Learning Temporal Evolution of Spatial Dependence with Generalized Spatiotemporal Gaussian Process Models

Shiwei Lan   SLAN@ASU.EDU
School of Mathematical and Statistical Sciences
Arizona State University
Tempe, AZ 85287, USA

Editor: Matthias Seeger

Abstract

A large number of scientific studies involve high-dimensional spatiotemporal data with complicated relationships. In this paper, we focus on a type of space-time interaction named *temporal evolution of spatial dependence (TESD)*, which is a zero time-lag spatiotemporal covariance. For this purpose, we propose a novel Bayesian nonparametric method based on non-stationary spatiotemporal Gaussian process (STGP). The classic STGP has a covariance kernel separable in space and time, failed to characterize TESD. More recent works on non-separable STGP treat location and time together as a joint variable, which is unnecessarily inefficient. We generalize STGP (gSTGP) to introduce time-dependence to the spatial kernel by varying its eigenvalues over time in the Mercer's representation. The resulting non-stationary non-separable covariance model bares a quasi Kronecker sum structure. Finally, a hierarchical Bayesian model for the joint covariance is proposed to allow for full flexibility in learning TESD. A simulation study and a longitudinal neuroimaging analysis on Alzheimer's patients demonstrate that the proposed methodology is (statistically) effective and (computationally) efficient in characterizing TESD. Theoretic properties of gSTGP including posterior contraction (for covariance) are also studied.

Keywords: Temporal Evolution of Spatial Dependence, Spatiotemporal Gaussian process, Non-stationary Non-separable Kernel, Quasi Kronecker Product/Sum Structure, Nonparametric Spatiotemporal Covariance Model

1. Introduction

Spatiotemporal data are ubiquitous in our daily life. For example, the climate data manifest a trend of global warming, and the traffic data feature a network structure in space and a periodic pattern in time. There is usually intricate interaction between space and time in such spatiotemporal process $y(\mathbf{x}, t)$. In particular, the dependence among spatial locations, $\mathrm{Cov}(y(\mathbf{x}, \cdot), y(\mathbf{x}', \cdot))$, may change over time. For example, in the study of dynamic brain connectivity in neuroscience (Cribben et al., 2012; Fiecas and Ombao, 2016; Lan et al., 2020), the spatial dependence among multi-site brain signals varies along certain cognitive processes. In the longitudinal analysis of brain images (Hyun et al., 2016), different brain regions also have changing connection in the progression of diseases such as Alzheimer. We formally define such *temporal evolution of spatial dependence (TESD)* for process $y(\mathbf{x}, t)$ as a zero time-lag $(t = t')$ covariance, i.e. $\mathrm{Cov}(y(\mathbf{x}, t), y(\mathbf{x}', t))$. In general, TESD is an important subject to understand complex relationships, to predict their progress, and to

extrapolate them to unknown territory. In this work, we propose a novel fully Bayesian nonparametric model based on a generalized spatiotemporal Gaussian process (STGP) to characterize TESD in spatiotemporal data.

STGP is a special type of Gaussian process that can model both spatial and temporal information simultaneously. There is a rich literature on STGP, of which a large body (Cressie and Huang, 1999; Gneiting, 2002; Sarkka and Hartikainen, 2012; Sarkka et al., 2013; Niu et al., 2015) imposes stationarity condition on the covariance $\mathcal{C}(\mathbf{x} - \mathbf{x}', t - t')$. However, stationary STGP would not work for learning TESD ($t = t'$) because temporal stationarity directly induces TESD constant in time, i.e. $\mathcal{C}(\mathbf{x} - \mathbf{x}', 0)$. Consequently, "independent and identically distributed (i.i.d.)" (Marco et al., 2015) is not an appropriate assumption for spatiotemporal observations with changing TESD. From a multivariate time series point of view, these observations can be seen as a realization of a vector process $\mathbf{y}(t) = (y(\mathbf{x}_1, t), \cdots, y(\mathbf{x}_I, t))$ that has time-invariant covariance (TESD).

Classical STGP often assumes a separable structure for the kernel such as $\mathcal{C}_{\mathbf{x}} \otimes \mathcal{C}_t$, e.g. Paciorek and Schervish (2003, 2006); Hartikainen et al. (2011); Sarkka and Hartikainen (2012); Sarkka et al. (2013); Niu et al. (2015); Kuzin et al. (2018); Todescato et al. (2020). Separable covariance in this setting ($t = t'$) reduces to a scalar factor by a spatial kernel, i.e. $\mathcal{C}_{\mathbf{x}}(\mathbf{x}, \mathbf{x}') \cdot \sigma_t^2$, thus loses full freedom in describing TESD (See more details in Section 2.1). To better reflect the space-time interactions, a large class of non-separable STGP models have been proposed by parametric construction (Cressie and Huang, 1999; Gneiting, 2002), by spectral representation (Fuentes et al., 2008), and by kernel convolution (Marco et al., 2015; Wang et al., 2020; Das and Bhattacharya, 2020) or mixing (Fonseca and Steel, 2011), etc. Some works (Datta et al., 2016; Hyun et al., 2016) treat spatial and temporal variables with no difference in the joint kernel, leading to a high-dimensional dense covariance matrix which is inefficient for learning TESD (See Section 2.4).

Even among non-stationary non-separable covariance models (Fuentes et al., 2008; Singh et al., 2010; Cressie and Wikle, 2011; Luttinen and Ilin, 2012; Marco et al., 2015; Datta et al., 2016; Hyun et al., 2016; Senanayake et al., 2016; Zhang and Cressie, 2020; Wang et al., 2020; Das and Bhattacharya, 2020), most of them focus on modeling and predicting mean functions but are not designed for efficient and flexible learning of TESD.

In the majority of existing works, the joint covariance kernel is built from parametric covariance functions such as exponential, Matérn, or their derivations. For example, Cressie and Huang (1999); Gneiting (2002) give conditions on admissible functions for such parametric construction of stationary non-separable models and Fuentes et al. (2008) generalize to the non-stationary non-separable case. Although regarded as nonparametric models for mean functions, they are limited in their capability of learning TESD (See Section 5.1). Even in the works of more recently proposed deep GP (Damianou and Lawrence, 2013; Salimbeni and Deisenroth, 2017; Dunlop et al., 2018; Zhao et al., 2021), covariance kernels bare very flexible structures but are parameterized by neural networks. In general, there is a lack of flexible and efficient Bayesian non-parametric models particularly for *spatiotemporal covariance learning* (e.g. TESD). This paper aims to fill the blank in the literature.

To learn TESD, we propose a time-dependent spatial kernel, $\mathcal{C}_{\mathbf{x}|t}$, to generalize classical separable STGP. Based on the Mercer's representation of the spatial kernel, $\mathcal{C}_{\mathbf{x}}$, we introduce the time-dependence by varying the eigenvalues of $\mathcal{C}_{\mathbf{x}}$ in time. Moreover, we endow an independent GP (hyper-)prior on these dynamic eigenvalues and obtain a nonparametric

model to flexibly capture complex TESD (See Section 2.3 and Section 5.1). To respect the time-changing nature of TESD, we construct a novel non-separable kernel with a quasi Kronecker sum structure (See model II in Section 2.2). This structure results in a highly sparse joint covariance matrix, making it amenable for efficient inference. As a reference, we also consider a quasi Kronecker product generalization more suitable for i.i.d. observations (See model I in Section 2.2). This is done to justify the careful design of covariance kernel, very essential in learning TESD with STGP.

The Mercer's representation and the Karhunen-Loéve expansion were considered in the literature of spatiotemporal modeling (Mike West, 1997; Wikle and Cressie, 1999; Wikle, 2002; Fontanella and Ippoliti, 2003; Cressie and Wikle, 2011). Similar ideas of mixture of basis functions have recently been explored in (coregionalization, Banerjee, 2015), (Dirichlet mixture, Gelfand et al., 2005; Das and Bhattacharya, 2020). The novelty of our construction lies in: 1) the quasi Kronecker sum formulation with a balanced structure designed for efficient covariance learning (Section 2.4), 2) random construction of dynamic spatial kernel flexible for modeling TESD (Section 2.3.2), and 3) theoretic guarantee for the Bayesian learning of TESD (Section 3.3). The proposed methodology also generalizes the semi-parametric approaches (Wilson and Ghahramani, 2011; Fox and Dunson, 2015; Lan et al., 2020).

Focusing on TESD in the spatiotemporal covariance learning, this work makes multi-fold contributions:

1. This is a novel non-separable, non-stationary, and fully nonparametric covariance model dedicated to learning TESD in spatiotemporal analysis;
2. The separable STGP is generalized by introducing the time-dependence to the spatial kernel via the Mercer's representation;
3. It provides a systematic comparison among multiple model structures of STGP based on (quasi) Kronecker product and sum both theoretically and numerically.

The proposed methodology sheds light on brain degradation of Alzheimer's patients in a longitudinal neuroimaging analysis. The potential utility of this work can be found in multiple areas including the genome-wide association study (GWAS), climate change, and investment management etc.

The rest of the paper is organized as follows. In Section 2, the classical separable STGP (baseline model 0) is reviewed and generalized by introducing the time-dependence to the spatial kernel. Two model structures, quasi Kronecker product (referenced model I) and quasi Kronecker sum (proposed model II), in the generalized STGP (gSTGP) are proposed and compared for learning TESD. Then we systematically investigate various theoretic properties of gSTGP in Section 3 and briefly discuss the posterior inference and the prediction of TESD in Section 4. In Section 5, we conduct a simulation study (Section 5.1) and apply the proposed methodology to analyze a series of positron emission tomography (PET) brain images of Alzheimer's patients (Section 5.2) obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI, 2003). By comparing with various non-stationary/non-separable models, we illustrate the effectiveness and efficiency of the proposed gSTGP in modeling and predicting TESD of spatiotemporal processes. Finally we conclude in Section 6 with a few comments on the methodology and some discussions of future directions.

2. Generalizing Spatiotemporal Gaussian Processes

In this section, we first define the (separable) STGP using the matrix normal distribution and explain why it fails to characterize TESD. This motivates the generalization of STGP to introduce the time-dependent spatial kernel. Two model structures based on quasi Kronecker product and sum are constructed and compared but the latter is found to be more theoretically effective and computationally efficient in learning TESD.

Let $\mathcal{X} \subset \mathbb{R}^d$ be a bounded spatial domain and let $\mathcal{T} \subset \mathbb{R}_+$ be a bounded temporal domain. Denote $\mathcal{Z} := \mathcal{X} \times \mathcal{T}$ as the joint domain and $\mathbf{z} := (\mathbf{x}, t)$ as the joint variable. The spatiotemporal data $\{y_{ij} | i = 1, \cdots, I; j = 1, \cdots, J\}$ are taken on a grid of points $\{\mathbf{z}_{ij} = (\mathbf{x}_i, t_j) | \mathbf{x}_i \in \mathcal{X}, t_j \in \mathcal{T}\}$ with the spatial discrete size I and the temporal discrete size J. A (centered) STGP is uniquely determined by its covariance kernel $\mathcal{C}_\mathbf{z} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, a bilinear symmetric positive-definite function. $\mathcal{C}_\mathbf{z}$ could be defined by exploring structures in space and time, to be detailed below.

2.1 Spatiotemporal Gaussian Process

The spatiotemporal data $\{y_{ij}\}$ have noise usually assumed i.i.d. and are often modeled using the standard STGP model:

$$
\begin{aligned}
y_{ij} &= f(\mathbf{x}_i, t_j) + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2) \\
f(\mathbf{z}) &\sim \mathcal{GP}(0, \mathcal{C}_\mathbf{z})
\end{aligned}
\tag{1}
$$

where the joint spatiotemporal kernel $\mathcal{C}_\mathbf{z}$ is in the form of a Kronecker product of spatial kernel $\mathcal{C}_\mathbf{x}$ and temporal kernel $\mathcal{C}_t$:

$$
\text{model 0 (separable)}: \quad \mathcal{C}_\mathbf{z} = \mathcal{C}_\mathbf{x} \otimes \mathcal{C}_t : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}, \quad (\mathbf{z}, \mathbf{z}') \mapsto \mathcal{C}_\mathbf{x}(\mathbf{x}, \mathbf{x}') \cdot \mathcal{C}_t(t, t') \tag{2}
$$

With such separable kernel, we define the classical STGP through the matrix normal distribution as follows.

Definition 1 (STGP) *A stochastic process $f(\mathbf{x}, t)$ is called (separable) spatiotemporal Gaussian process with a mean function $m(\mathbf{x}, t)$, a spatial kernel $\mathcal{C}_\mathbf{x}$ and a temporal kernel $\mathcal{C}_t$ if for any finite collection of locations $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^I$ and times $\mathbf{t} = \{t_j\}_{j=1}^J$,*

$$
\mathbf{F} = f(\mathbf{X}, \mathbf{t}) = [f(\mathbf{x}_i, t_j)]_{I \times J} \sim \mathcal{MN}_{I \times J}(\mathbf{M}, \mathbf{C}_\mathbf{x}, \mathbf{C}_t) \tag{3}
$$

where $\mathbf{M} = m(\mathbf{X}, \mathbf{t}) = [m(\mathbf{x}_i, t_j)]_{I \times J}$, $\mathbf{C}_\mathbf{x} = \mathcal{C}_\mathbf{x}(\mathbf{X}, \mathbf{X}) = [\mathcal{C}(\mathbf{x}_i, \mathbf{x}_{i'})]_{I \times I}$, $\mathbf{C}_t = \mathcal{C}_t(\mathbf{t}, \mathbf{t}) = [\mathcal{C}(t_j, t_{j'})]_{J \times J}$, and the matrix normal distribution $\mathcal{MN}$ is interpreted through the vectorization as in the following remark. We denote $f \sim \mathcal{GP}(m, \mathcal{C}_\mathbf{x}, \mathcal{C}_t)$.

Remark 1 *If we vectorize the matrix $\mathbf{F}_{I \times J}$, then we have*

$$
\text{vec}(\mathbf{F}) \sim \mathcal{N}(\text{vec}(\mathbf{M}), \mathbf{C}_t \otimes \mathbf{C}_\mathbf{x}), \quad \text{vec}(\mathbf{F}^\mathsf{T}) \sim \mathcal{N}(\text{vec}(\mathbf{M}^\mathsf{T}), \mathbf{C}_\mathbf{x} \otimes \mathbf{C}_t)
$$

where $\otimes$ is the regular Kronecker product for matrices. For this reason, we also denote a (centered) STGP as $f \sim \mathcal{GP}(0, \mathcal{C}_\mathbf{x} \otimes \mathcal{C}_t)$.

Denote $\sigma_t^2 := \mathcal{C}_t(t,t)$. For any fixed time $t \in \mathcal{T}$, the covariance of the separable STGP $f(\mathbf{x}, t)$ in the space domain is reduced to

$$\text{Cov}[f(\mathbf{x}, t), f(\mathbf{x}', t)] = \mathcal{C}_{\mathbf{x}}(\mathbf{x}, \mathbf{x}') \cdot \sigma_t^2, \quad \forall t \in \mathcal{T} \tag{4}$$

which at best changes by a scalar factor in time (without full freedom in describing the time change of spatial dependence), or at worst is constant in time (when a stationary $\mathcal{C}_t$ is adopted). Such drawback of the separable kernel makes the corresponding STGP unable to characterize TESD, defined as follows.

Definition 2 (TESD) *The temporal evolution of spatial dependence (TESD) of a spatiotemporal process $y(\mathbf{x}, t)$ is the spatial covariance conditioned on a common time $t = t' \in \mathcal{T}$, denoted as $\mathcal{C}_{y|t}$:*

$$\mathcal{C}_{y|t}(\mathbf{x}, \mathbf{x}') := \text{Cov}[y(\mathbf{x}, t), y(\mathbf{x}', t)], \quad for \; \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{5}$$

Remark 2 *TESD can be viewed as a zero time-lag covariance in the spatiotemporal analysis. Therefore, the (weak) stationarity (in time) is an inappropriate assumption as it induces constant (hence trivial) TESD regardless of the kernel structure. In the following we always assume non-stationarity for the process of interest unless specified otherwise.*

Because of the failure of separable kernel for learning TESD, we are motivated to remedy the classical STGP with time-dependence in the spatial covariance.

2.2 Generalized Spatiotemporal Gaussian Process

In this subsection, we generalize the classical STGP by introducing time-dependence to the spatial kernel $\mathcal{C}_{\mathbf{x}}$ while keeping the desirable structure in the joint kernel $\mathcal{C}_{\mathbf{z}}$. Two structures are constructed based on quasi Kronecker product (model I) and sum (model II) respectively. We promote model II as an effective and efficient tool for learning TESD and use model I as a reference to emphasize the importance of kernel design in STGP.

First, it is intuitive to replace the spatial $\mathcal{C}_{\mathbf{x}}$ kernel with a time-dependent analogy $\mathcal{C}_{\mathbf{x}|t}$ in the separable kernel (2) to derive the following non-separable joint kernel

$$\mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{C}_t \tag{6}$$

where $\mathcal{C}_{\mathbf{x}|t}$ and the quasi Kronecker product $\dot{\otimes}$ will be defined in Section 2.3. If we view $\{y_{ij}\}$ as observations with i.i.d. noise taken from a spatiotemporal process $y(\mathbf{z})$, then the marginal covariance for y in (1) becomes

$$\text{model I (qKron-prod)}: \quad \mathcal{C}_y^{\text{I}} = \underbrace{\mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{C}_t}_{prior} + \underbrace{\sigma_\varepsilon^2 \mathcal{I}_{\mathbf{x}} \otimes \mathcal{I}_t}_{likelihood} \tag{7}$$

where $\mathcal{I}_{\mathbf{x}}(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x} = \mathbf{x}')$, and $\mathcal{I}_t(t, t') = \delta(t = t')$ with $\delta(\cdot)$ being the Dirac function.

