identify associations between functional connectivity and behavioral scores.

8.1. Experimental setup

Human Connectome Data. The pre-processed resting-state functional MRI (rs-fMRI) data used in this experiment were obtained from the Human Connectome Project (HCP) S1200 data release (Smith et al., 2013) for 812 subjects whose fMRI data were complete and reconstructed using the improved $r227$ recon algorithm. Timeseries data were generated through the HCP preprocessing pipeline (WU-Minn, 2017) which yielded one representative timeseries across 4800 timestamps per independent component analysis (ICA) component for each subject at several different dimensionalities. Specifically, we used the rs-fMRI timeseries from 15 ICA components with a length of 4800.

Setup. We took the whole 4800 observations to estimate covariance matrices and computed the log likelihood at each timestamp. For $\mathcal{PWP}$, we selected 50 inducing points uniformly located in the whole time interval. Squared exponential covariance function was employed here to model the dynamics of covariance matrix of HCP data. We considered a weakly informative prior on the length scale parameter $\log \tau \sim \mathcal{N}(0, 10^2)$ and a data-driven prior on L, $L_{ij} \sim \mathcal{N}(0, 20^2)$ for $i \geq j$. On our server machine with 128G RAM (which is not small), $\mathcal{GWP}$ model failed to run on the HCP dataset due to its lack in scalability. Therefore, we compared our results with four parametric $\mathcal{DCC}$-GARCH models. Three of them employ a autoregression-moving-average model with order (1,1) for the mean but leverage different types of noise following multivariate Normal ($\mathcal{MVN}$), multivariate Student-t ($\mathcal{MVT}$) and multivariate Laplace distributions ($\mathcal{MVL}$). The last $\mathcal{DCC}$-GARCH model sets zero mean and has noise following multivariate Normal distributions ($\mathcal{MVN}0$).

Since the Markov chain Monte Carlo would yield less biased result than variational EM algorithm as shown in Table 3, to compare the performance with other models, we conducted the Markov chain Monte Carlo inference and estimated model parameters using the maximum a posteriori. Moreover, given those estimates, we reconstructed covariance matrices on the observed timestamps.

8.2. Individual functional connectivity construction

We randomly selected one participant (ID: 990366) for the demonstration of individual dynamic functional connectivity derivation. The log-likelihood of observation (i.e., ICA) at each timestamp was computed and plotted as a boxplot for

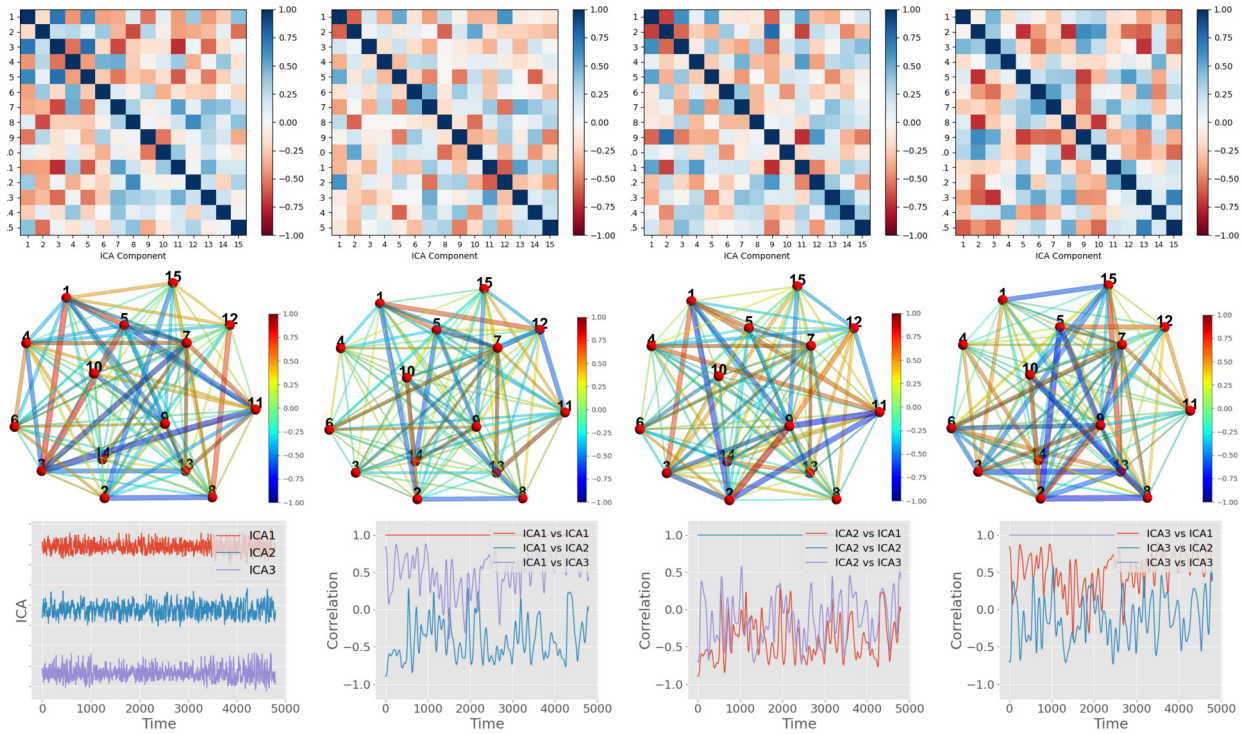


Fig. 3. Dynamic correlations between ICA components (i.e., dynamic functional connectivity) and corresponding network representations derived from the estimations of $\Sigma(x)$ at $x = 1001, 2001, 3001, 4800$ with HCP timeseries data. Top row: connectivity matrices; Middle row: corresponding network representations (thicker edge represents larger absolute edge values and the colormap renders the value of the edge from low to high); Bottom row: three true ICA components and corresponding inferred dynamic correlation processes. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

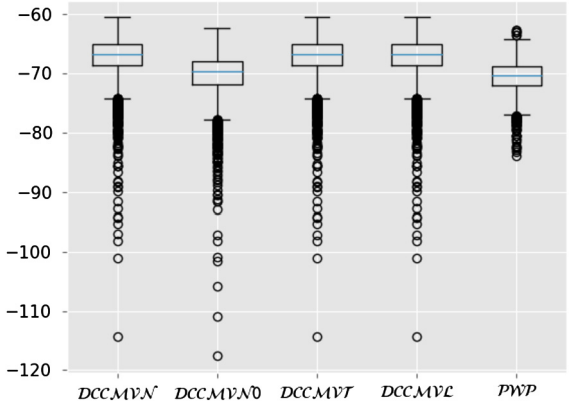


Fig. 4. Boxplots of log-Likelihood w.r.t. the whole 4800 timestamps (i.e., time) with the reconstructed covariance matrix. $\mathcal{PWP}$ shows better stability than $\mathcal{DCC}$s with less extreme outliers and lower variance.

all observations in Fig. 4. We also plotted the same boxplots of log-likelihoods estimated from $\mathcal{DCC}$ models. $\mathcal{PWP}$ and $\mathcal{DCC}\mathcal{MVN}0$ assume zero mean, which makes them comparable. The figure shows that $\mathcal{PWP}$ performs relatively worse than $\mathcal{DCC}$ models in terms of the mean of log-likelihood, but it provides more stable results than $\mathcal{DCC}$ models in the sense of less extreme outliers and lower variance.

In Fig. 3, we presented dynamic correlation matrices and the structural networks derived from the estimated $\Sigma(x)$ at timestamp $x = 1001, 2001, 3001, 4800$ to show the changes of their functional brain connectivity across time. This result proves the hypothesis in Hutchison et al. (2013) that the structure of covariance along time in functional connectivity may be significant, and shows a significant potential that our $\mathcal{PWP}$ is a very powerful tool to visualize the estimate of covariance in time-varying data. Moreover, to directly illustrate the temporal relation, we provided the plot of three ICA components as well as their corresponding inferred correlation processes in Fig. 3. It illustrates that the correlations between ICA components are not random and they have certain patterns.

Table 5
R^2 scores of linear model fitting with different features for different exogenous variables.