Alternatively, Lan et al. (2020) take a perspective from multivariate time series and model the noise in $\{y_{ij}\}$ as independent but not identically distributed (i.n.i.d.) (a more reasonable assumption for nonconstant TESD):

$$\mathbf{y}(t_j) := \{y(\mathbf{x}_i, t_j)\}_{i=1}^I = \mathbf{f}(t_j) + \boldsymbol{\varepsilon}_j, \quad \boldsymbol{\varepsilon}_j \overset{inid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{t_j})$$
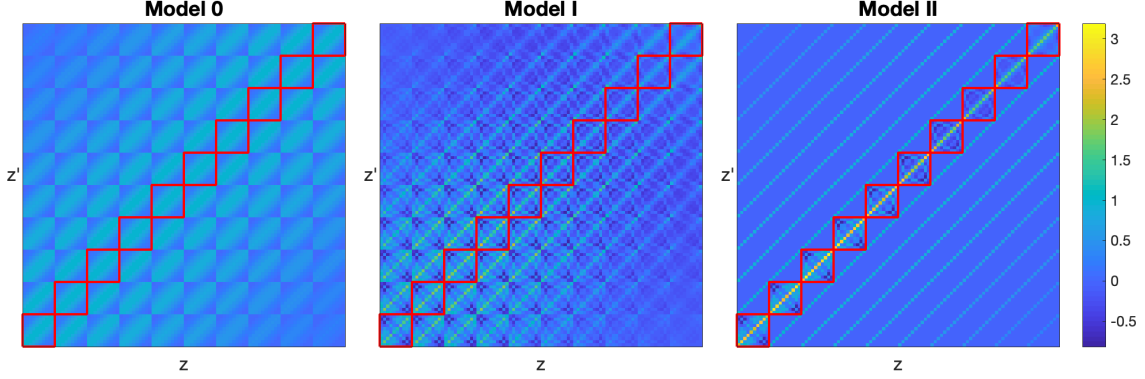$$f_i(t) \overset{iid}{\sim} \mathcal{GP}(0, \mathcal{C}_t), \quad i = 1, \cdots, I. \tag{8}$$

Figure 1: Joint kernels $\mathcal{C}_y$ specified by model 0 with separable structure $\mathcal{C}_{\mathbf{x}} \otimes \mathcal{C}_t$ (left), model I with quasi Kronecker product structure $\mathcal{C}_{\mathbf{x}|t} \,\dot{\otimes}\, \mathcal{C}_t$ (middle) and model II with quasi Kronecker sum structure $\mathcal{C}_{\mathbf{x}|t} \,\dot{\oplus}\, \mathcal{C}_t$ (right). Blocks in red frame illustrate the temporal evolution of spatial kernel $\mathcal{C}_{y|t}$ (TESD). Note $\mathbf{z} = (\mathbf{x}, t)$.

where the parametric covariance matrix $\mathbf{\Sigma}_t$ encodes the spatial dependence among multiple time series changing with time t. If we replace $\mathbf{\Sigma}_t$ with a time-dependent spatial kernel $\mathcal{C}_{\mathbf{x}|t}$, then (8) can be generalized to a fully non-parametric model:

$$
\begin{aligned}
y(\mathbf{x}, t) | m, \mathcal{C}_{\mathbf{x}|t} &\sim \mathcal{GP}_{\mathbf{x}}(m, \mathcal{C}_{\mathbf{x}|t} \,\dot{\otimes}\, \mathcal{I}_t) \\
m(\mathbf{x}, t) &\sim \mathcal{GP}_t(0, \mathcal{I}_{\mathbf{x}} \otimes \mathcal{C}_t)
\end{aligned}
\tag{9}
$$

which has the marginal covariance for y in a form of quasi Kronecker sum (see Section 2.3):

$$
\text{model II (qKron-sum)}: \qquad \mathcal{C}_y^{\text{II}} = \underbrace{\mathcal{I}_{\mathbf{x}} \otimes \mathcal{C}_t}_{prior} + \underbrace{\mathcal{C}_{\mathbf{x}|t} \,\dot{\otimes}\, \mathcal{I}_t}_{likelihood} =: \mathcal{C}_{\mathbf{x}|t} \,\dot{\oplus}\, \mathcal{C}_t
\tag{10}
$$

Figure 1 illustrates the structures of $\mathcal{C}_y$ in the three models. If we arrange the joint state as $\mathbf{Z} = \mathbf{X} \otimes \mathbf{t}$, each of the blocks in red frame indicates a covariance matrix in the space $\mathbf{X}$ at a time t_j, i.e. $\text{Cov}[y(\mathbf{X}, t_j), y(\mathbf{X}, t_j)]_{I \times I}$, and together they describe how $\mathcal{C}_{y|t}$ evolves in time (TESD). From the time horizon, TESD can be viewed as a kernel (matrix) valued function (blocks in red frame) of time t. Model 0 has constant TESD $\mathcal{C}_{y|t}$ (same blocks). On the other hand, both models I and II incorporate the time-dependence for the spatial kernel through $\mathcal{C}_{\mathbf{x}|t}$ (changing blocks). However, by allowing for a complex structure in the blocks off the main diagonal (not in red frame), model I would become overly complicated and expensive to fit. By contrast, model II provides enough flexibility in the main diagonal blocks (in red frame) for TESD, and yet results in a highly sparse joint kernel. See more comparison in Section 2.4. The following proposition states that the time-dependent spatial kernel $\mathcal{C}_{\mathbf{x}|t}$ is the *essence* of TESD $\mathcal{C}_{y|t}$ in these models.

Proposition 1 *If y is a spatiotemporal process according to one of the three models* (2) (7) (10), *then we have the following TESD's*

$$
\mathcal{C}_{y|t}^0 \equiv \underbrace{\mathcal{C}_{\mathbf{x}} \sigma_t^2}_{prior} + \underbrace{\sigma_\varepsilon^2 \mathcal{I}_{\mathbf{x}}}_{likelihood} \;, \quad \mathcal{C}_{y|t}^{\text{I}} = \underbrace{\mathcal{C}_{\mathbf{x}|t} \sigma_t^2}_{prior} + \underbrace{\sigma_\varepsilon^2 \mathcal{I}_{\mathbf{x}}}_{likelihood} \;, \quad \mathcal{C}_{y|t}^{\text{II}} = \underbrace{\mathcal{I}_{\mathbf{x}} \sigma_t^2}_{prior} + \underbrace{\mathcal{C}_{\mathbf{x}|t}}_{likelihood}
\tag{11}
$$

2.3 Construction of Time-Dependent Spatial Kernel

In both models I (7) and II (10), the time-dependent spatial kernel $\mathcal{C}_{\mathbf{x}|t}$ is the key to model TESD. In this subsection, we construct $\mathcal{C}_{\mathbf{x}|t}$ by dynamically varying eigenvalues of the spatial kernel $\mathcal{C}_{\mathbf{x}}$ in the Mercer's theorem.

First, the centered (spatial) GP $\mathcal{GP}(0, \mathcal{C}_{\mathbf{x}})$ is determined by its covariance kernel $\mathcal{C}_{\mathbf{x}}$ which defines a Hilbert-Schmidt integral operator on $L^2(\mathcal{X})$ as follows:

$$T_{\mathcal{C}_{\mathbf{x}}} : L^2(\mathcal{X}) \to L^2(\mathcal{X}), \quad \phi(\cdot) \mapsto \int \mathcal{C}_{\mathbf{x}}(\cdot, \mathbf{x}')\phi(\mathbf{x}')d\mathbf{x}' \tag{12}$$

Denote $\{\lambda_\ell^2, \phi_\ell(\mathbf{x})\}$ as the eigen-pairs of $T_{\mathcal{C}_{\mathbf{x}}}$ such that $T_{\mathcal{C}_{\mathbf{x}}}\phi_\ell(\mathbf{x}) = \lambda_\ell^2\phi_\ell(\mathbf{x})$. Then $\{\phi_\ell(\mathbf{x})\}$ serves as an orthonormal basis for $L^2(\mathcal{X})$. By the Mercer's theorem, we have the following representation of the spatial kernel $\mathcal{C}_{\mathbf{x}}$:

$$\mathcal{C}_{\mathbf{x}}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=1}^{\infty} \lambda_\ell^2 \phi_\ell(\mathbf{x})\phi_\ell(\mathbf{x}') \tag{13}$$

where the series converges in $L^2(\mathcal{X})$ norm.

2.3.1 DETERMINISTIC CONSTRUCTION

To introduce the time-dependence to the spatial kernel, thus denoted as $\mathcal{C}_{\mathbf{x}|t}$, we let the eigenvalues $\{\lambda_\ell^2\}$ change with time and denote them as $\{\lambda_\ell^2(t)\}$. That is, we define

$$\lambda_\ell^2(t) := \langle \phi_\ell(\mathbf{x}), \mathcal{C}_{\mathbf{x}|t}\phi_\ell(\mathbf{x}) \rangle, \quad T_{\mathcal{C}_{\mathbf{x}|t}}\phi_\ell(\mathbf{x}) = \lambda_\ell^2(t)\phi_\ell(\mathbf{x}) \tag{14}$$

Let $\lambda(t) := \{\lambda_\ell(t)\}_{\ell=1}^{\infty}$ for $\forall t \in \mathcal{T}$. To ensure the well-definedness of the generalization, we make the following assumption which essentially requires $\mathcal{C}_{\mathbf{x}|t}$ to be a trace $(\mathrm{tr}(\mathcal{C}_{\mathbf{x}|t}) = \sum_{\ell=1}^{\infty} \lambda_\ell^2(\cdot))$ class operator in $L^1(\mathcal{T})$.

Assumption 1 *We assume $\lambda_\ell \in L^2(\mathcal{T})$ for each $\ell \in \mathbb{N}$ and the infinite sequence λ satisfy*

$$\lambda \in \ell^2(L^2(\mathcal{T})), \quad i.e. \quad \|\lambda\|_{2,2}^2 := \sum_{\ell=1}^{\infty} \|\lambda_\ell(\cdot)\|_2^2 < +\infty \tag{15}$$

Under Assumption 1 we can have the following time-dependent spatial kernel $\mathcal{C}_{\mathbf{x}|t}$ well-defined through the series representation as in the Mercer's theorem (See Theorem 1):

$$\mathcal{C}_{\mathbf{x}|t}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=1}^{\infty} \lambda_\ell^2(t) \cdot \phi_\ell(\mathbf{x})\phi_\ell(\mathbf{x}') \tag{16}$$

where the eigen-basis $\{\phi_\ell(\mathbf{x})\}$ can be chosen as the eigen-functions of the spatial operator $\mathcal{C}_{\mathbf{x}}$ as in (13) or other (e.g. Fourier) basis functions. Such construction has some similarity to the "coregionalization" model (Banerjee, 2015) for which the spatial process $\mathbf{Y}(\mathbf{x}) = \mathbf{A}\mathbf{w}(\mathbf{x})$ has a temporal covariance $\mathbf{T} = \mathbf{A}\mathbf{A}^{\mathsf{T}}$ and independent component processes $w_\ell(\cdot) \sim \mathcal{GP}(0, \rho_\ell)$. Denote $\mathbf{T}_\ell = \mathbf{a}_\ell \mathbf{a}_j^{\mathsf{T}}$ with $\mathbf{a}_\ell$ as the ℓ-th column of $\mathbf{A}$. Thus $\mathbf{T} = \sum_{\ell=1}^{L} \mathbf{T}_\ell$ is finite-dimensional. The resulting covariance $\mathrm{Cov}(\mathbf{Y}(\mathbf{x}), \mathbf{Y}(\mathbf{x}')) = \sum_{\ell=1}^{L} \mathbf{T}_\ell \cdot \rho_\ell(\mathbf{x} - \mathbf{x}')$ is analogous to a finite truncation of (16).
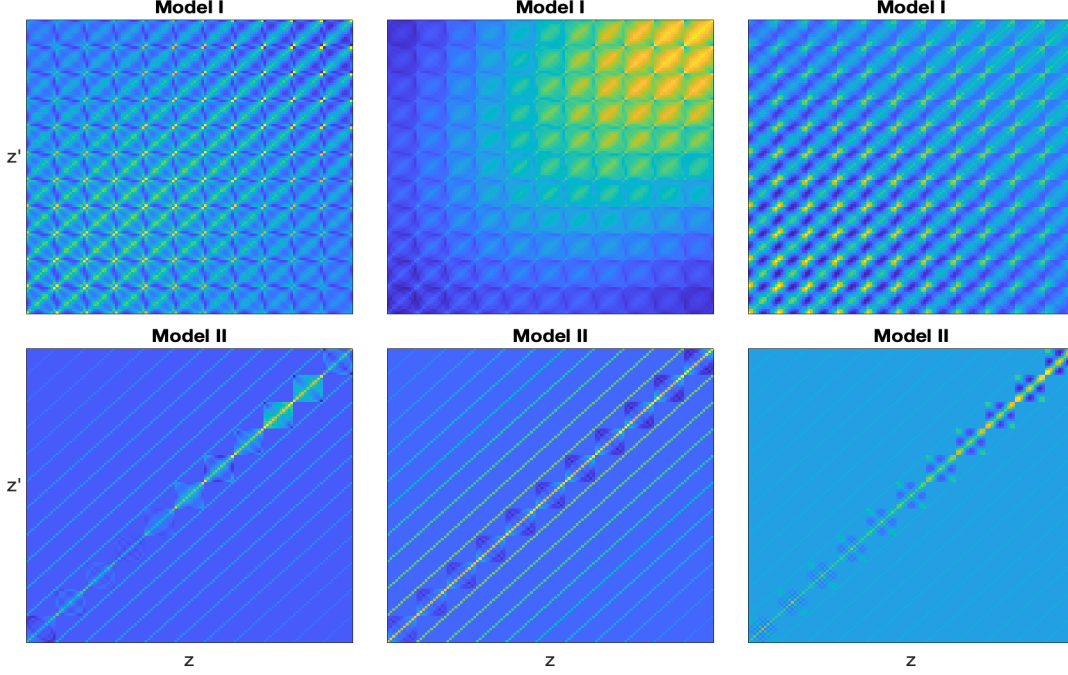
Figure 2: Random draws of the joint covariance $\mathcal{C}_y$ by model I (upper row) and model II (lower row) respectively.

With $\mathcal{C}_{\mathbf{x}|t}$ in (16) we can define the quasi Kronecker product for the prior kernel $\mathcal{C}_m = \mathcal{C}_{\mathbf{x}|t} \,\dot{\otimes}\, \mathcal{C}_t$ in model I and the likelihood kernel $\mathcal{C}_{y|m} = \mathcal{C}_{\mathbf{x}|t} \,\dot{\otimes}\, \mathcal{I}_t$ in model II repsectively:

$$\mathcal{C}_m^{\mathrm{I}}(\mathbf{z}, \mathbf{z}') = \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}} \mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \,\dot{\otimes}\, \mathcal{C}_t(\mathbf{z}, \mathbf{z}') = \sum_{\ell=1}^{\infty} \lambda_\ell(t) \mathcal{C}_t(t, t') \lambda_\ell(t') \phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}') \tag{17}$$

$$\mathcal{C}_{y|m}^{\mathrm{II}}(\mathbf{z}, \mathbf{z}') = \mathcal{C}_{\mathbf{x}|t} \,\dot{\otimes}\, \mathcal{I}_t(\mathbf{z}, \mathbf{z}') = \sum_{\ell=1}^{\infty} \lambda_\ell(t) \mathcal{I}_t(t, t') \lambda_\ell(t') \phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}') \tag{18}$$

2.3.2 RANDOM CONSTRUCTION

To avoid parametric construction of λ_ℓ, we adopt a full Bayesian approach and endow a (hyper) GP prior $\lambda_\ell(\cdot) \sim \mathcal{GP}(0, \mathcal{C}_{\lambda,\ell})$ for each $\ell \in \mathbb{N}$, similarly as what is done for the mean function m in standard GP regression. Now we consider random λ in the probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ with $\Omega = \ell^2(L^2(\mathcal{T}))$, σ-algebra $\mathcal{B}(\Omega)$ and probability measure $\mathbb{P}$ defined as finite product of Gaussian measures $\{\mathcal{GP}(0, \mathcal{C}_{\lambda,\ell})\}$ consistently extended to infinite product by Kolmogorov extension theorem (c.f. Theorem 29 in section A.2.1 of Dashti and Stuart, 2017). To reduce the modeling parameters, we let $\{\mathcal{C}_{\lambda,\ell}\}$ share a common factor $\mathcal{C}_u$ and model them with decaying magnitude, as detailed in the following assumption.

Assumption 2 *Let the random infinite sequence* $\lambda \in (\ell^2(L^2(\mathcal{T})), \mathcal{B}, \mathbb{P})$ *satisfy*

$$\lambda_\ell(\cdot) \sim \mathcal{GP}(0, \mathcal{C}_{\lambda,\ell}), \quad \mathcal{C}_{\lambda,\ell} = \gamma_\ell^2 \mathcal{C}_u, \qquad \sum_{\ell=1}^{\infty} \gamma_\ell^2 < \infty \tag{19}$$
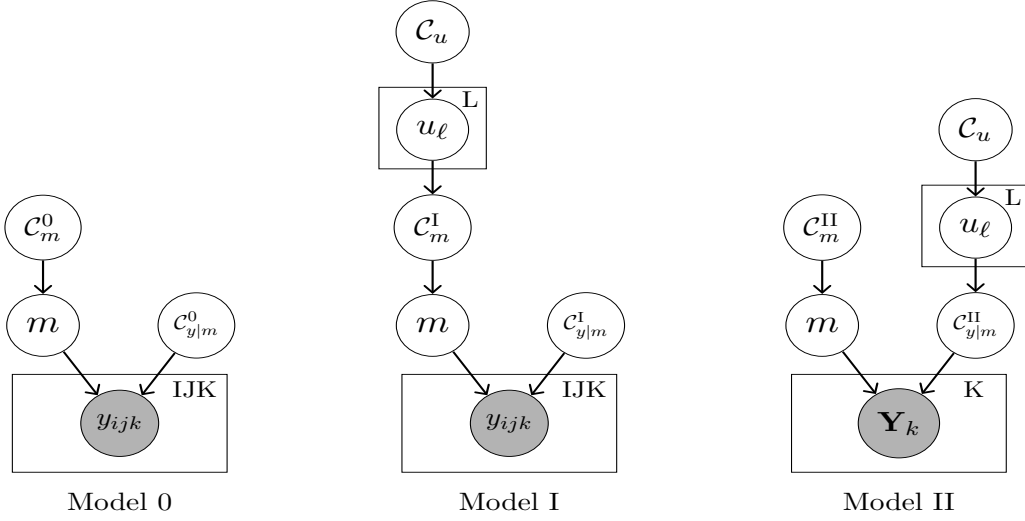
Figure 3: Graphical structures of STGP Model 0 (left), I (middle) and II (right) in (21).

In Practice, we could set for each $\ell \in \mathbb{N}$,

$$\lambda_\ell(t) = \gamma_\ell u_\ell(t), \quad u_\ell(\cdot) \overset{iid}{\sim} \mathcal{GP}(0, \mathcal{C}_u), \quad \gamma_\ell = \ell^{-\kappa/2} \text{ for some } \kappa > 1 \tag{20}$$

Under Assumption 2, Assumption 1 holds in $\mathbb{P}$ because $\mathbb{E}[\|\lambda\|_{2,2}^2] = \sum_{\ell=1}^\infty \gamma_\ell^2 \mathbb{E}[\|u_\ell\|_2^2] = \mathrm{tr}(\mathcal{C}_u) \sum_{\ell=1}^\infty \gamma_\ell^2 < \infty$. A few random joint kernels $\mathcal{C}_y$ by models I (upper row) and II (lower row) in Figure 2 illustrate the flexible prior candidates for the covariance $\mathcal{C}_y$ produced by gSTGP to capture complicated TESD (main diagonal blocks).