Feature	Linear Regression				
	MMSE	PSQI	PainIntens	PainInterf	Mars
Baseline features	0.21	0.19	0.16	0.18	0.19
$\mathcal{PWPMD}$	**0.48**	**0.50**	**0.45**	**0.48**	**0.58**

MMSE: Mini Mental Status Exam; PSQI: Pittsburgh Sleep Questionnaire; PainIntens: Pain Intensity Raw Score; PainInterf: Pain Interference T-score; Mars: Mars Contrast Sensitivity Test.

8.3. Multi-task learning on HCP data

In order to show the applicability of our dynamic brain connectivity features, we compared the fitting performances of $\mathcal{PWPMD}$ against baseline features.

Here we utilized all 812 subjects in the multi-task learning experiment, and considered a three-level multi-scale descriptor from (22) where the length scale parameter τ in the squared exponential covariance function is set to $500, 2000$ and 5000. We used the matrix from Cholesky decomposition of the sample covariance matrix as the baseline features for each subject as conducted in Van Den Heuvel and Pol (2010); Biswal (2012); Leonardi et al. (2013). Then we conducted the linear regression between the features extracted from the rs-fMRI timeseries and exogenous variables.

We considered five behavioral scores available in the HCP dataset as exogenous variables: MMSE, PSQI, PainIntens (raw), PainInterf T-score and Mars Log Scores. Specifically, Mini Mental Status Exam (MMSE) (Folstein et al., 1975; Crum et al., 1993) is a broad measure of cognitive status, Pittsburgh Sleep Questionnaire (PSQI) (Buysse et al., 1989) is a measure of sleep quality, Pain Intensity Raw Score (PainIntens) (Gershon et al., 2013) consists of a single item measuring immediate (i.e., acute) pain in adults, Pain Interference T-score (PainInterf) (Gershon et al., 2013) measures the degree to which pain interferes with other activities in life in adults, and Mars Contrast Sensitivity Test (Mars) (Arditi, 2005; Dougherty et al., 2005; Haymes et al., 2006) is a brief and reliable measure that assesses color contrast sensitivity.

The resulting R^2 scores from linear model fitting are reported in Table 5. It is apparent that the $\mathcal{PWPMD}$ achieves the best fitting performance across all five HCP behavioral measures. Notably, the $\mathcal{PWPMD}$ exhibits better performance by 39% when compared with the baseline feature on behavioral measurement Mars Log Score, and also outperforms the baseline by 27%, 31%, 29%, 30% on MMSE score, PSQI score, PainIntens raw score, PainInterf T-score, respectively. Our experiments illustrate that our proposed dynamic brain connectivity features $\mathcal{PWPMD}$ significantly improve the regression performance as compared with the baseline features. The promising results from these experiments on HCP dataset implicate a great potential for our $\mathcal{PWP}$ for multi-task learning in real-world clinical applications.

9. Conclusion

There is a significant interest in modeling time-varying changes of relationships between different variables in both theoretical and application-wise perspectives. As previous stochastic approaches heavily suffer from computational burden, we introduced a novel stochastic process, i.e., $\mathcal{PWP}$, which can model dynamic covariance matrices accurately and efficiently. Not only we provide theoretical guarantee that it is a well defined process, but also illustrate that it is easy to be incorporated into different models such as hierarchical Gaussian model and multi-task model. Moreover, we empirically evaluate our ideas and its usefulness with two independent sets of experiments. Especially for the real experiment on HCP data, features derived from dynamic functional connectivity can be useful for multi-task learning over traditional approaches extracting features from covariance matrices. We believe there is a significant potential that $\mathcal{PWP}$ can be further utilized in various areas where time-varying associations between variables need to effectively characterized.

Although $\mathcal{PWP}$ can handle considerably long time series, it does not necessarily emphasize approximation and inference for high dimensional data. As we noted in Section 7.3, the inference would be more difficult when the dimension of channels for the time-series increases. One would need to leverage factor analysis (Cunningham and Yu, 2014; Meng and Bouchard, 2021) or introduce the sparsity via shrinkage priors (Huber and Feldkircher, 2019) for covariance matrix modeling to make $\mathcal{PWP}$ make suitable for high dimensional data. Distributed learning for PWP may be another feasible approach as the data sizes of various recent datasets are continuously increasing, but it is beyond the scope of current work and remains as a future direction to consider.

Acknowledgement

Part of this research was carried when Fan Yang and Won Hwa Kim were at the University of Texas at Arlington. This research was supported by NSF IIS CRII 1948510, and partially supported by grants from South Korea funded by Ministry of Science and ICT (MSIT) including IITP-2019-0-01906 (AI Graduate Program at POSTECH, 10%), IITP-2022-2020-0-01461 (ITRC, 10%), IITP-2022-0-00290 (Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense, 10%) from the Institute of Information & communications Technology Planning & Evaluation, Ministry

of Health & Welfare including HU22C0168 (10%) and HU22C0171 (10%) from Korea Health Industry Development Institute (KHIDI), and National Research Foundation (NRF) NRF-2022R1A2C2092336 (50%).

Appendix A. Theorem proving

A.1. Proof of Theorem 1

Here we present the proofs of Theorem 1 as below.

Proof. In the construction of $\mathcal{PWP}$, $\{\tilde{u}_{vd}\}$ have independent predictive process priors. Therefore, we have

$$\tilde{\boldsymbol{u}}_v(x) = (\tilde{u}_{v1}(x), \ldots, \tilde{u}_{v\mathcal{D}}(x))^T \sim \mathcal{N}_{\mathcal{D}}(\mathbf{0}, B), \tag{A.1}$$

where B is the diagonal matrix with elements $b_d = \tilde{C}_d(x, x)$. Because of $C_d(x, x) = 1$ and the property (5), $b_d \leq 1$ for $d = 1, \ldots, \mathcal{D}$. According to the property of multivariate Gaussian distribution, it immediately follows that

$$L\tilde{\boldsymbol{u}}_v(x) \sim \mathcal{N}_{\mathcal{D}}(\mathbf{0}, S^*), \tag{A.2}$$

where $S^* = LBL^T$. Due to (A.3) and according to the definition of Wishart distribution, we have $\Sigma(x) \sim \mathcal{W}_{\mathcal{D}}(\mathcal{V}, S^*)$. Since $\mathcal{V} \geq \mathcal{D}$ in the construction, this Wishart distribution is well defined. □

A.2. Proof of Theorem 2

Here we present the proofs of Theorem 2 as below.

Proof. We denote the diagonal elements of L as $(l_1, \ldots, l_D)$, then according to

$$\begin{aligned}\Sigma(x) &= L\tilde{U}(x)\tilde{U}(x)^T L^T \\ &= \sum_{v=1}^{\mathcal{V}} L\tilde{\boldsymbol{u}}_v(x)\tilde{\boldsymbol{u}}_v^T(x)L^T,\end{aligned} \tag{A.3}$$

the $(i, j)^{\text{th}}$ element of the covariance $\Sigma(x)$ is given as

$$\Sigma_{ij}(x) = \sum_{v=1}^{V} l_i \tilde{u}_{vi}\tilde{u}_{vj} l_j\,. \tag{A.4}$$

According to (7), we let $\tilde{u}_{0d} \overset{iid}{\sim} \mathcal{PP}(0, \tilde{C}(x, x'))$, and then we have