With all these definitions, we summarize the foregoing STGP models in the following unified form and illustrate their graphic structures in Figure 3.

$$y(\mathbf{z})|m, \mathcal{C}_{y|m} \sim \mathcal{GP}(m, \mathcal{C}_{y|m})$$
$$m(\mathbf{z}) \sim \mathcal{GP}(0, \mathcal{C}_m)$$

$$\begin{aligned}
\text{model 0 (baseline)}: \quad & \mathcal{C}_{y|m}^0 = \sigma_\varepsilon^2 \mathcal{I}_\mathbf{x} \otimes \mathcal{I}_t, & & \mathcal{C}_m^0 = \mathcal{C}_\mathbf{x} \otimes \mathcal{C}_t \\
\text{model I (reference)}: \quad & \mathcal{C}_{y|m}^\mathrm{I} = \sigma_\varepsilon^2 \mathcal{I}_\mathbf{x} \otimes \mathcal{I}_t, & & \mathcal{C}_m^\mathrm{I} = \mathcal{C}_{\mathbf{x}|t}(\lambda) \dot\otimes \mathcal{C}_t \\
\text{model II (proposed)}: \quad & \mathcal{C}_{y|m}^\mathrm{II} = \mathcal{C}_{\mathbf{x}|t}(\lambda) \dot\otimes \mathcal{I}_t, & & \mathcal{C}_m^\mathrm{II} = \mathcal{I}_\mathbf{x} \otimes \mathcal{C}_t \\
& \lambda_\ell(t) = \gamma_\ell u_\ell(t), & & u_\ell(\cdot) \overset{iid}{\sim} \mathcal{GP}(0, \mathcal{C}_u) \text{ for } \ell \in \mathbb{N}
\end{aligned} \tag{21}$$

2.4 Comparison of Two Models

Before concluding this section, we compare the generalized STGP models I (7) and II (10) and explain the superiority of model II compared with model I for learning TESD.

From the modeling perspective, model I puts much more weight on mean regression than on covariance learning (See Figure 3). The STGP model (21) can be equivalently viewed as decomposing the process into mean and residual: $y(\mathbf{z}) = m(\mathbf{z}) + \varepsilon(\mathbf{z})$ with $m(\mathbf{z}) \sim \mathcal{GP}(0, \mathcal{C}_m)$ and $\varepsilon(\mathbf{z}) \sim \mathcal{GP}(0, \mathcal{C}_{y|m})$. The residual process $\varepsilon(\mathbf{z})$ could possibly be inter-correlated in space over time but this is not correctly reflected with the likelihood kernel $\mathcal{C}_{y|m}^\mathrm{I} = \sigma_\varepsilon^2 \mathcal{I}_\mathbf{x} \otimes \mathcal{I}_t$ in

model I under i.i.d. assumption. To make it worse, with more and more data, the posterior of the covariance function may not contract to the true value as it becomes dominated by the likelihood not sufficiently modeled. On the other hand, model II with a balanced structure (Figure 3) between the prior kernel $\mathcal{C}_m^{\mathrm{II}}$ and the likelihood kernel $\mathcal{C}_{y|m}^{\mathrm{II}}$ achieves a good trade-off. Moreover, the posterior concentrates on the likelihood with the kernel structure $\mathcal{C}_{y|m}^{\mathrm{II}} = \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{I}_t$ that could correctly capture TESD given enough data (See Figure 1 and more numerical evidence in Section 5.1.2).

From the computing perspective, model II has significantly lower computational complexity than model I. As illustrated in Figure 1, discretizing $\mathcal{C}_y^{\mathrm{II}}$ leads to a highly sparse covariance matrix formed by a block-diagonal matrix $\mathbf{C}_{\mathbf{x}|t}$ and a sparse matrix $\mathbf{C}_t \otimes \mathbf{I}_{\mathbf{x}}$. On the contrary, the discretized $\mathcal{C}_y^{\mathrm{I}}$ is in general a dense matrix that involves intensive computation. Therefore, model II is more computationally efficient for learning TESD. See more details in Appendix B.3 and Section 5.1.2.

3. Theories

Now we study the theoretic properties of the time-dependent spatial kernel $\mathcal{C}_{\mathbf{x}|t}$ and the generalized STGP. The key elements are the spatial basis $\{\phi_\ell(\mathbf{x})\}$ and the dynamic eigenvalues $\{\lambda_\ell(t)\}$ (could be random) for which we will make a few more assumptions. Readers can skip or postpone reading this section without interruption.

3.1 Well-definedness

First we prove the well-definedness of kernels (17) and (18) in the Mercer's representation in the following theorem.

Theorem 1 *[Mercer's Kernels] Under Assumption 1, both $\mathcal{C}_m^I = \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}} \mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t$ and $\mathcal{C}_{y|m}^{II} = \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{I}_t$ are well defined non-negative definite kernels on $\mathcal{Z}$.*

Proof See Appendix A. ∎

With the Mercer's kernels (17) and (18), we could represent STGP in the spatial basis with random time-varying coefficients similarly as in the Karhunen-Loéve theorem (Fukunaga, 1990; Dong et al., 2006).

Theorem 2 *[Karhunen-Loéve Expansion] Under Assumption 1, STGP $f(\mathbf{x}, t) \sim \mathcal{GP}(0, \mathcal{C}_{\mathbf{z}})$ has the following representation of series expansion:*

$$f(\mathbf{x}, t) = \sum_{\ell=1}^{\infty} f_\ell(t)\phi_\ell(\mathbf{x}), \quad f_\ell(t) = \int_{\mathcal{X}} f(\mathbf{x}, t)\phi_\ell(\mathbf{x})d\mathbf{x} \tag{22}$$

where $\{f_\ell\}_{\ell=1}^{\infty}$ are random processes with mean functions $\mathbb{E}[f_\ell(t)] = 0$ and covariance functions as follows

- *if $\mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}} \mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t$, then $\mathbb{E}[f_\ell(t)f_{\ell'}(t')] = \lambda_\ell(t)\mathcal{C}_t(t, t')\lambda_\ell(t')\delta_{\ell\ell'}$.*

- *if $\mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{I}_t$, then $\mathbb{E}[f_\ell(t)f_{\ell'}(t')] = \lambda_\ell^2(t)\delta(t = t')\delta_{\ell\ell'}$.*

Proof See Appendix A. ∎

3.2 Regularity of Random Functions

For the convenience of discussion, we introduce the following general (k, s, p)-norm [1] to the infinite-sequence functions $\lambda = \{\lambda_\ell\}_{\ell=1}^\infty$ for $k, s > 0$ and $0 < p \le \infty$.

$$\|\lambda\|_{k,s,p} = \left(\sum_{\ell=1}^\infty \ell^{ks}\|\lambda_\ell\|_p^k\right)^{\frac{1}{k}} \tag{23}$$

And we denote the space $\ell^{k,s}(L^p(\mathcal{T})) := \{\lambda|\|\lambda\|_{k,s,p} < +\infty\}$. Note, the norm in Assumption 1 corresponds to the special case $k = 2, s = 0, p = 2$. For a given spatial basis $\{\phi_\ell(\mathbf{x})\}_{\ell=1}^\infty$, there is one-one correspondence $f(\mathbf{x}, t) \leftrightarrow \{f_\ell(t)\}_{\ell=1}^\infty$ in (22). Therefore, we could also define (k, s, p)-norm (23) for $f \in \ell^{k,s}(L^p(\mathcal{T}))$. Note, when $p = 2$, $f \in \ell^{k,s}(L^2(\mathcal{T}))$ with a fixed spatial basis $\{\phi_\ell(\mathbf{x})\}_{\ell=1}^\infty$ also implies $f \in \ell^{k,s}(L^2(\mathcal{Z}))$ regardless of spatial basis (normalized in $L^2(\mathcal{X})$) because $\left(\sum_{\ell=1}^\infty \ell^{ks}\|f_\ell(t)\phi_\ell(\mathbf{x})\|_2^k\right)^{\frac{1}{k}} = \left(\sum_{\ell=1}^\infty \ell^{ks}\|f_\ell(t)\|_2^k\|\phi_\ell(\mathbf{x})\|_2^k\right)^{\frac{1}{k}} = \|f\|_{k,s,2}$. For the rest of this section, we consider the case $k = p = 2$. In the following, notation $\lesssim$ ($\gtrsim$) means "smaller (greater) than or equal to a universal constant times".

If the dynamic eigenvalues λ decay in order $\kappa > 1$, the following proposition states that they fall in a subset of $\ell^2(L^2(\mathcal{T}))$.

Proposition 2 *Under Assumption 2-(20), $\lambda \in \ell^{2,s}(L^2(\mathcal{T}))$ in $\mathbb{P}$ for $s < (\kappa - 1)/2$.*

Proof It is straightforward to verify that

$$\mathbb{E}[\|\lambda\|_{2,s,2}^2] = \sum_{\ell=1}^\infty \ell^{2s}\mathbb{E}[\|\lambda_\ell\|_2^2] = \sum_{\ell=1}^\infty \ell^{2s}\gamma_\ell^2\mathbb{E}[\|u_\ell\|_2^2] \lesssim \text{tr}(\mathcal{C}_u)\sum_{\ell=1}^\infty \ell^{2s-\kappa} < \infty$$

if $2s - \kappa < -1$, i.e. $s < (\kappa - 1)/2$. ∎

To discuss the regularity of random functions drawn from STGP, we need the following assumptions on the spatial basis $\{\phi_\ell(\mathbf{x})\}_{\ell=1}^\infty$ and the dynamic eigenvalues λ.

Assumption 3 *We assume the spatial basis $\{\phi_\ell(\mathbf{x})\}_{\ell=1}^\infty$ are bounded in $L^\infty(\mathcal{X})$ and are Lipschitz with controlled growth rate in the Lipschitz constants $\text{Lip}(\phi_\ell)$:*

$$\sup_{\ell\in\mathbb{N}}\|\phi_\ell\|_\infty + \ell^{-1}\text{Lip}(\phi_\ell) \le C, \quad for\ some\ C > 0 \tag{24}$$

Define $Q_{\lambda,\mathcal{C}}(t, t') := \lambda^2(t)\mathcal{C}(t, t) - 2\lambda(t)\mathcal{C}(t, t')\lambda(t') + \lambda^2(t)\mathcal{C}(t, t)$. We need the following additional assumption on λ for the regularity of the full function $f(\mathbf{x}, t)$

$$\lambda \in \ell^{2,s}(L^\infty(\mathcal{T})), \quad \sup_{\ell\in\mathbb{N}}\ell^{-2}\sup_{t,t'\in\mathcal{T}}\frac{Q_{\lambda_\ell,\mathcal{C}}(t, t')}{\|\lambda\|_\infty^2|t - t'|^2} \le C, \quad for\ some\ C > 0 \tag{25}$$

---

[1] When $k = 2$, this is related to Sobolev norm in the frequency domain and Hilbert scales.

The following theorem states that the regularity of random function f in (22) depends on the decay rate s of dynamic eigenvalues λ.

Theorem 3 *[Regularity of Random Functions] Assume $\lambda \in \ell^{2,s}(L^2(\mathcal{T}))$. If $f(\mathbf{x}, t) \sim \mathcal{GP}(0, \mathcal{C}_\mathbf{z})$ as in Theorem 2, then $f = \sum_{\ell=1}^\infty f_\ell(t)\phi_\ell(\mathbf{x}) \in \ell^{2,s}(L^2(\mathcal{Z}))$ in probability.*

Moreover, under Assumption 3-(24), there is a version ^{2} $\tilde{f}(\mathbf{x})$ of $f(\mathbf{x}) := \int_\mathcal{T} f(\mathbf{x}, t)dt$ in $C^{0,s'}(\mathcal{X})$ for $s' < s$. If further $\mathcal{C}_\mathbf{z} = \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}} \mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t$ and $\{Q_{\lambda_\ell, \mathcal{C}_t}\}$ satisfies Assumption 3-(25), then there is a version $\tilde{f}(\mathbf{z})$ of $f(\mathbf{z})$ in $C^{0,s'}(\mathcal{Z})$ for $s' < s$.

Proof See Appendix A. ∎

Remark 3 *For random λ satisfying Assumption 2-(20), the above results still hold for $s < (\kappa - 1)/2$ by Proposition 2.*

Corollary 1 *If $f(\mathbf{x}, t) \sim \mathcal{GP}(0, \mathcal{C}_\mathbf{z})$ has a continuous version, then $\{f_\ell\}_{\ell=1}^\infty$ as in Theorem 2 are GP's defined on $\mathcal{T}$.*

Proof See Appendix A. ∎

3.3 Posterior Contraction

Now we consider the posterior properties. On the separable Banach space $(\mathbb{B} = \ell^2(L^2(\mathcal{T})), \|\cdot\|_{2,2})$, we consider a Gaussian random element λ satisfying Assumption 2-(20) and denote its associated reproducing kernel Hilbert space (RKHS) as $(\mathbb{H}, \|\cdot\|_\mathbb{H})$. We assume $\mathrm{tr}(\mathcal{C}_u) = 1$ by rescaling in (20). Then RKHS is $\mathbb{H} = \ell^{2,\kappa/2}(L^2(\mathcal{T}))$ with the following inner product and norm

$$\langle h, h' \rangle_\mathbb{H} = \sum_{\ell=1}^\infty \langle \gamma_\ell^{-1} h_\ell, \gamma_\ell^{-1} h'_\ell \rangle, \ \forall h, h' \in \mathbb{H}, \qquad \|\cdot\|_\mathbb{H} = \langle \cdot, \cdot \rangle_\mathbb{H}^{\frac{1}{2}} \tag{26}$$

Define the contraction rate of λ at λ_0 as follows

$$\varphi_{\lambda_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - \lambda_0\|_{2,2} \leq \varepsilon} \frac{1}{2} \|h\|_\mathbb{H}^2 - \log \Pi(\|\lambda\|_{2,2} < \varepsilon) \tag{27}$$

where Π is the prior measure on λ. Let p be a centered (assume $m \equiv 0$ for simplicity) Gaussian model, which is uniquely determined by its covariance $\mathcal{C}_{\mathbf{x}|t} = \sum_{\ell=1}^\infty \lambda_\ell^2(t)\phi_\ell \otimes \phi_\ell$. For a fixed spatial basis $\{\phi_\ell\}$, the model density p is parametrized by λ, hence denoted as p_λ. Let $n = I \wedge J$. Denote $P_\lambda^{(n)} := \bigotimes_{j=1}^n P_{\lambda,j}$ as the product measure on $\bigotimes_{j=1}^n (\mathcal{X}_j, \mathcal{B}_j, \mu_j)$. Each $P_{\lambda,j}$ has a density p_{λ_j} with respect to the σ-finite measure μ_j. Define the average Hellinger distance as $d_{n,H}^2(\lambda, \lambda') = \frac{1}{n} \sum_{j=1}^n \int (\sqrt{p_{\lambda,j}} - \sqrt{p_{\lambda',j}})^2 d\mu_j$. To bound the Hellinger distance between the modeling parameter λ and its true value λ_0, we make the following assumption.

---

^{2}A version/modification of stochastic process $\tilde{f}(\mathbf{x})$ of $f(\mathbf{x})$ means $\mathbb{P}[\tilde{f}(\mathbf{x}) = f(\mathbf{x})] = 1$ for $\forall \mathbf{x} \in \mathcal{X}$.

Assumption 4 *Let $\lambda \in \ell^{1,s}(L^\infty(\mathcal{T}))$ with some $s > 0$. Assume λ satisfy the following bounds*

$$c_\ell := \inf_{t \in \mathcal{T}} |\lambda_\ell(t)| \gtrsim \ell^{-s/2}, \quad C := \sup_{\ell \in \mathbb{N}} \|\lambda_\ell\|_\infty < +\infty \tag{28}$$

Denote the observations $Y^{(n)} = \{Y_j\}_{j=1}^n$ with $Y_j = y(\mathbf{X}, t_j)$. Note they are i.i.d. in model I and i.n.i.d. in model II conditioned on the mean. Now we prove the following posterior contraction about $\mathcal{C}_{\mathbf{x}|t}$ in model II, which generalizes Theorem 2.2 of Lan et al. (2020).

Theorem 4 *[Posterior Contraction of $\mathcal{C}_{\mathbf{x}|t}$ in model II] Let λ be a Borel measurable, zero-mean, tight Gaussian random element in $\Theta = \ell^2(L^2(\mathcal{T}))$ satisfying Assumption 4 and $P_\lambda^{(n)} = \bigotimes_{j=1}^n P_{\lambda,j}$ be the product measure of $Y^{(n)}$ parametrized by λ. If the true value $\lambda_0 \in \Theta$ is in the support of λ, and ε_n satisfies the rate equation $\varphi_{\lambda_0}(\varepsilon_n) \leq n\varepsilon_n^2$ with $\varepsilon_n \geq n^{-\frac{1}{2}}$, then there exists $\Theta_n \subset \Theta$ such that $\Pi_n(\lambda \in \Theta_n : d_{n,H}(\lambda, \lambda_{n,0}) > M_n \varepsilon_n | Y^{(n)}) \to 0$ in $P_{\lambda_{n,0}}^{(n)}$-probability for every $M_n \to \infty$.*

Proof See Appendix A. ∎

Remark 4 *Note the above posterior contraction theorem is for the covariance operator $\mathcal{C}_{\mathbf{x}|t}$ (characterized through the dynamic eigen-values λ), not for the mean function (which has already been extensively studied in Ghosal and van der Vaart (2017)). Therefore, the non-parametric model II of (21) for $\mathcal{C}_{\mathbf{x}|t}$ is not in the conjugate setting: because of the position of λ inside $\mathcal{C}_{\mathbf{x}|t}$, the likelihood for λ is not Gaussian.*

Remark 5 *The observations $Y^{(n)}$ in model I are conditional iid and the likelihood model is determined by $\mathcal{C}_{y|m} = \sigma_\varepsilon^2 \mathcal{I}_{\mathbf{x}} \otimes \mathcal{I}_t$ which does not contain λ, thus the posterior of λ in model I cannot contract to the correct value.*

Although the above theorem dictates that the posterior of λ contracts to the true value λ_0 at certain rate ε_n, it does not provide the details of ε_n. The following theorem specifies the contraction rate, which depends on the regularity of both the truth and the prior used.

Theorem 5 *[Posterior Contraction Rate of $\mathcal{C}_{\mathbf{x}|t}$ in model II] Let λ be a Gaussian random element satisfying Assumption 2-(20) and $\mathrm{tr}(\mathcal{C}_u) = 1$. The rest settings are the same as in Theorem 4. If the true value $\lambda_0 \in \ell^{2,s}(L^2(\mathcal{T}))$, then we have the rate of posterior contraction $\varepsilon_n = \Theta(n^{-(\frac{\kappa-1}{2} \wedge s)/\kappa})$.*

Proof See Appendix A. ∎

4. Posterior Inference and Predictions

Suppose we are given the spatiotemporal data $\mathcal{D} := \{\mathbf{Z}, \mathbf{Y}\}$, and there are K independent trials in the data $\mathbf{Y} := \{\mathbf{Y}_k\}_{k=1}^K$ with each $\mathbf{Y}_k = y(\mathbf{Z}) = [y(\mathbf{x}_i, t_j)]_{I \times J}$. Denote $\mathbf{M}_{I \times J} :=$

$m(\mathbf{Z})$, $\mathbf{C_M} := \mathcal{C}_m(\mathbf{Z}, \mathbf{Z})$ and $\mathbf{C_{Y|M}} := \mathcal{C}_{y|m}(\mathbf{Z}, \mathbf{Z})$. We truncate the Mercer's representation of $\mathcal{C}_{\mathbf{x}|t}$ for L terms. Consider the model (21) with hyper-parameters which are in turn given priors respectively, summarized as follows

$$\mathrm{vec}(\mathbf{Y}_k)|\mathbf{M}, \mathbf{C_{Y|M}} \sim \mathcal{N}(\mathrm{vec}(\mathbf{M}), \mathbf{C_{Y|M}}), \quad \text{i.i.d. for } k = 1, \cdots, K$$

$$m(\mathbf{z}) \sim \mathcal{GP}(0, \mathcal{C}_m), \quad \mathcal{C}_{\mathbf{x}|t}(\mathbf{z}, \mathbf{z}') = \sum_{\ell=1}^{L} \lambda_\ell(t)\lambda_\ell(t')\phi_\ell(\mathbf{x})\phi_\ell(\mathbf{x}')$$

$$\lambda_\ell(t) = \gamma_\ell u_\ell(t), \quad u_\ell(\cdot) \overset{iid}{\sim} \mathcal{GP}(0, \mathcal{C}_u) \text{ for } \ell = 1, \cdots, L \tag{29}$$

$$\mathcal{C}_* = \sigma_*^2 \exp(-0.5\| * - *' \|^s / \rho_*^s)$$

$$\sigma_*^2 \sim \Gamma^{-1}(a_*, b_*), \quad \log \rho_* \sim \mathcal{N}(m_*, V_*), \quad * = \mathbf{x}, t, \text{ or } u$$

where the likelihood kernel $\mathcal{C}_{y|m}$ and the prior kernel $\mathcal{C}_m$ are specified in (21). We adopt the Metropolis-Within-Gibbs scheme and use the slice samplers (Neal, 2003; Murray et al., 2010) for the posterior inference. More details can be found in Appendix B.