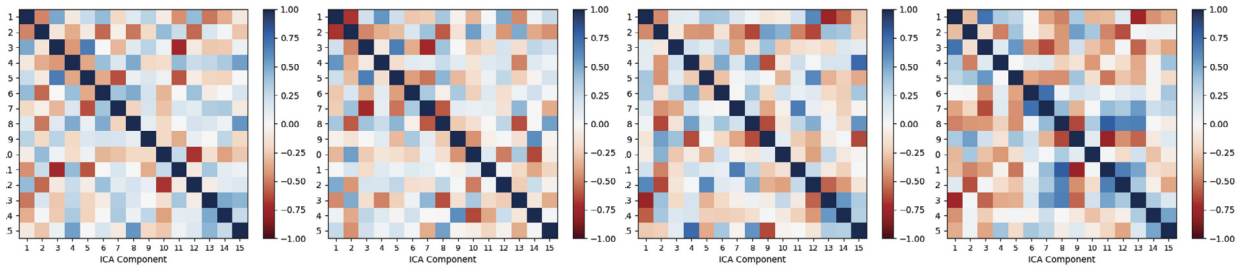
$$\begin{aligned}&\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) \\ =&\sum_{v=1}^{\mathcal{V}} l_i l_j l_k l_l \operatorname{cov}(\tilde{u}_{vi}(x)\tilde{u}_{vj}(x), \tilde{u}_{vk}(x')\tilde{u}_{vl}(x')) \\ =&\mathcal{V} l_i l_j l_k l_l \operatorname{cov}(\tilde{u}_{0i}(x)\tilde{u}_{0j}(x), \tilde{u}_{0k}(x')\tilde{u}_{0l}(x')).\end{aligned} \tag{A.5}$$

Because of the symmetric property of covariance, let $s \neq t$, and we only need to consider three classes summarized as the following three cases:

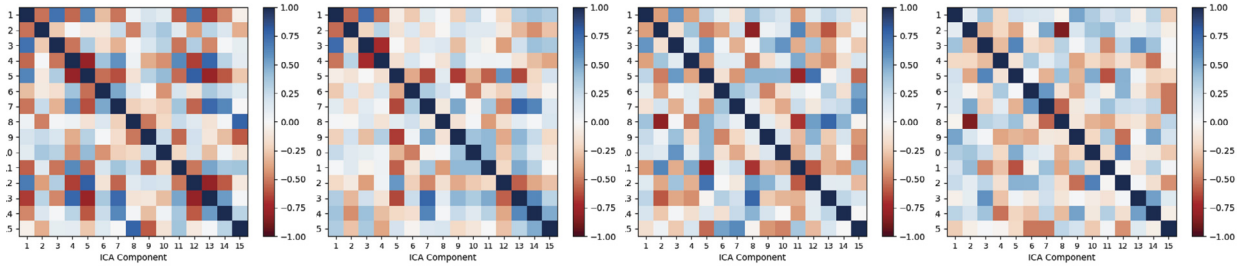
(i) $\operatorname{cov}(\Sigma_{ss}(x), \Sigma_{ss}(x'))$.
(ii) $\operatorname{cov}(\Sigma_{st}(x), \Sigma_{st}(x'))$ and $\operatorname{cov}(\Sigma_{st}(x), \Sigma_{ts}(x'))$.
(iii) Otherwise.

For the first case, without loss of generality, we assume $i = j = k = l$, then we rewrite (A.5) as

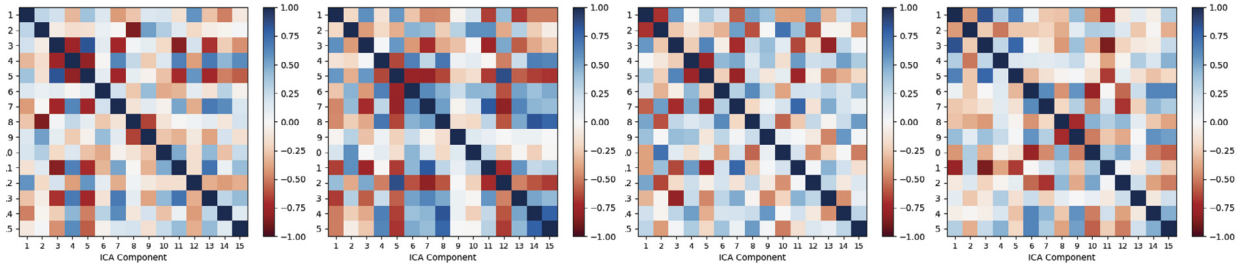
$$\begin{aligned}&\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) \\ =&\mathcal{V} l_i l_j l_k l_l \big(\mathrm{E}(\tilde{u}_{0i}^2(x)\tilde{u}_{0i}^2(x')) - \mathrm{E}(\tilde{u}_{0i}^2(x))\mathrm{E}(\tilde{u}_{0i}^2(x'))\big) \\ =&\mathcal{V} l_i l_j l_k l_l \big(\tilde{C}(x, x)\tilde{C}(x', x') + 2\tilde{C}^2(x, x') - \tilde{C}(x, x)\tilde{C}(x', x')\big) \\ =&2\mathcal{V} l_i^4 \tilde{C}^2(x, x').\end{aligned} \tag{A.6}$$



(a) Dynamic correlations on participant with ID 169949



(b) Dynamic correlations on participant with ID 199958



(c) Dynamic correlations on participant with ID 668361

Fig. B.5. Dynamic correlations (i.e., dynamic functional connectivity between ICA components) derived from the estimations of $\Sigma(x)$ at $x = 1001$, 2001, 3001 and 4800 with HCP timeseries data. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

In the second case, without loss of generality, we assume $i = k \neq j = l$, then we rewrite (A.5) as

$$\begin{aligned}
&\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) \\
=&\mathcal{V} l_i l_j l_k l_l \big(\mathrm{E}(\tilde{u}_{0i}(x)\tilde{u}_{0i}(x'))\mathrm{E}(\tilde{u}_{0j}(x)\tilde{u}_{0j}(x')) \\
&- \mathrm{E}(\tilde{u}_{0i}(x)\tilde{u}_{0j}(x))\mathrm{E}(\tilde{u}_{0i}(x')\tilde{u}_{0j}(x'))\big) \\
=&\mathcal{V} l_i^2 l_j^2 \tilde{C}^2(x, x'). \qquad \text{(A.7)}
\end{aligned}$$

The third case includes two situations: (a) $i \neq j, k, l$, or (b) $i = j \neq k = l$. As for situation (a), (A.5) is rewritten as

$$\begin{aligned}
&\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) \\
=&\mathcal{V} l_i l_j l_k l_l \big(\mathrm{E}(\tilde{u}_{0i}(x)\tilde{u}_{0j}(x)\tilde{u}_{0k}(x')\tilde{u}_{0l}(x')) \\
&- \mathrm{E}(\tilde{u}_{0i}(x)\tilde{u}_{0j}(x))\mathrm{E}(\tilde{u}_{0k}(x')\tilde{u}_{0l}(x'))\big) \\
=&\mathcal{V} l_i l_j l_k l_l \big(\mathrm{E}(\tilde{u}_{0i}(x))\mathrm{E}(\tilde{u}_{0j}(x)\tilde{u}_{0k}(x')\tilde{u}_{0l}(x')) \\
&- \mathrm{E}(\tilde{u}_{0i}(x))\mathrm{E}(\tilde{u}_{0j}(x))\mathrm{E}(\tilde{u}_{0k}(x')\tilde{u}_{0l}(x'))\big) = 0. \qquad \text{(A.8)}
\end{aligned}$$

And it is trivial that situation (b) has the same result. □

Appendix B. Dynamic correlation matrices on more participants

We also display the dynamic correlation matrices derived from the estimated $\Sigma(x)$ at timestamp $x = 1001, 2001, 3001$ and 4800 on more randomly selected participants with IDs 169946, 199958 and 668361 in Fig. B.5 part (a), (b) and (c),

respectively. These plots show the changes of brain connectivity across time as well and further provide evidences that the structure of covariance/correlation may be significantly time-varying.

References

Arditi, A., 2005. Improving the design of the letter contrast sensitivity test. Investig. Ophthalmol. Vis. Sci. 46 (6), 2225–2229.

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc., Ser. B, Stat. Methodol. 70 (4), 825–848.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M., et al., 2003. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. Bayesian Stat. 7 (453–464), 210.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn. Reson. Med. 34 (4), 537–541.

Biswal, B.B., 2012. Resting state fMRI: a personal history. Neuroimage 62 (2), 938–944.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: a review for statisticians. J. Am. Stat. Assoc. 112 (518), 859–877.

Bru, M.F., 1991. Wishart processes. J. Theor. Probab. 4 (4), 725–751.

Buysse, D.J., Reynolds III, C.F., Monk, T.H., Berman, S.R., Kupfer, D.J., 1989. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res. 28 (2), 193–213.