With the posteriors we can consider the following various prediction problems at new data points $(\mathbf{x}_*, t_*)$, $(\mathbf{x}, t_*)$ or $(\mathbf{x}_*, t)$:

$$m(\mathbf{x}_*, t_*)|\mathcal{D}, \quad m(\mathbf{x}_*, t)|\mathcal{D}, \quad m(\mathbf{x}, t_*)|\mathcal{D}, \quad \mathcal{C}_{\mathbf{x}|t_*}(\mathbf{x}, \mathbf{x}')|\mathcal{D}, \quad \mathcal{C}_{\mathbf{x}|t}(\mathbf{x}, \mathbf{x}_*)|\mathcal{D} \tag{30}$$

4.1 Prediction of Mean

We only consider the prediction $m(\mathbf{x}_*, t_*)|\mathcal{D}$ because the other two predictions $m(\mathbf{x}_*, t)|\mathcal{D}$, $m(\mathbf{x}, t_*)|\mathcal{D}$ are sub-problems of it. The prediction of mean has been well studied in the literature. The following proposition gives the predictive distribution of the mean function $m(\mathbf{x}, t)$ in our set-up.

Proposition 3 *Fit the spatiotemporal data* $\mathcal{D} = \{\mathbf{Z}, \mathbf{Y}\}$ *with the model* (29). *Then given a new point* $\mathbf{z}_* = (\mathbf{x}_*, t_*)$ *we have*

$$m(\mathbf{z}_*)|\mathcal{D} \sim \mathcal{N}(m', C')$$
$$m' = c_*^\mathsf{T}(\mathbf{C_M} + K^{-1}\mathbf{C_{Y|M}})^{-1}\overline{\mathbf{Y}}, \quad C' = C_{m_*} - c_*^\mathsf{T}(\mathbf{C_M} + K^{-1}\mathbf{C_{Y|M}})^{-1}c_*$$

where we denote

$$\overline{\mathbf{Y}} := \frac{1}{K}\sum_{k=1}^{K}\mathrm{vec}(\mathbf{Y}_k), \quad C_{m_*} := \mathcal{C}_m(\mathbf{z}_*, \mathbf{z}_*), \quad c_* := \mathcal{C}_m(\mathbf{Z}, \mathbf{z}_*), \quad c_*^\mathsf{T} := \mathcal{C}_m(\mathbf{z}_*, \mathbf{Z})$$

Proof See Appendix A. ∎

4.2 Prediction of Covariances

Now we consider two types of prediction for covariances of particular interest. The first one, $\mathcal{C}_{\mathbf{x}|t_*}(\mathbf{x}, \mathbf{x}')|\mathcal{D}$, evolves TESD among existing locations to other (future) time(s) t_*. The second one, $\mathcal{C}_{\mathbf{x}|t}(\mathbf{x}, \mathbf{x}_*)|\mathcal{D}$, extends TESD to new (neighboring) location(s) $\mathbf{x}_*$. Both

predictions have practical meaning and useful applications. For example, the former could predict how the brain connection evolves during some memory process, or in the progression of brain degradation of Alzheimer's disease. With the latter we could extend our knowledge of climate change from observed regions to unobserved territories.

Note, the prediction of TESD to new locations is exclusive to the proposed fully non-parametric model. The semi-parametric methods of dynamic covariance modeling (Wilson and Ghahramani, 2011; Fox and Dunson, 2015; Lan et al., 2020) with a covariance matrix (instead of a kernel) for the spatial dependence do not have this feature because the discrete spatial size (the size of covariance matrix) has been fixed.

4.2.1 Evolve Spatial Dependence to Future Time

Note from the definition (16), we know that $\mathcal{C}_{\mathbf{x}|t}$ is a function of dynamic eigenvalues $\{\lambda_\ell(t)\}$ with fixed spatial basis $\{\phi_\ell(\mathbf{x})\}$. Therefore, the prediction of the kernel $\mathcal{C}_{\mathbf{x}|t}$ in the time direction can be reduced to predicting $\lambda_\ell(t_*)|\mathcal{D}$ as follows. Denote $\boldsymbol{\lambda}_\ell := \lambda_\ell(\mathbf{t})$.

$$p(\lambda_\ell(t_*)|\mathcal{D}) = \int p(\lambda_\ell(t_*), \boldsymbol{\lambda}_\ell|\mathcal{D})d\boldsymbol{\lambda}_\ell = \int p(\lambda_\ell(t_*)|\boldsymbol{\lambda}_\ell)p(\boldsymbol{\lambda}_\ell|\mathcal{D})d\boldsymbol{\lambda}_\ell$$

where $p(\lambda_\ell(t_*)|\boldsymbol{\lambda}_\ell)$ is the standard GP predictive distribution. We can use the standard GP predictive mean and covariance to predict and quantify the associated uncertainty for $(\lambda_\ell(t_*)|\boldsymbol{\lambda}_\ell^{(s)})^2$ with $\boldsymbol{\lambda}_\ell^{(s)} \sim p(\boldsymbol{\lambda}_\ell|\mathcal{D})$, and then take average over all the posterior samples to get an approximation of $\lambda_\ell^2(t_*)|\mathcal{D}$. Therefore, $\mathcal{C}_{\mathbf{x}|t_*}(\mathbf{x}, \mathbf{x}')|\mathcal{D}$ can be obtained/approximated by substituting $\lambda_\ell^2(t)$ with $\lambda_\ell^2(t_*)|\mathcal{D}$ in (16)

$$\mathcal{C}_{\mathbf{x}|t_*}(\mathbf{x}, \mathbf{x}')|\mathcal{D} = \sum_{\ell=1}^{\infty}(\lambda_\ell^2(t_*)|\mathcal{D})\,\phi_\ell(\mathbf{x})\phi_\ell(\mathbf{x}')$$

$$\approx \frac{1}{S}\sum_{s=1}^{S}\sum_{\ell=1}^{L}(\lambda_\ell(t_*)|\boldsymbol{\lambda}_\ell^{(s)})^2\phi_\ell(\mathbf{x}; \eta_{\mathbf{x}}^{(s)})\phi_\ell(\mathbf{x}'; \eta_{\mathbf{x}}^{(s)}), \quad \boldsymbol{\lambda}_\ell^{(s)} \sim p(\boldsymbol{\lambda}_\ell|\mathcal{D})$$

where $\lambda_\ell(t_*)|\boldsymbol{\lambda}_\ell^{(s)} = \mathcal{C}_u(t^*, \mathbf{t})\mathcal{C}_u(\mathbf{t}, \mathbf{t})^{-1}\boldsymbol{\lambda}_\ell^{(s)}$.

4.2.2 Extend Evolution of Spatial Dependence to Neighbors

Recall that in the definition of $\mathcal{C}_{\mathbf{x}|t}$, the fixed basis $\{\phi_\ell(\mathbf{x})\}$ is taken from the eigenfunctions of the spatial kernel $\mathcal{C}_{\mathbf{x}}$. To extend $\mathcal{C}_{\mathbf{x}|t}$ as a function of time to other locations based on existing knowledge informed by data, one could predict the basis at a new position, namely $\phi_\ell(\mathbf{x}_*)$, using its known values $\boldsymbol{\phi}_\ell := \phi_\ell(\mathbf{X})$ as in the conditional Gaussian:

$$\phi_\ell(\mathbf{x}_*)|\boldsymbol{\phi}_\ell = \mathcal{C}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{X})\mathcal{C}_{\mathbf{x}}(\mathbf{X}, \mathbf{X})^{-1}\boldsymbol{\phi}_\ell = \mathcal{C}_{\mathbf{x}}(\mathbf{x}_*, \mathbf{X})\lambda_\ell^{-2}\boldsymbol{\phi}_\ell, \quad \forall \ell = 1, \cdots, L$$

Then, $\mathcal{C}_{\mathbf{x}|t}(\mathbf{x}, \mathbf{x}_*)|\mathcal{D}$ can be predicted/approximated by substituting $\phi_\ell(\mathbf{x}')$ with $\phi_\ell(\mathbf{x}_*)|\boldsymbol{\phi}_\ell$ in (16) as follows

$$\mathcal{C}_{\mathbf{x}|t}(\mathbf{x}, \mathbf{x}_*)|\mathcal{D} = \sum_{\ell=1}^{\infty}\lambda_\ell^2(t)\phi_\ell(\mathbf{x})(\phi_\ell(\mathbf{x}_*)|\mathcal{D})$$

$$\approx \frac{1}{S}\sum_{s=1}^{S}\sum_{\ell=1}^{L}(\lambda_\ell^{(s)}(t))^2\phi_\ell(\mathbf{x}; \eta_{\mathbf{x}}^{(s)})(\phi_\ell(\mathbf{x}_*)|\boldsymbol{\phi}_\ell), \quad \lambda_\ell^{(s)}(t) \sim p(\lambda_\ell(t)|\mathcal{D})$$

5. Numerical Experiments

In this section, we compare the proposed generalized STGP models (21) against a complete spectrum of models with different combinations in stationarity and separability including:

0. classical stationary separable (stat-sep) model (2);

1. stationary non-separable model (Gneiting, 2002):

$$\text{stat-nonsep}: \qquad C(\mathbf{z}, \mathbf{z}') = \frac{\sigma^2}{(a|t - t'|^{2\alpha} + 1)^{\beta d/2}} \exp\left(-\frac{c\|\mathbf{x} - \mathbf{x}'\|^{2\gamma}}{(a|t - t'|^{2\alpha} + 1)^{\beta\gamma}}\right) \quad (31)$$

where for simplicity we fix $a = d = 1$, $\alpha = \gamma = \frac{1}{2}$, $\beta = 2$, and let $c = \frac{1}{2\rho}$.

2. non-stationary separable model (Paciorek and Schervish, 2003):

$$\text{nonstat-sep}: \qquad C(\mathbf{z}, \mathbf{z}') = \sigma^2 A(\mathbf{x}, \mathbf{x}') \exp\left(-Q(\mathbf{x}, \mathbf{x}')\right) * A(t, t) \exp\left(-Q(t, t')\right) \quad (32)$$

where $A(r, r') = |\Sigma_r(\rho)|^{\frac{1}{4}} |\Sigma_{r'}(\rho)|^{\frac{1}{4}} |(\Sigma_r(\rho) + \Sigma_{r'}(\rho))/2|^{-\frac{1}{2}}$ and $Q(r, r') = (r - r')^{\mathsf{T}}[(\Sigma_r(\rho) + \Sigma_{r'}(\rho))/2]^{-1}(r - r')$ with r being $\mathbf{x}$ or t. $\Sigma_r(\rho)$ is the covariance of Gaussian kernel centered at r with correlation length ρ.

3. non-stationary non-separable model (Wang et al., 2020):

$$\text{nonstat-nonsep}: \qquad \begin{aligned} C(\mathbf{z}, \mathbf{z}') =& \sigma^2 A(\mathbf{x}, \mathbf{x}') A(t, t) \left(1 + Q(\mathbf{x}, \mathbf{x}') + Q(t, t')\right)^{s_0} \cdot \\ & \left(1 + Q(\mathbf{x}, \mathbf{x}')\right)^{s_1} \left(1 + Q(t, t')\right)^{s_2} \end{aligned} \quad (33)$$

Table 1 lists all the models for comparison with their authors and properties.

Model	Author(s)	Non-stationary	Non-separable	Non-parametric	Sparse
stat-sep	–	✗	✗	✓(mean) ✗(covariance)	✗
stat-nonsep	Gneiting (2002)	✗	✓	✓(mean) ✗(covariance)	✗
nonstat-sep	Paciorek and Schervish (2003)	✓	✗	✓(mean) ✗(covariance)	✗
nonstat-nonsep	Wang et al. (2020)	✓	✓	✓(mean) ✗(covariance)	✗
qKron-prod (I)	S.Lan	✓	✓	✓(mean) ✓(covariance)	✗
qKron-sum (II)	S.Lan	✓	✓	✓(mean) ✓(covariance)	✓

Table 1: Spatiotemporal models for comparison.

We evaluate the performance of all the above models in fitting and predicting mean and covariance as functions of time using a simulated spatiotemporal process and an analysis of real brain imaging data. As a Bayesian non-parametric method for covariance modeling, model II with the quasi Kronecker sum structure (qKron-sum) is shown to be the best in characterizing TESD. All the numerical codes are publicly available at GitHub repository.

5.1 Simulation

In this section, we study a simulated example of non-stationary and non-separable spatiotemporal process. One can find a similar study for a simulated stationary and non-separable process in Appendix C.1.
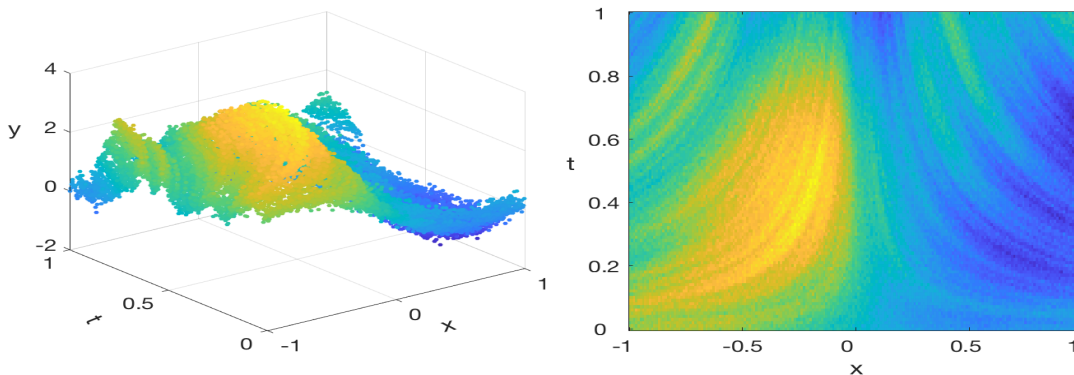
Figure 4: Simulated spatiotemporal data over $[-1,1] \times [0,1]$, viewed in 3d (left) and projected in space-time domains (right).

5.1.1 DATA GENERATION

We consider the following spatiotemporal process with a non-stationary, non-separable covariance for spatial dimension $d = 1$:

$$
\begin{aligned}
y(x,t) &\sim \mathcal{GP}(m, \mathcal{C}_y), \quad x \in \mathcal{X} = [-1,1], \ t \in \mathcal{T} = [0,1] \\
m(x,t) &= \cos(\pi x) \sin(2\pi t) \\
\mathcal{C}_y^{\mathrm{ns}}(\mathbf{z}, \mathbf{z}') &= \exp\left(-\frac{|x - x'|^2}{2\ell_x} - \frac{|t - t'|^2}{2\ell_t} - \frac{|xt - x't'|}{2\ell_{xt}} \right) + \sigma_\varepsilon^2 \delta(\mathbf{z} = \mathbf{z}')
\end{aligned}
\tag{34}
$$

To generate observations, we discretize the domain by dividing $\mathcal{X}$ into $N_x = 200$ equal subintervals and $\mathcal{T}$ into $N_t = 100$ equal subintervals. Setting $\ell_x = 0.5$, $\ell_t = 0.3$, $\ell_{xt} = \sqrt{\ell_x \ell_t} \approx 0.39$ and $\sigma_\varepsilon^2 = 10^{-2}$, we generate 20301 data points $\{y_{ij}\}$ over the mesh grid. Such random process can be repeated for K trials and we plot one of them in Figure 4.

5.1.2 MODEL FIT

For simplicity we use a subset of these 20301 data points taken on an equally spaced submesh with $I = 5$ and $J = 101$. Now we fit the data with STGP models (21) respectively. We set $\gamma_\ell = \ell^{-\kappa/2}$ with $\kappa = 1.2$, $a = [1,1,1]$, $m = [0,0,0]$ for all models; $b = [5,10,10]$, $V = [0.1, 0.1, 0.01]$ for models 0 and I and $b = [0.1, 1, 5]$, $V = [1,1,1]$ for model II. The truncation number of Mercer's kernel expansion is set to $L = I = 5$. For each experiment, we run MCMC to collect 2.4×10^4 samples, burn in the first 4000, and subsample every other. The resulting 10^4 posterior samples are used to estimate the mean function $m|\mathbf{Y}$ and the covariance function $\mathcal{C}_{y|t}|\mathbf{Y}$. For other parametric covariance models (the first four) listed in Table 1 we let $\sigma^2 \sim \Gamma(a', b')$ and $\log \rho \sim \mathcal{N}(m', V')$. Though not originally designed to learn TESD, these models are set in the framework of (29) with the same hyper-parameter setting as model I. We want to test these models in recovering mean function $m(\mathbf{x}, t)$ and more importantly learning TESD $\mathcal{C}_{y|t}$ which has the following truth:

$$
C_{y|t}^{\mathrm{ns}}(x, x') := \mathrm{Cov}[y(x,t), y(x',t)] = \exp\left(-\frac{|x - x'|^2}{2\ell_x} - \frac{|x - x'|t}{2\ell_{xt}} \right) + \sigma_\varepsilon^2 \delta(x = x')
\tag{35}
$$

(a) Stat-sep model

(b) Stat-nonsep model

(c) Nonstat-sep model

(d) Nonstat-nonsep model

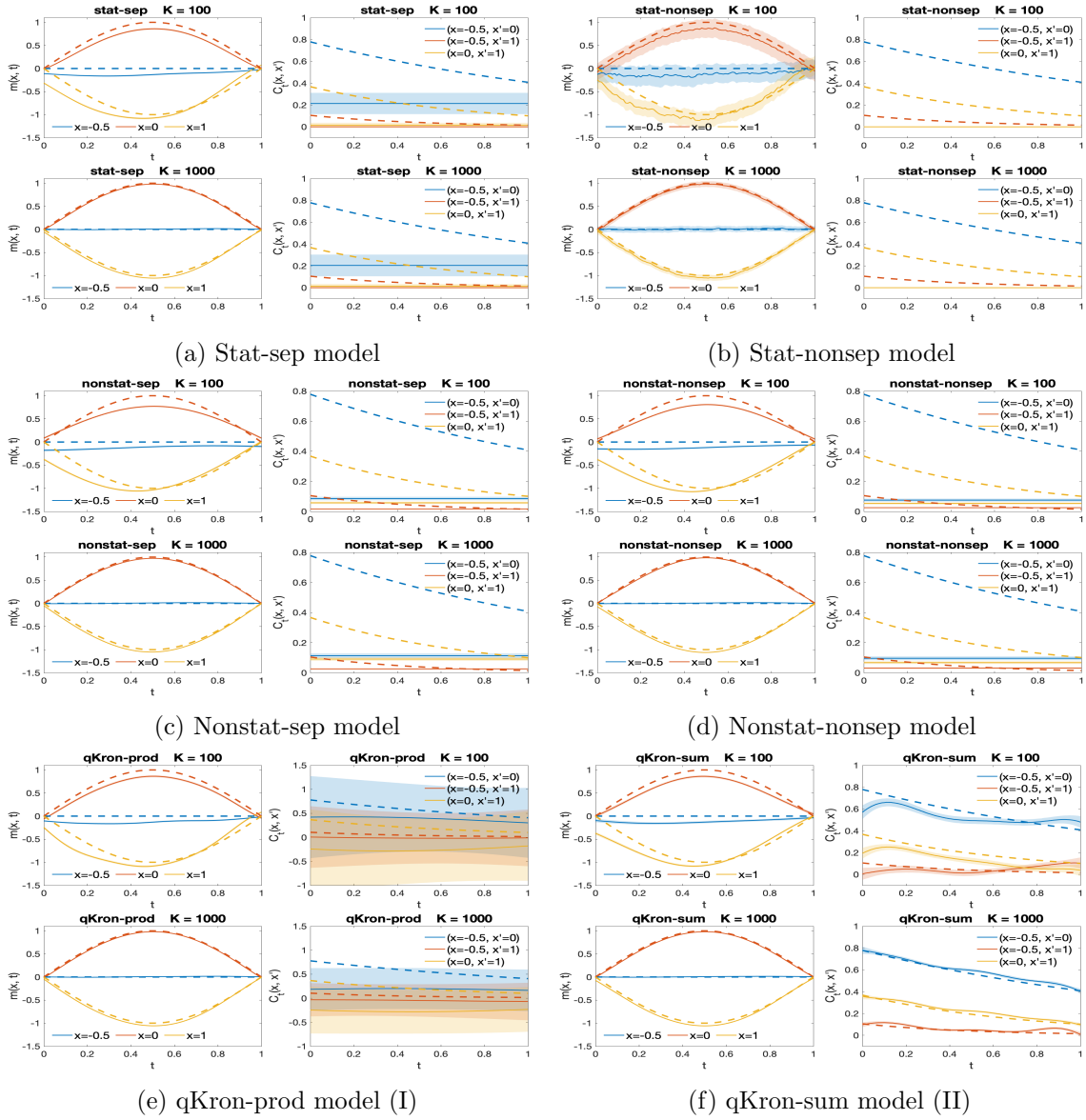(e) qKron-prod model (I)

(f) qKron-sum model (II)

Figure 5: A simulated process: selective mean functions $m(x, t)$ (left column) and covariance functions $C_{y|t}^{ns}(x, x')$ (right column) fitted by various models with $K = 100$ trials of data (upper row) and $K = 1000$ trials of data (lower row) on each panel. Dashed lines are true values, solid curves are estimates with shaded credible regions indicating their uncertainty.

The posterior estimates by different models are plotted at selective locations in Figure 5. With growing data information (increasing trial number K) [3], most posterior estimates contract towards some values (with decreasing width of credible bands), especially for mean

---

[3]More repeated trials (K) can be viewed as increasing data (n) though they are stacked over the same discrete time points $\{t_j\}_{j=1}^J$. Better contraction (closer to truth with better credible band coverage) can be achieved by increasing J distinct time points (as increasing N in Figure 5 of Lan et al., 2020).

functions (left column of each panel). However, most of them fail to reconstruct correct TESD (right column of each panel). According to Proposition 1, model 0 yields static TESD, and the result by model I is inevitably pulled towards the wrong constant σ_ε^2 with increasing data due to the poorly structured likelihood kernel $\mathcal{C}_{y|m}^{\mathrm{I}}$. Model II, by contrast, characterizes TESD with reasonable accuracy (Panel 5f). Although all the parametric covariance models recover the true mean functions, none of them generates correct time-varying TESD, even it is the non-stationary non-separable (Wang et al., 2020).

We repeat the experiments for 10 times and summarize their model performance in terms of mean squared errors (MSE) in Table 2. As we can see, model II renders comparable MSE in estimating mean functions as other models, but $2 \sim 3$ orders of magnitude smaller MSE for estimating TESD, using much less time.

	$K = 100$			$K = 1000$		
Model	mean	TESD	time	mean	TESD	time
stat-sep	2.17e-2 (1.2e-5)	0.143 (2e-5)	2410 (230)	1.14e-3 (9.3e-7)	0.143 (2e-5)	2450 (520)
stat-nonsep	2.18e-2 (1.1e-5)	0.142 (1.9e-05)	2340 (350)	1.15e-3 (6.5e-7)	0.142 (1.3e-5)	2330 (310)
nonstat-sep	2.34e-2 (1.1e-5)	0.111 (2.5e-05)	1770 (200)	1.1e-3 (8.6e-7)	9.52e-2 (1.9e-5)	1840 (250)
nonstat-nonsep	2.27e-2 (1.4e-5)	0.114 (2.1e-05)	3230 (410)	1.13e-3 (7.9e-7)	0.105 (2.4e-5)	3240 (420)
qKron-prod (I)	2.19e-2 (8.7e-4)	0.15 (0.11)	3850 (1700)	1.14e-3 (6.1e-7)	0.189 (2.7e-2)	3760 (1400)
qKron-sum (II)	2.06e-2 (2.6e-5)	**5.14e-3** (9.5e-5)	**1720** (160)	1.13e-3 (1.5e-6)	**3.85e-4** (4.1e-5)	**2380** (890)

Table 2: Mean squared errors (MSE) for fitting mean $m(x,t)$ and TESD $C_{y|t}^{\mathrm{ns}}(x,x')$. Table values are median estimates of 10 repeated results with sample standard deviation in the parentheses.

5.1.3 MODEL PREDICTION

Now we consider the predictions described in Section 4. We train the models based on 85% of the data (shorter gray ticks) and hold out the rest 15% (black dash-dot ticks) for testing as illustrated in Figure 6 that plots results only for models I (qKron-prod) and II (qKron-sum). We first compare the mean prediction in the time direction, i.e. $m(x, t_*)$, which is also a well studied object. In Figure 6, thick dash-dot lines are the predicted values of the mean function with dark shaded regions as the corresponding credible bands. In the left columns of these two panels, we see that both models give comparable results in predicting mean functions, with better accuracy for more data (trials K).

Next, we consider the prediction of covariance functions, particularly in two types, namely TESD to future (Section 4.2.1) and TESD to neighbor (Section 4.2.2). The middle columns on the two panels of Figure 6 show the first type of covariance prediction, i.e. $\mathcal{C}_{y|t_*}(x, x)$. The predicted TESD in time follows the trend of the fitted values with model II significantly better than model I (closer to truth with smaller credible bands). Note there is higher uncertainty (wider credible bands) when such prediction comes into 'no-data' (extrapolation) zone compared with that in the interpolation zone where there are still nearby training points. Lastly, we compare the second type of covariance prediction, $\mathcal{C}_{y|t}(x, x_*)$, TESD to a new neighbor $x_* = 0.1$ in the right columns on two panels of Figure 6. Note, we do not have any data at this location $x_* = 0.1$, yet model II can still extend TESD (thick dash-dot lines) to new locations with decent precision, in reference to the truth (dash lines). On the contrary, model I yields less desirable results with much higher uncertainty.
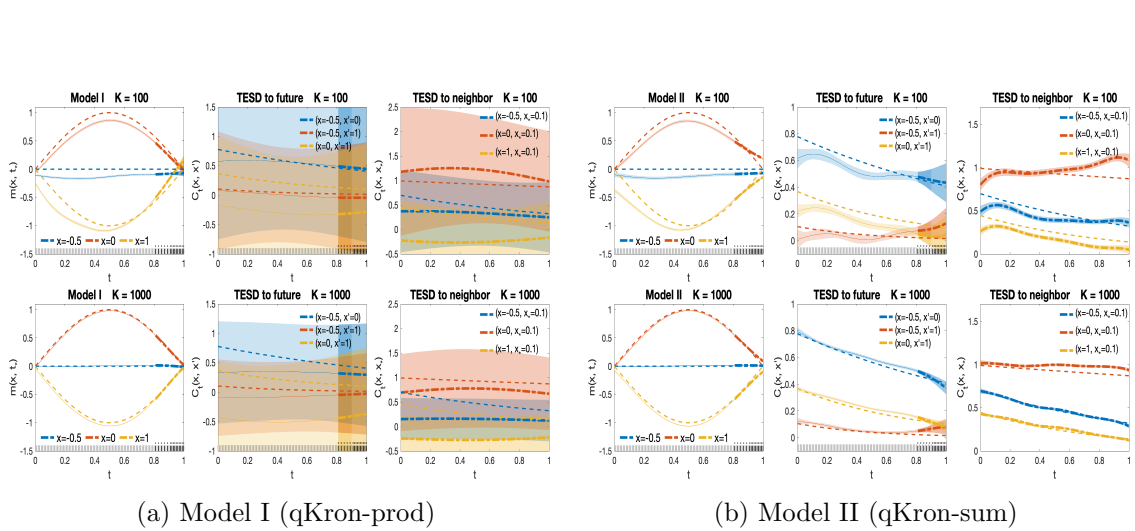
(a) Model I (qKron-prod) (b) Model II (qKron-sum)

Figure 6: Prediction of mean functions $m(x, t_*)$ (left column), TESD at future times t_*, $C_{y|t_*}^{ns}(x, x')$ (middle column), and at neighboring locations x_*, $C_{y|t}^{ns}(x, x_*)$ (right column), by (a) model I and (b) model II based on $K = 100$ trials (upper row) and $K = 1000$ trials (lower row) of data on each panel. Dashed lines are true values. Solid lines with light shaded regions are results (estimates with credible bands) of the models trained on data indicated by short gray ticks. Thick dash-dot lines with dark shaded regions are results (predictions with credible bands) of the models on testing data indicated by black dash-dot ticks.

We extend the experiments to all models and repeat each for 10 times. Table 3 summarizes their mean squared prediction errors (MSPE). Again we find model II generating significantly smaller MSPE for TESD using less time compared with the other models. All the numeric evidences strongly support that model II is both effective and efficient in modeling and predicting TESD.

Trials	Model	mean	TESD to future	TESD to neighbor	time
	stat-sep	5.09e-3 (1.4e-6)	7.05e-2 (1.6e-5)	0.308 (2.4e-5)	2280 (140)
	stat-nonsep	2e-2 (3.4e-6)	7e-2 (1.2e-3)	0.307 (2.2e-3)	2100 (290)
	nonstat-sep	1.63e-2 (1.4e-5)	5.07e-2 (6.2e-4)	0.247 (2.6e-4)	1600 (170)
$K = 100$	nonstat-nonsep	1.31e-2 (7.7e-6)	5.21e-2 (7.7e-4)	0.257 (2.3e-4)	2820 (370)
	qKron-prod (I)	8.32e-3 (7.3e-4)	0.109 (0.68)	0.327 (0.66)	4170 (770)
	qKron-sum (II)	8.65e-3 (2e-4)	**1.04e-2** (0.34)	**7.06e-3** (1.3e-4)	**1800** (96)
	stat-sep	7.95e-4 (3.1e-7)	7.05e-2 (2.2e-5)	0.308 (2.4e-5)	2290 (430)
	stat-nonsep	1.26e-2 (9.2e-7)	6.95e-2 (8.4e-4)	0.308 (2.4e-3)	2150 (290)
	nonstat-sep	7.56e-4 (4.5e-7)	3.5e-2 (4e-4)	0.212 (6.4e-5)	1670 (210)
$K = 1000$	nonstat-nonsep	1.45e-3 (9.3e-7)	4.23e-2 (1.2e-3)	0.238 (2.7e-4)	2820 (370)
	qKron-prod (I)	4.18e-4 (5.7e-6)	0.131 (2.4e-2)	0.187 (0.17)	3380 (1200)
	qKron-sum (II)	1.6e-3 (1.3e-5)	**4.39e-3** (1.3e-2)	**1.48e-3** (4.3e-5)	**2310** (810)

Table 3: Mean squared prediction errors (MSPE) for predicting mean $m(x, t_*)$ and TESD's $C_{y|t_*}^{ns}(x, x)$ and $C_{y|t}^{ns}(x, x_*)$. Table values are median estimates of 10 repeated results with sample standard deviation in the parentheses.

5.2 Longitudinal Analysis of Alzheimer's Brain Images

In this section we will apply the generalized STGP model to Alzheimer's neuroimaging data to study the association of brain regions in the progression of this disease. Since other models fail to characterize TESD and involve prohibitive computation in this example, we will mainly focus on model II in the following unless stated otherwise.

Alzheimer's disease (AD) is a chronic neurodegenerative disease that affects patients' brain functions including memory, language, orientation, etc. in the elder population generally above 65. According to the World Alzheimer Report (Report, 2018), there were about 50 million people worldwide living with dementia in 2018, and this figure is expected to skyrocket to 132 million by 2050. Yet the cause of AD is poorly understood. Longitudinal studies have collected high resolution neuroimaging data, genetic data and clinical data in order to better understand the progress of brain degradation. In this section, we analyze the positron emission tomography (PET) brain imaging data from the Alzheimer's Disease Neuroimaging Initiative project (ADNI, 2003) with two aims: (i) to characterize the change of the brain structure and function over time; and (ii) to detect the spatial correlation between brain regions and describe its temporal evolution (TESD).

5.2.1 POSITRON EMISSION TOMOGRAPHY (PET) DATA

We obtain PET scans scheduled at the baseline, 6 months, 1 year, 18 months, 2 years and 3 years from the ADNI study. There are 51 subjects in this data set, with 14 Cognitively Normal (CN), 27 Mild Cognitive Impairment (MCI) and 19 Alzheimer's Disease (AD). Among these patients, only the MCI group has data at 18 months and the AD group is followed up until 2 years. PET brain image scans are obtained and processed by coregsitering to have the same position, averaging over 6 five-minute frames, standardizing to $160 \times 160 \times 96$ voxel image grid, and smoothing to have uniform resolution. A detailed description of PET protocol and acquisition can be found at http://adni.loni.usc.edu.

We focus on a (48-th) slice in the middle (horizontal section) and model the images of size 160×160. For each subject k at a specific time point t during the study, the response function $y_k(x, t)$ represents the pixel value of location x in the image being read. Therefore the discrete data $\{y_{ijk}\}$ have dimension $I \times J \times K$, with $I = 160^2 = 25600$, $J \in \{5, 6, 4\}$ and $K \in \{14, 27, 19\}$. To study the spatial dependency in these brain images, we need a kernel with discrete size 25600×25600, which is enormous if it is a dense matrix. We introduce a spatial kernel based on the graph Laplacian (Shen et al., 2010; Ng et al., 2012; Hu et al., 2015; Huang et al., 2018; Dunlop et al., 2020). The resulting precision matrix is highly sparse (with 0.035% non-zero entries) and thus amenable to an efficient learning of TESD. See more details in Appendix B.4.

5.2.2 MODEL FIT

Now we fit model II (qKron-sum) with the graph-Laplacian based spatial kernel (B.4) to the PET brain images. The following setting for hyper-parameters is used: $a = [1, 1, 1]$, $b = [0.1, 1, 0.1]$, $m = [0, 0, 0]$ and $V = [0.1, 1, 1]$; however the results are not sensitive to the setting. The smoothness of actual time-varying spatial dependence in the brain regions is unknown, thus it is difficult to specify a prior that matches the regularity of the truth. Therefore we choose $\kappa = 0$ in the prior model (20)-(16) for $\mathcal{C}_{\mathbf{x}|t}$. It results in an improper
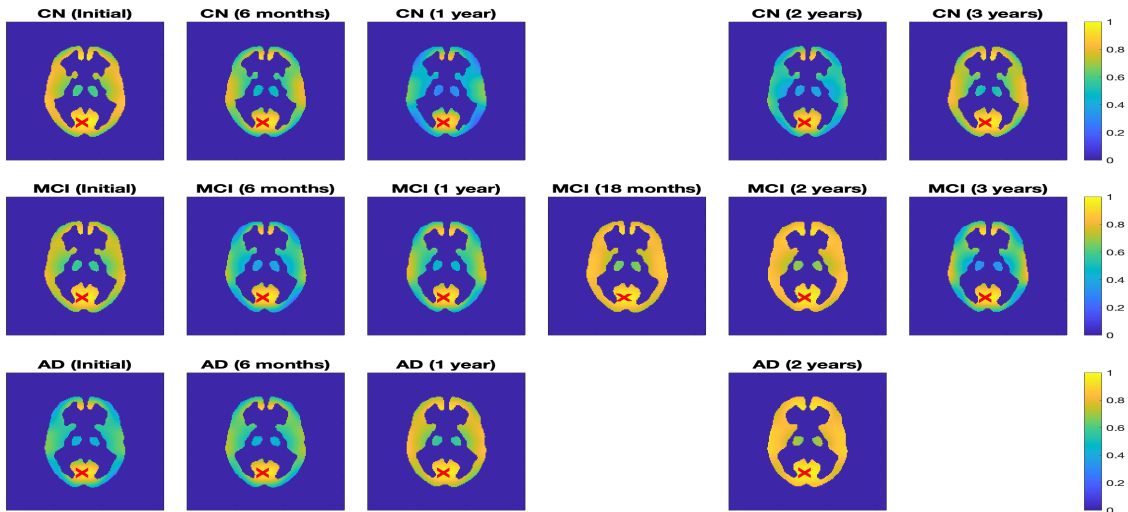
Figure 7: Estimated correlation between the brain region of interest (ROI) and a selected point of interest (POI, red cross) for CN (top row), MCI (middle row) and AD (bottom row) respectively by the proposed model II (qKron-sum). The color at a point in ROI indicates the correlation between that point and the selected POI in the scale of $[0, 1]$.

prior, however regularized by the likelihood (See more details in Figure C.3 and Appendix C.2). Smoother (and more informative) priors tend to blur TESD found here (results not shown). The truncation number of the Mercer's kernel expansion is set to $L = 100$. We run MCMC to collect 2.4×10^4 samples, burn in the first 4000, and subsample every other. The resulting 10^4 samples are used to obtain posterior estimates of mean functions $\mathbf{M}(t)$ and covariance functions $\mathbf{C}_y(t)$. Figure C.1 shows the fitted brain images at 6 scheduled times. The estimated brain images of patients in the control group (CN) have higher pixel values than the other two groups with bigger (blue) hollow regions. This can be seen more clearly from the summary of their estimated pixel values in Figure C.2.

Next, we investigate TESD in the PET brain images. TESD over this discretize field is a matrix valued function of time t. At each time the spatial covariance matrix is of size 25600×25600, too big for a direct visualization. Instead, we examine one row of the spatial correlation matrix by selecting a region of interest (ROI) (chosen based on the pixel values above the 83.5% quantile) and a point of interest (POI) (based on the overall variance, marked as red cross) in the occipital lobe. Figure 7 plots the ROI-POI correlations across time. As these spatial correlations evolve with time, the POI is highly correlated to its nearest region across all the time. It is also interesting to note the high correlation between the POI and some area in the frontal lobe.