Cappiello, L., Engle, R.F., Sheppard, K., 2006. Asymmetric dynamics in the correlations of global equity and bond returns. J. Financ. Econom. 4 (4), 537–572.

Chai, B., Walther, D.B., Beck, D.M., Fei-Fei, L., 2009. Exploring functional connectivity of the human brain using multivariate information analysis. Neural Inf. Process. Ser. 22, 270–278.

Chib, S., Nardari, F., Shephard, N., 2006. Analysis of high dimensional multivariate stochastic volatility models. J. Econom. 134 (2), 341–371.

Crum, R.M., Anthony, J.C., Bassett, S.S., Folstein, M.F., 1993. Population-based norms for the mini-mental state examination by age and educational level. JAMA 269 (18), 2386–2391.

Cunningham, J.P., Yu, B.M., 2014. Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. 17 (11), 1500–1509.

Dai, M., Zhang, Z., Srivastava, A., 2016. Testing stationarity of brain functional connectivity using change-point detection in fMRI data. In: CVPR Workshop, pp. 19–27.

Dougherty, B.E., Flom, R.E., Bullimore, M.A., 2005. An evaluation of the Mars letter contrast sensitivity test. Optom. Vis. Sci. 82 (11), 970–975.

Engle, R., 2002. Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. J. Bus. Econ. Stat. 20 (3), 339–350.

Engle, R.F., Kroner, K.F., 1995. Multivariate simultaneous generalized arch. Econom. Theory 11 (1), 122–150.

Engle, R.F., Sheppard, K., 2001. Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. Tech. rep. National Bureau of Economic Research.

Finley, A.O., Sang, H., Banerjee, S., Gelfand, A.E., 2009. Improving the performance of predictive process modeling for large datasets. Comput. Stat. Data Anal. 53 (8), 2873–2884.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. “Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12 (3), 189–198.

Fox, E.B., Dunson, D.B., 2015. Bayesian nonparametric covariance regression. J. Mach. Learn. Res. 16 (1), 2501–2542.

Fox, E.B., West, M., 2011. Autoregressive models for variance matrices: stationary inverse Wishart processes. arXiv preprint. arXiv:1107.5239.

Gelfand, A.E., Banerjee, S., Gamerman, D., 2005. Spatial process modelling for univariate and multivariate dynamic spatial data. Environmetrics 16 (5), 465–479.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6 (6), 721–741.

Gershon, R.C., Wagster, M.V., Hendrie, H.C., Fox, N.A., Cook, K.F., Nowinski, C.J., 2013. NIH toolbox for assessment of neurological and behavioral function. Neurology 80 (11), S2–S6.

Gouriéroux, C., Jasiak, J., Sufana, R., 2009. The Wishart autoregressive process of multivariate stochastic volatility. J. Econom. 150 (2), 167–181.

Greicius, M., 2008. Resting-state functional connectivity in neuropsychiatric disorders. Curr. Opin. Neurol. 21 (4), 424–430.

Haymes, S.A., Roberts, K.F., Cruess, A.F., Nicolela, M.T., LeBlanc, R.P., Ramsey, M.S., Chauhan, B.C., Artes, P.H., 2006. The letter contrast sensitivity test: clinical evaluation of a new design. Investig. Ophthalmol. Vis. Sci. 47 (6), 2739–2745.

Hindriks, R., Adhikari, M.H., Murayama, Y., Ganzetti, M., Mantini, D., Logothetis, N.K., Deco, G., 2016. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? NeuroImage 127, 242–256.

Huber, F., Feldkircher, M., 2019. Adaptive shrinkage in bayesian vector autoregressive models. J. Bus. Econ. Stat. 37 (1), 27–39.

Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Della Penna, S., Duyn, J.H., Glover, G.H., Gonzalez-Castillo, J., et al., 2013. Dynamic functional connectivity: promise, issues, and interpretations. NeuroImage 80, 360–378.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K., 1999. An introduction to variational methods for graphical models. Mach. Learn. 37 (2), 183–233.