To better summarize TESD as full-size covariance/correlation matrices, we threshold the (25600^2) absolute correlation values at the top 10%, then we obtain the adjacency matrix (25600×25600) based on the nonzero values of the correlations. Finally we define the *graph connection* as the diagonal of the degree matrix (row sums of the adjacency) projected back to 160×160 mesh. Therefore, the value of each point on the graph connection indicates how many nodes are connected to it. Figure 8 plots the graph connections of the brain ROI
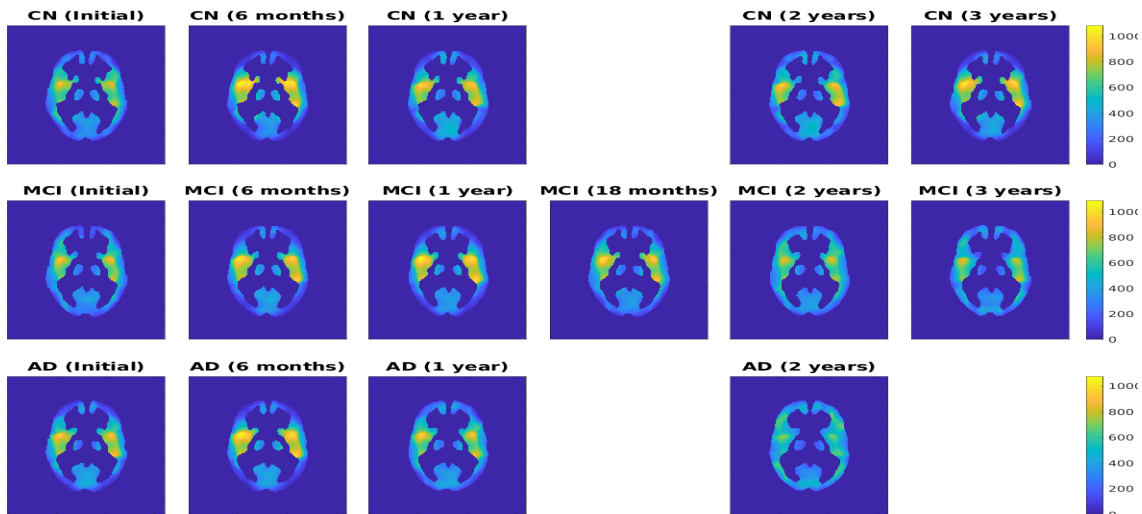
Figure 8: Estimated graph connection of the brain images for CN (top row), MCI (middle row) and AD (bottom row) respectively by the proposed model II (qKron-sum). The color at each point represents the number of nodes connected to it, indicating the brain activeness.

for different groups. For each of these graph connections in Figure 8, the truncation at any value yields a network of connected nodes that are the most active. As seen from Figure 8, these networks are most likely to concentrate on certain region in the temporal lobe. We successfully characterize the dynamic changing of such connectivity network of in these brain images. Note that the connectivity becomes weaker (thus the network of connected nodes become smaller) in the later stage for the MCI group (2 and 3 years) and the AD group (2 years), which could serve as an indicator of brain degradation.

To compare the generalized STGP models with other spatiotemporal models with parametric covariance listed in Table 1, we have to reduce the image size to 40×40 for learning TESD (spatial covariance of size 1600×1600) with available computing resources. Note that qKron-sum is the only sparse model. All the other dense models would otherwise require more than 200 GB memory for the original spatial covariance of size 25600×25600 and take more than 1 week to gather only a few thousands of samples even on GPU.

Working with the coarsen images (by sub-sampling image pixels) for the AD patients, we conduct the similar comparison as in Section 5.1.2. For each model, we collect 1.2×10^4 samples, burn in the first 2000, and subsample every 5. The resulted 2000 samples are used to obtain posterior estimates. Unlike simulation, there is no true mean or covariance in this example. Therefore, we compare the log-likelihood, weighted MSE and time consumption in Table 4. The values of log-likelihood for different models are not very comparable because the log-likelihood is dominated by the log-determinant of the likelihood kernel, which scales linearly with the problem size (IJK). We thus consider the weighted MSE, the positive quadratic form of log-likelihood, which is also interpreted as the Mahalanobis distance between actual image and the fitted mean, weighted by the likelihood kernel. The proposed STGP model qKron-sum attains the lowest weighted MSE score using the least amount of time (almost 2% of that for stat-nonsep and nonstat-nonsep models).

23

Model	log-likelihood	weighted MSE	Time (seconds)
stat-sep	247059.02 (13.04)	59075.74 (247.80)	57650.46
stat-nonsep	247359.03 (16.73)	57731.13 (241.00)	270202.49
nonstat-sep	143714.69 (63.98)	60663.36 (238.87)	177338.31
nonstat-nonsep	179503.71 (151.24)	60490.79 (234.76)	254509.56
qKron-prod (I)	97405.81 (45591.06)	75654.64 (58212.91)	163502.18
qKron-sum (II)	9079.57 (64.19)	**4545.54 (75.61)**	**5087.06**

Table 4: Log-likelihood, weighted mean squared error (MSE) and time for AD by various models. Table values are median estimates of 2000 posterior samples with sample standard deviation in the parentheses.

To investigate the TESD results output by different spatiotemporal models, we plot ROI-POI correlations (as in Figure 7) in Figure C.5 and graph connections (as in Figure 8) in Figure 9 by various models only for the AD patients. Only the proposed qKron-sum model captures time-varying spatial correlations. Note we observe the same brain degradation in the AD group as in Figure 8 illustrated by decreasing brain connection.

5.2.3 MODEL PREDICTION

Next we hold out the data at the last time point for testing. For each group, the generalized STGP model II (qKron-sum) is built based on the rest of the data. Then we predict the mean and covariance functions of the brain image at the held-out time point. Figure C.6 compares the actual individuals' brain images (upper row) with the predicted brain images (lower row) at the last time point. We can see that the prediction reflects the basic feature of the brain structure in each group. Next, Figure 10 plots the correlation (TESD) between the brain ROI and the selected POI (marked as red cross) predicted at the last time point. Note that the POI is less correlated to the middle region (thalamus), especially for the AD group. This is consistent with the fitted results shown in Figure 7. Readers can find more numerical results of the neuroimaging analysis in Appendix C.2.

6. Conclusion

In this paper, we generalize the separable STGP to model TESD in spatiotemporal data. Instead of treating the space variable $\mathbf{x}$ and the time variable t as a joint variable $\mathbf{z} = (\mathbf{x}, t)$, we introduce time-dependence to the spatial kernel by varying its eigenvalues in the Mercer's representation and propose a novel Bayesian nonparametric non-stationary non-separable model for covariance learning. Theoretic properties of such time-dependent spatial kernel, including the convergence, the regularity of random prior draws and the posterior contraction, have been systematically investigated. We construct and compare two joint kernels with quasi Kronecker product and sum structures respectively. The latter is highly sparse and well structured for learning TESD. The advantage of the proposed (quasi Kronecker sum) model is demonstrated by a simulation study of spatiotemporal process. It is then applied to analyze PET brain images of Alzheimer's patients to describe
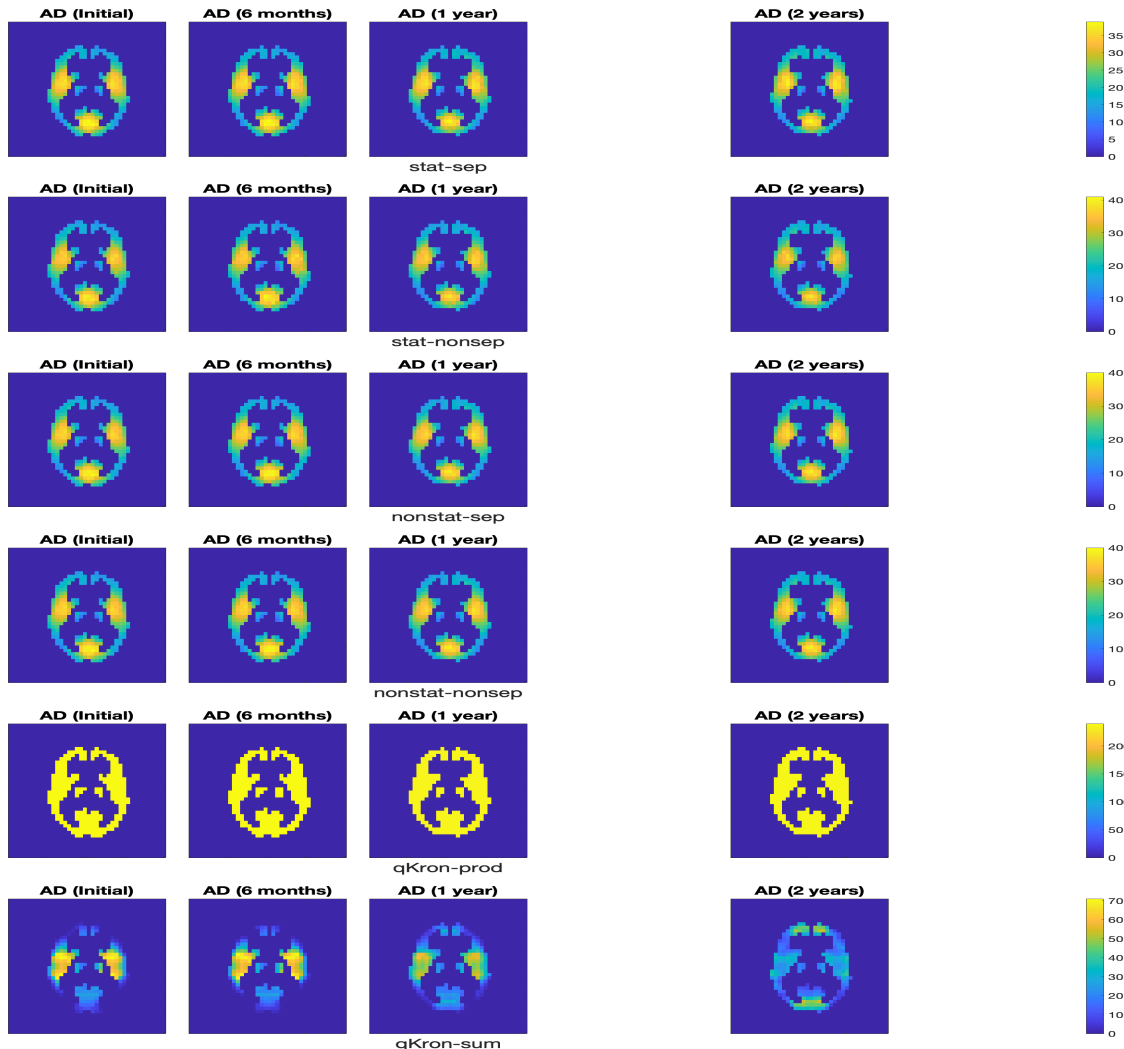
Figure 9: Estimated graph connection of the brain images for AD by various models. The color at each point represents the number of nodes connected to it, indicating the brain activeness.

and predict the change of the brain structure in these patients and uncover TESD in their brain regions in the past and for the future. The numerical evidences have verified the effectiveness and efficiency of the proposed model in characterizing TESD.

There are multiple future directions. For example, we can model the regularity of the priors through the decaying rate of dynamic eigenvalues (20) to learn from data, e.g., $\gamma_\ell \sim \Gamma^{-1}(a_\ell, b_\ell)$ with $b_\ell/a_\ell = \mathcal{O}(\ell^{-\kappa/2})$ or $(\kappa - 1) \sim \Gamma(a, b)$. The proposed model uses full data on the grid of the space and time but can be readily relaxed to handle missing data. The model can also be generalized to include covariates (regression) to explain the response variable (process) (Hyun et al., 2016). Therefore, the estimation and prediction of TESD, e.g. in the brain images, can be done at individual level. We can further incorporate such information in the covariance and investigate the effect of covariates on TESD.
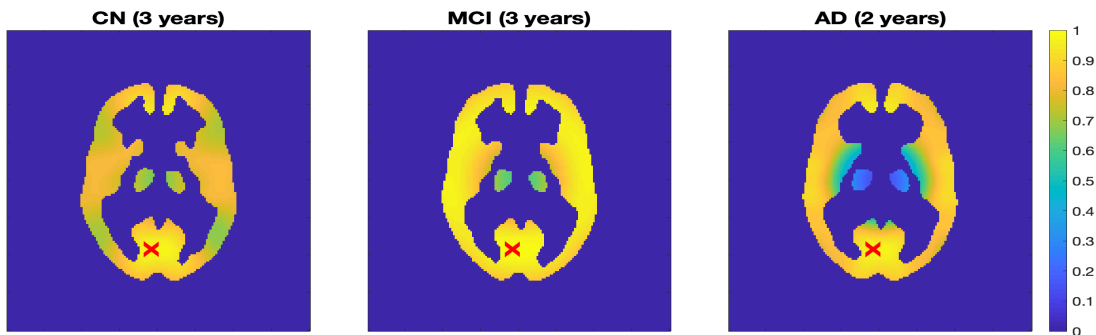
Figure 10: Predicted correlation between the brain region of interest (ROI) and a selected point of interest (POI, red cross) for CN (left), MCI (middle) and AD (right) respectively by the proposed model II (qKron-sum). The color at a point in ROI indicates the correlation between that point and the selected POI in the scale of $[0, 1]$.

The proposed scheme is designed to learn space-time interaction, more specifically, the spatial dependence conditioned commonly evolving time. The other space-time interaction, the spatial variation of temporal correction (SVTC), can be studied in the same vein by introducing space-dependence to the temporal kernel. This is more related and comparable to the "coregionalization" model (Banerjee, 2015) and could have potential applications in the study of animal migration or climate change. More generally, we could conduct similar analysis (conditional evolution) of interaction between any two types of information that could go beyond space and time, or even among more than two types, which involves tensor analysis. We leave them to future exploration.

For the longitudinal analysis of AD patients' brain images, subjects who are followed up for the whole study are limited in number. There are more subjects dropped in the middle or missing scheduled scans from the ADNI (ADNI, 2003) study thus discarded in the paper. Therefore, there could be large variance in the estimation of TESD due to the insufficient data. As ADNI continues collecting more data, they are expected to contribute a more accurate description of TESD in AD brain images that could facilitate the understanding of the mechanism behind this disease. Another important topic is the diagnosis of AD. It would be interesting to investigate TESD of the subjects' brains before and after being diagnosed with AD, which could shed more light on the reason of such disease.

Acknowledgments

SUPPLEMENTARY MATERIAL

Appendix A. Proofs

Theorem 1 *[Mercer's Kernels] Under Assumption 1, both $\mathcal{C}_m^I = \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}} \mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t$ and $\mathcal{C}_{y|m}^{II} = \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{I}_t$ are well defined non-negative definite kernels on $\mathcal{Z}$.*

Proof [Proof of Theorem 1] We first prove both series (17) and (18) converge in $L^1(\mathcal{Z} \times \mathcal{Z})$. Note for (17) we have

$$
\begin{aligned}
&\sum_{\ell=1}^{\infty} \left| \int_{\mathcal{Z}} \int_{\mathcal{Z}} \lambda_\ell(t) \mathcal{C}_t(t, t') \lambda_\ell(t') \phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}') d\mathbf{z} d\mathbf{z}' \right| \\
&\leq \sum_{\ell=1}^{\infty} \left| \int_{\mathcal{T}} \int_{\mathcal{T}} \lambda_\ell(t) \mathcal{C}_t(t, t') \lambda_\ell(t') dt dt' \right| \int_{\mathcal{X}} \int_{\mathcal{X}} |\phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}')| d\mathbf{x} d\mathbf{x}' \\
&\lesssim \sum_{\ell=1}^{\infty} |\langle \lambda_\ell, \mathcal{C}_t \lambda_\ell \rangle| (\|\phi_\ell(\mathbf{x})\|_2^2 + \|\phi_\ell(\mathbf{x}')\|_2^2)/2 \\
&\leq \|\mathcal{C}_t\| \sum_{\ell=1}^{\infty} \|\lambda_\ell\|_2^2 = \|\mathcal{C}_t\| \|\lambda\|_{2,2}^2 < +\infty
\end{aligned}
$$

And for (18) we can bound

$$
\begin{aligned}
&\sum_{\ell=1}^{\infty} \left| \int_{\mathcal{Z}} \int_{\mathcal{Z}} \lambda_\ell^2(t) \delta_t(t') \phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}') d\mathbf{z} d\mathbf{z}' \right| \\
&\leq \sum_{\ell=1}^{\infty} \int_{\mathcal{T}} \lambda_\ell^2(t) dt \int_{\mathcal{X}} \int_{\mathcal{X}} |\phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}')| d\mathbf{x} d\mathbf{x}' \lesssim \|\lambda\|_{2,2}^2 < +\infty
\end{aligned}
$$

The convergence of series (17) and (18) follows by the dominated convergence theorem.

Now we prove the non-negativeness. $\forall f(\mathbf{z}) \in L^2(\mathcal{Z})$, denote $f_\ell(t) := \int_{\mathcal{X}} f(\mathbf{z}) \phi_\ell(\mathbf{x}) d\mathbf{x}$. Then we have

$$
\begin{aligned}
\langle f(\mathbf{z}), \mathcal{C}_m^I f(\mathbf{z}') \rangle &= \langle f(\mathbf{z}), \int_{\mathcal{Z}} \sum_{\ell=1}^{\infty} \lambda_\ell(t) \mathcal{C}_t(t, t') \lambda_\ell(t') \phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}') f(\mathbf{z}') d\mathbf{z}' \rangle \\
&= \langle f(\mathbf{z}), \sum_{\ell=1}^{\infty} \int_{\mathcal{T}} \lambda_\ell(t) \mathcal{C}_t(t, t') \lambda_\ell(t') \phi_\ell(\mathbf{x}) f_\ell(t') dt' \rangle \\
&= \sum_{\ell=1}^{\infty} \int_{\mathcal{T}} \int_{\mathcal{T}} f_\ell(t) \lambda_\ell(t) \mathcal{C}_t(t, t') \lambda_\ell(t') f_\ell(t') dt dt' \\
&= \sum_{\ell=1}^{\infty} \langle f_\ell \lambda_\ell, \mathcal{C}_t \lambda_\ell f_\ell \rangle \geq 0
\end{aligned}
$$

where the convergence can be shown as above. Similarly we have

$$\langle f(\mathbf{z}), \mathcal{C}_{y|m}^{\mathrm{II}} f(\mathbf{z}')\rangle = \langle f(\mathbf{z}), \int_{\mathcal{Z}} \sum_{\ell=1}^{\infty} \lambda_{\ell}^2(t)\delta(t=t')\phi_{\ell}(\mathbf{x})\phi_{\ell}(\mathbf{x}')f(\mathbf{z}')d\mathbf{z}'\rangle$$

$$= \langle f(\mathbf{z}), \sum_{\ell=1}^{\infty} \int_{\mathcal{T}} \lambda_{\ell}^2(t)\delta(t=t')\phi_{\ell}(\mathbf{x})f_{\ell}(t')dt'\rangle = \sum_{\ell=1}^{\infty} \int_{\mathcal{T}} \lambda_{\ell}^2(t)f_{\ell}^2(t)dt \geq 0$$

Therefore we complete the proof. ∎

Theorem 2 *[Karhunen-Loéve Expansion] Under Assumption 1, STGP $f(\mathbf{x}, t) \sim \mathcal{GP}(0, \mathcal{C}_{\mathbf{z}})$ has the following representation of series expansion:*

$$f(\mathbf{x}, t) = \sum_{\ell=1}^{\infty} f_{\ell}(t)\phi_{\ell}(\mathbf{x}), \quad f_{\ell}(t) = \int_{\mathcal{X}} f(\mathbf{x}, t)\phi_{\ell}(\mathbf{x})d\mathbf{x} \tag{22}$$

where $\{f_{\ell}\}_{\ell=1}^{\infty}$ are random processes with mean functions $\mathbb{E}[f_{\ell}(t)] = 0$ and covariance functions as follows

- *if $\mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t$, then $\mathbb{E}[f_{\ell}(t)f_{\ell'}(t')] = \lambda_{\ell}(t)\mathcal{C}_t(t, t')\lambda_{\ell}(t')\delta_{\ell\ell'}$.*

- *if $\mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{I}_t$, then $\mathbb{E}[f_{\ell}(t)f_{\ell'}(t')] = \lambda_{\ell}^2(t)\delta(t=t')\delta_{\ell\ell'}$.*

Proof [Proof of Theorem 2] Note that $\{\phi_{\ell}(\mathbf{x})\}_{\ell=1}\infty$ is an orthonormal basis for $L^2(\mathcal{X})$, therefore we have the series representation (22) of $f(\cdot, t)$ for each $t \in \mathcal{T}$. Then we can calculate for $\mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t$ as in (17)

$$\mathbb{E}[f_{\ell}(t)] = \mathbb{E}\left[\int_{\mathcal{X}} f(\mathbf{x}, t)\phi_{\ell}(\mathbf{x})d\mathbf{x}\right] = \int_{\mathcal{X}} \mathbb{E}[f(\mathbf{x}, t)]\phi_{\ell}(\mathbf{x})d\mathbf{x} = 0$$

$$\mathbb{E}[f_{\ell}(t)f_{\ell'}(t')] = \mathbb{E}\left[\int_{\mathcal{X}} f(\mathbf{x}, t)\phi_{\ell}(\mathbf{x})d\mathbf{x} \int_{\mathcal{X}} f(\mathbf{x}', t')\phi_{\ell'}(\mathbf{x}')d\mathbf{x}'\right]$$

$$= \int_{\mathcal{X}} \int_{\mathcal{X}} \mathbb{E}[f(\mathbf{x}, t)f(\mathbf{x}', t')]\phi_{\ell}(\mathbf{x})\phi_{\ell'}(\mathbf{x}')d\mathbf{x}d\mathbf{x}'$$

$$= \int_{\mathcal{X}} \int_{\mathcal{X}} \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \otimes \mathcal{C}_t(\mathbf{z}, \mathbf{z}')\phi_{\ell}(\mathbf{x})\phi_{\ell'}(\mathbf{x}')d\mathbf{x}d\mathbf{x}'$$

$$= \sum_{\tilde{\ell}=1}^{\infty} \lambda_{\tilde{\ell}}(t)\mathcal{C}_t(t, t')\lambda_{\tilde{\ell}}(t') \int_{\mathcal{X}} \int_{\mathcal{X}} \phi_{\tilde{\ell}}(\mathbf{x})\phi_{\tilde{\ell}}(\mathbf{x}')\phi_{\ell}(\mathbf{x})\phi_{\ell'}(\mathbf{x}')d\mathbf{x}d\mathbf{x}'$$

$$= \lambda_{\ell}(t)\mathcal{C}_t(t, t')\lambda_{\ell}(t')\delta_{\ell\ell'}$$

Similarly we have $\mathbb{E}[f_{\ell}(t)f_{\ell'}(t')] = \lambda_{\ell}^2(t)\delta(t=t')\delta_{\ell\ell'}$ for $\mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{I}_t$ using the definition (18).

Lastly, we prove the convergence of the infinite sum. Denote $F_L(\mathbf{x}, t) = \sum_{\ell=1}^{L} f_\ell(t)\phi_\ell(\mathbf{x})$. We have for $\mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t$

$$\mathbb{E}[|f - F_L|^2] = \mathbb{E}[f^2] + \mathbb{E}[F_L^2] - 2\mathbb{E}[fF_L]$$

$$= \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t(\mathbf{z}, \mathbf{z}) + \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{\ell'=1}^{L} f_\ell(t)f_{\ell'}(t)\phi_\ell(\mathbf{x})\phi_{\ell'}(\mathbf{x})\right] - 2\mathbb{E}\left[f\sum_{\ell=1}^{L} f_\ell(t)\phi_\ell(\mathbf{x})\right]$$

$$= \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t(\mathbf{z}, \mathbf{z}) + \sum_{\ell=1}^{L} \lambda_\ell^2(t)\mathcal{C}_t(t,t)\phi_\ell^2(\mathbf{x}) - 2\sum_{\ell=1}^{L}\int_{\mathcal{X}} \mathbb{E}[f(\mathbf{x}, t)f(\mathbf{x}', t)]\phi_\ell(\mathbf{x}')\phi_\ell(\mathbf{x})d\mathbf{x}'$$

$$= \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t(\mathbf{z}, \mathbf{z}) + \sum_{\ell=1}^{L} \lambda_\ell^2(t)\mathcal{C}_t(t,t)\phi_\ell^2(\mathbf{x})$$

$$- 2\sum_{\ell=1}^{L}\sum_{\ell'=1}^{L} \lambda_{\ell'}^2(t)\mathcal{C}_t(t,t)\int_{\mathcal{X}} \phi_\ell(\mathbf{x}')\phi_\ell(\mathbf{x})\phi_{\ell'}(\mathbf{x})\phi_{\ell'}(\mathbf{x}')d\mathbf{x}'$$

$$= \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t(\mathbf{z}, \mathbf{z}) - \sum_{\ell=1}^{L} \lambda_\ell^2(t)\mathcal{C}_t(t,t)\phi_\ell^2(\mathbf{x}) = \sum_{\ell=L+1}^{\infty} \lambda_\ell^2(t)\mathcal{C}_t(t,t)\phi_\ell^2(\mathbf{x}) \to 0, \text{ as } L \to \infty$$

The same argument (by replacing $\mathcal{C}(t, t')$ with $\delta(t = t')$) yields the $L_{\mathbb{P}}^2$ convergence of the expansion (22) for $\mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{I}_t$. ∎

Theorem 3 *[Regularity of Random Functions] Assume $\lambda \in \ell^{2,s}(L^2(\mathcal{T}))$. If $f(\mathbf{x}, t) \sim \mathcal{GP}(0, \mathcal{C}_{\mathbf{z}})$ as in Theorem 2, then $f = \sum_{\ell=1}^{\infty} f_\ell(t)\phi_\ell(\mathbf{x}) \in \ell^{2,s}(L^2(\mathcal{Z}))$ in probability.*

Moreover, under Assumption 3-(24), there is a version [4] $\tilde{f}(\mathbf{x})$ of $f(\mathbf{x}) := \int_{\mathcal{T}} f(\mathbf{x}, t)dt$ in $C^{0,s'}(\mathcal{X})$ for $s' < s$. If further $\mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t$ and $\{Q_{\lambda_\ell, \mathcal{C}_t}\}$ satisfies Assumption 3-(25), then there is a version $\tilde{f}(\mathbf{z})$ of $f(\mathbf{z})$ in $C^{0,s'}(\mathcal{Z})$ for $s' < s$.

Proof [Proof of Theorem 3] We compute the expectation of the $(2, s, 2)$-norm of f

$$\mathbb{E}[\|f\|_{2,s,2}^2] = \sum_{\ell=1}^{\infty} \ell^{2s}\mathbb{E}[\|f_\ell\|_2^2] = \sum_{\ell=1}^{\infty} \ell^{2s}\int_{\mathcal{T}} \mathbb{E}[f_\ell^2(t)]dt$$

$$= \begin{cases} \sum_{\ell=1}^{\infty} \ell^{2s}\int_{\mathcal{T}} \mathcal{C}_t(t,t)\lambda_\ell^2(t)dt \lesssim \sum_{\ell=1}^{\infty} \ell^{2s}\|\lambda_\ell\|_2^2, & \text{if } \mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}}\mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \dot{\otimes} \mathcal{C}_t \\ \sum_{\ell=1}^{\infty} \ell^{2s}\|\lambda_\ell\|_2^2, & \text{if } \mathcal{C}_{\mathbf{z}} = \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{I}_t \end{cases}$$

$$= \|\lambda\|_{2,s,2}^2 < +\infty$$

where the equality on the second line follows from Theorem 2. This implies $f \in \ell^{2,s}(L^2(\mathcal{T}))$ in probability.

Now we prove the Hölder continuity of the random function f using Kolmogorov's celebrated continuity test (Theorem 3.42 of Hairer, 2009) and (Theorem 30 in section A.2.5

---

[4] A version/modification of stochastic process $\tilde{f}(\mathbf{x})$ of $f(\mathbf{x})$ means $\mathbb{P}[\tilde{f}(\mathbf{x}) = f(\mathbf{x})] = 1$ for $\forall \mathbf{x} \in \mathcal{X}$.

of Dashti and Stuart, 2017). First, we consider the marginal function. By Jensen's inequality

$$\mathbb{E}[|f(\mathbf{x}) - f(\mathbf{x}')|^2] \leq \int_{\mathcal{T}} \mathbb{E}[|f(\mathbf{x}, t) - f(\mathbf{x}', t)|^2] dt = \sum_{\ell, \ell'} \int_{\mathcal{T}} \mathbb{E}[f_\ell(t) f_{\ell'}(t)] dt \Delta \phi_\ell \Delta \phi_{\ell'}$$

$$= \sum_{\ell=1}^{\infty} \|\lambda_\ell\|_2^2 |\phi_\ell(\mathbf{x}) - \phi_\ell(\mathbf{x}')|^2 \leq \sum_{\ell=1}^{\infty} \|\lambda_\ell\|_2^2 \min\{2\|\phi_\ell\|_\infty^2, \mathrm{Lip}(\phi_\ell)^2 |\mathbf{x} - \mathbf{x}'|^2\}$$

$$\leq 2 \sum_{\ell=1}^{\infty} \|\lambda_\ell\|_2^2 \|\phi_\ell\|_\infty^{2-\delta} \mathrm{Lip}(\phi_\ell)^\delta |\mathbf{x} - \mathbf{x}'|^\delta \lesssim \sum_{\ell=1}^{\infty} \ell^\delta \|\lambda_\ell\|_2^2 |\mathbf{x} - \mathbf{x}'|^\delta$$

$$\leq \|\lambda\|_{2,s,2}^2 |\mathbf{x} - \mathbf{x}'|^\delta \ \text{for } \delta < 2s$$

where we used that $\min\{a, bx^2\} \leq a^{1-\frac{\delta}{2}} b^{\frac{\delta}{2}} |x|^\delta$ for $\delta \in [0, 2]$. Then by Kolmogorov's continuity theorem there is a modification $\tilde{f}(\mathbf{x})$ of $f(\mathbf{x})$ in $C^{0,s'}(\mathcal{X})$ for $s' < \delta/2 < s$.

Lastly, we consider the full function $f(\mathbf{x}, t)$.

$$\mathbb{E}[|f(\mathbf{x}, t) - f(\mathbf{x}', t')|^2] = \mathbb{E}\left[\left|\sum_{\ell=1}^{\infty} f_\ell(t) \Delta \phi_\ell + \Delta f_\ell \phi_\ell(\mathbf{x}')\right|^2\right]$$

$$= \sum_{\ell, \ell'} \mathbb{E}[f_\ell(t) f_{\ell'}(t)] \Delta \phi_\ell \Delta \phi_{\ell'} + 2 \sum_{\ell, \ell'} \mathbb{E}[f_\ell(t) \Delta f_{\ell'}] \Delta \phi_\ell \phi_{\ell'}(\mathbf{x}') + \sum_{\ell, \ell'} \mathbb{E}[\Delta f_\ell \Delta f_{\ell'}] \phi_\ell(\mathbf{x}') \phi_{\ell'}(\mathbf{x}')$$

$$\leq 2 \left(\sum_{\ell=1}^{\infty} \lambda_\ell^2(t) |\Delta \phi_\ell|^2 + \sum_{\ell=1}^{\infty} \mathbb{E}[|\Delta f_\ell|^2] \phi_\ell^2(\mathbf{x}')\right) \leq 2 \left(\sum_{\ell=1}^{\infty} \|\lambda_\ell\|_\infty^2 |\Delta \phi_\ell|^2 + \sum_{\ell=1}^{\infty} Q_{\lambda_\ell, \mathcal{C}_t}(t, t') \|\phi_\ell\|_\infty^2\right)$$

$$\lesssim \sum_{\ell=1}^{\infty} \ell^\delta \|\lambda_\ell\|_\infty^2 |\mathbf{x} - \mathbf{x}'|^\delta + \sum_{\ell=1}^{\infty} \min\{\lambda_\ell^2(t) + \lambda_\ell^2(t'), C\ell^2 \|\lambda_\ell\|_\infty^2 |t - t'|^2\}$$

$$\lesssim \sum_{\ell=1}^{\infty} \ell^\delta \|\lambda_\ell\|_\infty^2 |\mathbf{x} - \mathbf{x}'|^\delta + \sum_{\ell=1}^{\infty} \ell^\delta \|\lambda_\ell\|_\infty^2 |t - t'|^\delta$$

$$\lesssim \|\lambda\|_{2,s,\infty}^2 |\mathbf{z} - \mathbf{z}'|^\delta \ \text{for } \delta < 2s$$

where $\Delta f := f(t) - f(t')$, $\Delta \phi := \phi(\mathbf{x}) - \phi(\mathbf{x}')$, and the first inequality is due to Cauchy-Schwarz inequality and $2ab \leq a^2 + b^2$. The conclusion follows by Kolmogorov's continuity theorem. ∎

Corollary 1 *If $f(\mathbf{x}, t) \sim \mathcal{GP}(0, \mathcal{C}_\mathbf{z})$ has a continuous version, then $\{f_\ell\}_{\ell=1}^{\infty}$ as in Theorem 2 are GP's defined on $\mathcal{T}$.*

Proof [Proof of Corollary 1] $f_\ell(t)$ can be viewed as infinite weighted sum of GP $f(\mathbf{x}, t)$ thus becomes another GP. This can be made rigorous by approximating $\phi_\ell(\mathbf{x})$ with a sequence of simple functions $\phi_{n,\ell} = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ with disjoint $\{A_i\}$:

$$f_\ell(t) = \int_{\mathcal{X}} f(\mathbf{x}, t) \phi_\ell(\mathbf{x}) d\mathbf{x} = \lim_{n \to +\infty} \int_{\mathcal{X}} f(\mathbf{x}, t) \phi_{n,\ell}(\mathbf{x}) d\mathbf{x} = \sum_{i=1}^{\infty} a_i \int_{A_i} f(\mathbf{x}, t) d\mathbf{x}$$

Note for $\forall t \in \mathcal{T}$, $f_{A_i}(t) := \int_{A_i} f(\mathbf{x}, t)d\mathbf{x}$ coincides with the Riemann integral. Thus $\{f_{A_i}(t)\}$ are jointly normal as a limit of (Riemann) sum of (weighted) joint Gaussian random variables. Therefore $f_\ell(t)$ is normal for any fixed $t \in \mathcal{T}$. The same argument applies to $\mathbf{t} = (t_1, \cdots, t_k)$ replacing t. Thus it concludes the proof. ∎

For the dynamic spatial kernels $\mathcal{C}_i = \sum_{\ell=1}^{\infty} \lambda_{i,\ell}^2(t)\phi_\ell \otimes \phi_\ell$, we consider the Gaussian likelihood models $p_i \sim \mathcal{N}_n(\mathbf{m}_i(t), \mathbf{C}_i(t))$, with $\mathbf{C}_i = \sum_{\ell=1}^{n} \lambda_{i,\ell}^2(t)\phi_\ell(\mathbf{x}) \otimes \phi_\ell(\mathbf{x}') = \boldsymbol{\Phi}\boldsymbol{\Lambda}_i\boldsymbol{\Phi}^\mathsf{T}$, for $i = 0, 1$. For $\lambda_i \in \ell^{1,s}(L^\infty(\mathcal{T}))$ with some $s > 0$, we can bound the Hellinger distance d_H, Kullback-Leibler (K-L) divergence ($K(p_0, p_1) := \mathrm{E}_0(\log(p_0/p_1))$) and K-L variation ($V(p_0, p_1) := \mathrm{E}_0(\log(p_0/p_1))^2$) between two models with their difference in eigenvalues measured by $\|\cdot\|_{1,s,\infty}$ in the following lemma.