Kastner, G., Frühwirth-Schnatter, S., Lopes, H.F., 2017. Efficient bayesian inference for multivariate factor stochastic volatility models. J. Comput. Graph. Stat. 26 (4), 905–917.

Keilholz, S.D., 2014. The neural basis of time-varying resting-state functional connectivity. Brain Connect. 4 (10), 769–779.

Lee, N., Kim, J.-M., 2021. Dynamic functional connectivity analysis based on time-varying partial correlation with a copula-dcc-GARCH model. Neurosci. Res. 169, 27–39.

Leonardi, N., Richiardi, J., Gschwind, M., Simioni, S., Annoni, J.-M., Schluep, M., Vuilleumier, P., Van De Ville, D., 2013. Principal components of functional connectivity: a new approach to study dynamic brain connectivity during rest. NeuroImage 83, 937–950.

Li, L., Pluta, D., Shahbaba, B., Fortin, N., Ombao, H., Baldi, P., 2019. Modeling dynamic functional connectivity with latent factor gaussian processes. In: Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/bf499a12e998d178afd964adf64a60cb-Paper.pdf.

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

Meng, R., Bouchard, K., 2021. Bayesian inference in high-dimensional time-series with the orthogonal stochastic linear mixing model. arXiv preprint. arXiv: 2106.13379.

Meng, R., Soper, B., Lee, H.K., Liu, V.X., Greene, J.D., Ray, P., 2021. Nonstationary multivariate gaussian processes for electronic health records. J. Biomed. Inform. 117, 103698.

Meng, R., Lee, H., Bouchard, K., 2022. Stochastic Collapsed Variational Inference for Structured Gaussian Process Regression Network. arXiv preprint. arXiv: 2106.00719.

Monti, R.P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C., Montana, G., 2014. Estimating time-varying brain connectivity networks from functional mri time series. NeuroImage 103, 427–443.
Orskaug, E., 2009. Multivariate DCC-GARCH model:-with various error distributions. Master's thesis. Institutt for Matematiske Fag.
Pourahmadi, M., 1999. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. Biometrika 86 (3), 677–690.
Seiler, C., Holmes, S., 2017. Multivariate heteroscedasticity models for functional brain connectivity. Front. Neurosci. 11, 696.
Smith, S.M., 2012. The future of fmri connectivity. NeuroImage 62 (2), 1257–1266.
Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., et al., 2013. Resting-state fMRI in the human connectome project. NeuroImage 80, 144–168.
Van Den Heuvel, M.P., Pol, H.E.H., 2010. Exploring the brain network: a review on resting-state fMRI functional connectivity. Eur. Neuropsychopharmacol. 20 (8), 519–534.
Varoquaux, G., Gramfort, A., Poline, J.-B., Thirion, B., 2010. Brain covariance selection: better individual functional connectivity models using population prior. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems, vol. 2. NIPS'10. Curran Associates Inc., Red Hook, NY, USA, pp. 2334–2342.
Warnick, R., Guindani, M., Erhardt, E., Allen, E., Calhoun, V., Vannucci, M., 2018. A bayesian approach for estimating dynamic functional network connectivity in fmri data. J. Am. Stat. Assoc. 113 (521), 134–151.
Wilson, A.G., Ghahramani, Z., 2010. Generalised Wishart processes. arXiv preprint. arXiv:1101.0240.
WU-Minn, H., 2017. 1200 subjects data release reference manual. https://www.humanconnectome.org.
Yin, J., Geng, Z., Li, R., Wang, H., 2010. Nonparametric covariance model. Stat. Sin. 20, 469.
Zhang, K., Tsang, I.W., Kwok, J.T., 2008. Improved Nyström low-rank approximation and error analysis. In: ICML, pp. 1232–1239.
Zhang, W., Leng, C., 2012. A moving average Cholesky factor model in covariance modelling for longitudinal data. Biometrika 99 (1), 141–150.
Zhu, Y., Zhu, X., Kim, M., Kaufer, D., Laurienti, P.J., Wu, G., 2019. Characterizing dynamic functional connectivity using data-driven approaches and its application in the diagnosis of Alzheimer's disease. In: Connectomics. Elsevier, pp. 181–197.