Lemma A.1 *Let $p_i \sim \mathcal{N}_n(0, \mathbf{C}_i(t))$ be Gaussian models for $i = 0, 1$, with $\{\lambda_{i,\ell}^2(t)\}$ being the eigenvalues of $\mathbf{C}_i = \boldsymbol{\Phi}\boldsymbol{\Lambda}_i\boldsymbol{\Phi}^\mathsf{T}$ satisfying Assumption 4. Then we have*

- $d_H(p_0, p_1) \lesssim \|\lambda_0 - \lambda_1\|_{1,s,\infty}^{\frac{1}{2}}$

- $K(p_0, p_1) \lesssim \|\lambda_0 - \lambda_1\|_{1,s,\infty}$

- $V(p_0, p_1) \lesssim \|\lambda_0 - \lambda_1\|_{1,s,\infty}^2$

Proof First we calculate the Kullback-Leibler divergence

$$K(p_0, p_1) = \frac{1}{2}\left\{\mathrm{tr}(\mathbf{C}_1^{-1}\mathbf{C}_0 - \mathbf{I}) + (\mathbf{m}_1 - \mathbf{m}_0)^\mathsf{T}\mathbf{C}_1^{-1}(\mathbf{m}_1 - \mathbf{m}_0) + \log\frac{|\mathbf{C}_1|}{|\mathbf{C}_0|}\right\}$$

Consider $\mathbf{m}_i \equiv 0$. By the non-negativity of K-L divergence we have for general $\mathbf{C}_i > 0$,

$$\log\frac{|\mathbf{C}_0|}{|\mathbf{C}_1|} \leq \mathrm{tr}(\mathbf{C}_1^{-1}\mathbf{C}_0 - \mathbf{I}) \tag{A.1}$$

Therefore we can bound K-L divergence

$$K(p_0, p_1) \leq \frac{1}{2}\{\mathrm{tr}(\mathbf{C}_1^{-1}\mathbf{C}_0 - \mathbf{I}) + \mathrm{tr}(\mathbf{C}_0^{-1}\mathbf{C}_1 - \mathbf{I})\} \leq 2C\|\lambda_0 - \lambda_1\|_{1,s,\infty}$$

where we use

$$\mathrm{tr}(\mathbf{C}_1^{-1}\mathbf{C}_0 - \mathbf{I}) = \sum_{\ell=1}^{n} \lambda_{1,\ell}^{-2}(t)(\lambda_{0,\ell}^2(t) - \lambda_{1,\ell}^2(t)) \leq 2C\sum_\ell c_\ell^{-2}\|\lambda_{0,\ell} - \lambda_{1,\ell}\|_\infty \leq 2C\|\lambda_0 - \lambda_1\|_{1,s,\infty} \tag{A.2}$$

Now we calculate the following K-L variation

$$V(p_0, p_1) = \frac{1}{2}\mathrm{tr}((\mathbf{C}_1^{-1}\mathbf{C}_0 - \mathbf{I})^2) + (\mathbf{m}_1 - \mathbf{m}_0)^\mathsf{T}\mathbf{C}_1^{-1}\mathbf{C}_0\mathbf{C}_1^{-1}(\mathbf{m}_1 - \mathbf{m}_0) + K^2(p_0, p_1)$$

Consider $\mathbf{m}_i \equiv 0$ and we can bound it by the similar argument as (A.2)

$$V(p_0, p_1) \leq C^2\|\lambda_0 - \lambda_1\|_{2,s,\infty}^2 + 4C^2\|\lambda_0 - \lambda_1\|_{1,s,\infty}^2 \lesssim \|\lambda_0 - \lambda_1\|_{1,s,\infty}^2$$

It is easy to see that the centered K-L variation $V_0(p_0, p_1) = \mathrm{Var}_0(\log(p_0/p_1)) = \mathrm{E}_0(\log(p_0/p_1) - K(p_0, p_1))^2$ can be bounded

$$V_0(p_0, p_1) \leq C^2 \|\lambda_0 - \lambda_1\|_{2, s, \infty}^2$$

Lastly, the squared Hellinger distance for multivariate Gaussians can be calculated

$$h^2(p_0, p_1) = 1 - \frac{|\mathbf{C}_0 \mathbf{C}_1|^{1/4}}{\left|\frac{\mathbf{C}_0 + \mathbf{C}_1}{2}\right|^{1/2}} \exp\left\{-\frac{1}{8}(\mathbf{m}_0 - \mathbf{m}_1)^\mathsf{T} \left(\frac{\mathbf{C}_0 + \mathbf{C}_1}{2}\right)^{-1} (\mathbf{m}_0 - \mathbf{m}_1)\right\}$$

Consider $\mathbf{m}_i \equiv 0$. Notice that $1 - x \leq -\log x$, and by (A.1) we can bound the squared Hellinger distance using the similar argument in (A.2)

$$h^2(p_0, p_1) \leq \log \frac{\left|\frac{\mathbf{C}_0 + \mathbf{C}_1}{2}\right|^{1/2}}{|\mathbf{C}_0 \mathbf{C}_1|^{1/4}} \leq \frac{1}{2}\mathrm{tr}(\mathbf{C}_0^{-\frac{1}{2}} \mathbf{C}_1^{-\frac{1}{2}}(\mathbf{C}_0 + \mathbf{C}_1)/2 - \mathbf{I})$$

$$\leq \frac{1}{4}\{\mathrm{tr}(\mathbf{C}_1^{-\frac{1}{2}} \mathbf{C}_0^{\frac{1}{2}} - \mathbf{I}) + \mathrm{tr}(\mathbf{C}_0^{-\frac{1}{2}} \mathbf{C}_1^{\frac{1}{2}} - \mathbf{I})\} \leq \frac{1}{2}\|\lambda_0 - \lambda_1\|_{1, s, \infty}$$

$\blacksquare$

Following Ghosal and van der Vaart (2017), now we prove the following posterior contraction about $\mathcal{C}_{\mathbf{x}|t}$ in model II. For the convenience of discussion, we fix all hyper-parameters at their optimal values. One can refer to van der Vaart and van Zanten (2009, 2011) for varying them to scale GP.

Theorem 4 [Posterior Contraction of $\mathcal{C}_{\mathbf{x}|t}$ in model II] Let λ be a Borel measurable, zero-mean, tight Gaussian random element in $\Theta = \ell^2(L^2(\mathcal{T}))$ satisfying Assumption 4 and $P_\lambda^{(n)} = \bigotimes_{j=1}^n P_{\lambda, j}$ be the product measure of $Y^{(n)}$ parametrized by λ. If the true value $\lambda_0 \in \Theta$ is in the support of λ, and ε_n satisfies the rate equation $\varphi_{\lambda_0}(\varepsilon_n) \leq n\varepsilon_n^2$ with $\varepsilon_n \geq n^{-\frac{1}{2}}$, then there exists $\Theta_n \subset \Theta$ such that $\Pi_n(\lambda \in \Theta_n : d_{n,H}(\lambda, \lambda_{n,0}) > M_n\varepsilon_n|Y^{(n)}) \to 0$ in $P_{\lambda_{n,0}}^{(n)}$-probability for every $M_n \to \infty$.

Proof [Proof of Theorem 4] We use Theorem 1 of Ghosal and van der Vaart (2007) and it suffices to verify the following two conditions (the entropy condition (2.4), and the prior mass condition (2.5)) for some universal constants $\xi, K > 0$ and sufficiently large $k \in \mathbb{N}$:

$$\sup_{\varepsilon > \varepsilon_n} \log N(\xi\varepsilon/2, \{\lambda \in \Theta_n : d_{n,H}(\lambda, \lambda_{n,0}) < \varepsilon\}, d_{n,H}) \leq n\varepsilon_n^2 \tag{A.3}$$

$$\frac{\Pi_n(\lambda \in \Theta_n : k\varepsilon_n < d_{n,H}(\lambda, \lambda_{n,0}) < 2k\varepsilon_n)}{\Pi_n(\overline{B}_n(\lambda_{n,0}, \varepsilon_n))} \leq e^{Kn\varepsilon_n^2 k^2/2} \tag{A.4}$$

where the left side of (A.3) is called *Le Cam dimension* (LeCam, 1973, 1975), logarithm of the minimal number of $d_{n,H}$-balls of radius $\xi\varepsilon/2$ needed to cover a ball of radius ε around the true value $\lambda_{n,0}$; $\overline{B}_n(\lambda_{n,0}, \varepsilon) := \{\lambda \in \Theta : \frac{1}{n}\sum_{j=1}^n K_j(\lambda_{n,0}, \lambda) \leq \varepsilon^2, \frac{1}{n}\sum_{j=1}^n V_j(\lambda_{n,0}, \lambda) \leq \varepsilon^2\}$, with $K_j(\lambda_{n,0}, \lambda) = K(P_{\lambda_{n,0},j}, P_{\lambda,j})$ and $V_j(\lambda_{n,0}, \lambda) = V(P_{\lambda_{n,0},j}, P_{\lambda,j})$. For each $1 \leq \ell \leq n$, define the coordinate rate function

$$\varphi_{\lambda_0, \ell}(\varepsilon_{n,\ell}) = \inf_{h \in \mathbb{H}_\ell : \|h - \lambda_{0,\ell}\|_\infty \leq \varepsilon_{n,\ell}} \frac{1}{2}\|h\|_{\mathbb{H}_\ell}^2 - \log \Pi_\ell(\|\lambda_\ell\|_\infty < \varepsilon_{n,\ell}) \tag{A.5}$$

For each Gaussian random element $\lambda_\ell \in \mathbb{B}_\ell = L^\infty(\mathcal{T})$, we have $\lambda_{0,\ell} \in \overline{\overline{\mathbb{H}}}_\ell$ and the measurable set $B_{n,\ell} \subset \mathbb{B}_\ell$ (c.f. Theorem 2.1 of van der Vaart and van Zanten, 2008) such that

$$\log N(3\varepsilon_{n,\ell}, B_{n,\ell}, \|\cdot\|_\infty) \leq 6Cn\varepsilon_{n,\ell}^2 \tag{A.6}$$

$$\Pi_\ell(\lambda_\ell \notin B_{n,\ell}) \leq e^{-Cn\varepsilon_{n,\ell}^2} \tag{A.7}$$

$$\Pi_\ell(\|\lambda_\ell - \lambda_{0,\ell}\|_\infty < 2\varepsilon_{n,\ell}) \geq e^{-n\varepsilon_{n,\ell}^2} \tag{A.8}$$

Now let $\varepsilon_{n,\ell} = 2^{-\ell}\ell^{-s}\varepsilon_n^2$ for $\ell = 1, \cdots, n$. Set $\Theta_n = \{\lambda \in \Theta \cap \ell^{1,s}(L^\infty(\mathcal{T})) : \lambda_\ell \in B_{n,\ell}\} \subset \Theta$, and $N(\varepsilon_n, \Theta_n, d_{n,H}) = \max_{1 \leq \ell \leq n} N(3\varepsilon_{n,\ell}, B_{n,\ell}, \|\cdot\|_\infty)$. By Lemma A.1 and (A.6), we have the following global entropy bound because $d_{n,H}^2(\lambda, \lambda') \leq \|\lambda - \lambda'\|_{1,s,\infty} \leq \varepsilon_n^2$ for $\forall \lambda, \lambda' \in \Theta_n$.

$$\log N(\varepsilon_n, \Theta_n, d_{n,H}) \leq 6Cn(2^{-\ell}\ell^{-s}\varepsilon_n^2)^2 \leq Cn\varepsilon_n^4 \leq n\varepsilon_n^2$$

which is stronger than the local entropy condition (A.3). Now by Lemma A.1 and (A.8) we have

$$\Pi_n(\overline{B}_n(\lambda_{n,0}, \varepsilon_n)) \geq \Pi_n(\|\lambda_{n,0} - \lambda\|_{1,s,\infty} \leq \varepsilon_n^2, \|\lambda_{n,0} - \lambda\|_{1,s,\infty}^2 \leq \varepsilon_n^2)$$

$$= \Pi_n(\|\lambda_{n,0} - \lambda\|_{1,s,\infty} \leq \varepsilon_n^2) \geq \exp\left\{\sum_{\ell=1}^n \log \Pi_\ell(\|\lambda_\ell - \lambda_{0,\ell}\|_\infty < 2\varepsilon_{n,\ell})\right\}$$

$$\geq e^{-n\sum_{\ell=1}^n \varepsilon_{n,\ell}^2} = e^{-Knk^2\varepsilon_n^4/2}, \quad \text{with } K = 2, \; k^2 = \sum_{\ell=1}^n 2^{-2\ell}\ell^{-2s}$$

Then (A.4) is immediately satisfied because the numerator is bounded by 1. Therefore the proof is completed. ∎

Remark 6 *This theorem generalizes Theorem 2.2 of Lan et al. (2020) where the spatial domain has fixed size D. Therefore the Hellinger metric, KL divergence and variance are easier to bound (Lemma B.1). Note we do not have the complementary assertion as in Lemma 1 of Ghosal and van der Vaart (2007) thus the resulting contraction is only on Θ_n, weaker than that in Theorem 2.2 of Lan et al. (2020).*

Theorem 5 *[Posterior Contraction Rate of $\mathcal{C}_{\mathbf{x}|t}$ in model II] Let λ be a Gaussian random element satisfying Assumption 2-(20) and $\text{tr}(\mathcal{C}_u) = 1$. The rest settings are the same as in Theorem 4. If the true value $\lambda_0 \in \ell^{2,s}(L^2(\mathcal{T}))$, then we have the rate of posterior contraction $\varepsilon_n = \Theta(n^{-(\frac{\kappa-1}{2}\wedge s)/\kappa})$.*

Proof [Proof of Theorem 5] First, we prove that the negative logarithm of small ball probability $\varphi_0(\varepsilon) = -\log \Pi(\|\lambda\|_{2,2} < \varepsilon) = \mathcal{O}(\varepsilon^{-\frac{2}{\kappa-1}})$. Apply Karhunen-Loéve theorem to u_ℓ in model (20) to get $u_\ell(t) = \sum_{i=1}^\infty Z_{\ell,i}\xi_i\phi_i(t)$ with $Z_{\ell,i} \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\{\xi_i^2, \phi_i\}$ being the eigen-pairs of $\mathcal{C}_u$. Note $\mathbb{E}[\|u_\ell\|_2^2] = \sum_{i=1}^\infty \xi_i^2 = \text{tr}(\mathcal{C}_u) = 1$. Because normal densities with

standard deviations $\sigma \geq \tau$ satisfy $\phi_\sigma(z)/\phi_\tau(z) \geq \tau/\sigma$ for every $z \in \mathbb{R}$, we have

$$
\mathrm{P}\left(\sum_{\ell \leq L}\gamma_\ell^2\|u_\ell\|_2^2 < \varepsilon^2\right) = \mathrm{P}\left(\sum_{\ell \leq L}\gamma_\ell^2 \sum_{i=1}^\infty Z_{\ell,i}^2\xi_i^2 < \varepsilon^2\right) = \int_{\sum_{\ell \leq L}\sum_{i \in \mathbb{N}} z_{\ell,i}^2 < \varepsilon^2} \prod_{\ell \leq L, i \in \mathbb{N}} \phi_{\gamma_\ell \xi_i}(z_{\ell,i})dz_{\ell,i}
$$

$$
\geq \prod_{\ell=1}^L \frac{\gamma_L}{\gamma_\ell}\mathrm{P}\left(\sum_{\ell=1}^L\sum_{i=1}^\infty \gamma_L^2 Z_{\ell,i}^2\xi_i^2 < \varepsilon^2\right) \gtrsim \left(\frac{L!}{L^L}\right)^{\frac{\kappa}{2}}\mathrm{P}\left(\gamma_L^2\sum_{\ell=1}^L\|u_\ell\|_2^2 < \varepsilon^2\right)
$$

$$
\geq e^{-L\kappa/2}\frac{1}{2}
$$

for L large enough such that $\gamma_L^{-2}L^{-1}\varepsilon^2 \geq 1$ by the central limit theorem. This is satisfied when $L \gtrsim \varepsilon^{-2/(\kappa-1)}$. On the other hand, by Markov's inequality,

$$
\mathrm{P}\left(\sum_{\ell>L}\gamma_\ell^2\|u_\ell\|_2^2 < \varepsilon^2\right) \geq 1 - \frac{1}{\varepsilon^2}\sum_{\ell>L}\mathbb{E}[\gamma_\ell^2\|u_\ell\|_2^2] \geq 1 - \frac{1}{\varepsilon^2}\int_L^\infty x^{-\kappa}dx = 1 - \frac{1}{(\kappa-1)L^{\kappa-1}\varepsilon^2} \geq \frac{1}{2}
$$

for L large enough such that $(\kappa-1)L^{\kappa-1}\varepsilon^2 \geq 2$, i.e., $L \geq \varepsilon^{-2/(\kappa-1)}\left(\frac{2}{\kappa-1}\right)^{\frac{1}{\kappa-1}} \geq \varepsilon^{-2/(\kappa-1)}e^{-1/(2e)}$. Therefore we have $\mathrm{P}\left(\sum_{\ell=1}^\infty \gamma_\ell^2\|u_\ell\|_2^2 < 2\varepsilon^2\right) \gtrsim e^{-L\kappa/2}\frac{1}{2^2}$. Thus the best upper bound for $\varphi_0(\varepsilon) \lesssim L\kappa/2 \lesssim \varepsilon^{-2/(\kappa-1)}$.

Next, we show the de-centering function $\inf_{h \in \mathbb{H}:\|h-\lambda\|_{2,2}\leq \varepsilon}\|h\|_{\mathbb{H}}^2 \leq \|\lambda\|_{2,s,2}^{\kappa/s}\varepsilon^{-(\kappa-2s)/s}$ if $\lambda \in \ell^{2,s}(L^2(\mathcal{T}))$ for $s < \kappa/2$. For every $L \in \mathbb{N}$, $\lambda^L := \{\lambda_\ell\}_{\ell=1}^L \in \mathbb{H}$. Its square $(2,2)$-distance to λ and square RHKS-norm satisfy

$$
\|\lambda^L - \lambda\|_{2,2}^2 = \sum_{\ell>L}\|\lambda_\ell\|_2^2 \leq L^{-2s}\|\lambda\|_{2,s,2}^2
$$

$$
\|\lambda^L\|_{\mathbb{H}}^2 = \sum_{\ell=1}^L \gamma_\ell^{-2}\|\lambda_\ell\|_2^2 \leq \|\lambda\|_{2,s,2}^2 \max_{1\leq\ell\leq L}\gamma_\ell^{-2}\ell^{-2s} \lesssim \|\lambda\|_{2,s,2}^2 \max_{1\leq\ell\leq L}\ell^{\kappa-2s}
$$

Choosing the minimal integer L such that $L \geq \|\lambda\|_{2,s,2}^{1/s}\varepsilon^{-1/s}$ yields the result.

Finally, when the true parameter $\lambda_0 \in \ell^{2,s}(L^2(\mathcal{T}))$, then we get the minimal solution to the rate equation $\varphi_{\lambda_0}(\varepsilon_n) \leq n\varepsilon_n^2$ by setting both $\varepsilon_n^{-2/(\kappa-1)} \lesssim n\varepsilon_n^2$ and $\varepsilon_n^{-(\kappa-2s)/s} \lesssim n\varepsilon_n^2$, which gives the rate of posterior contraction $n^{-(\frac{\kappa-1}{2}\wedge s)/\kappa}$. ∎

Remark 7 *Posterior contraction rate is the minimal solution to the rate equation $\varphi_{\lambda_0}(\varepsilon_n) \leq n\varepsilon_n^2$. Therefore any rate slower than the result given above is also 'a' contraction rate. The minimax rate $n^{-s/(2s+1)}$ can be attained if and only if $(\kappa-1)/2 = s$, when the prior regularity matches that of the truth. When this does not happen, we can only expect suboptimal rates.*

Proposition 3 *Fit the spatiotemporal data $\mathcal{D} = \{\mathbf{Z}, \mathbf{Y}\}$ with the model (29). Then given a new point $\mathbf{z}_* = (\mathbf{x}_*, t_*)$ we have*

$$
m(\mathbf{z}_*)|\mathcal{D} \sim \mathcal{N}(m', C')
$$

$$
m' = c_*^{\mathsf{T}}(\mathbf{C_M} + K^{-1}\mathbf{C_{Y|M}})^{-1}\overline{\mathbf{Y}}, \quad C' = C_{m_*} - c_*^{\mathsf{T}}(\mathbf{C_M} + K^{-1}\mathbf{C_{Y|M}})^{-1}c_*
$$

where we denote

$$\overline{\mathbf{Y}} := \frac{1}{K} \sum_{k=1}^{K} \text{vec}(\mathbf{Y}_k), \quad C_{m_*} := \mathcal{C}_m(\mathbf{z}_*, \mathbf{z}_*), \quad c_* := \mathcal{C}_m(\mathbf{Z}, \mathbf{z}_*), \quad c_*^{\mathsf{T}} := \mathcal{C}_m(\mathbf{z}_*, \mathbf{Z})$$

Proof [Proof of Proposition 3] Compute using the following formula

$$p(m(\mathbf{z}_*)|\mathcal{D}) = \int p(m(\mathbf{z}_*), \mathbf{M}|\mathcal{D})d\mathbf{M} = \int p(m(\mathbf{z}_*)|\mathbf{M})p(\mathbf{M}|\mathcal{D})d\mathbf{M}$$

$$\propto \int p(m(\mathbf{z}_*)|\mathbf{M})p(\mathbf{M})p(\mathcal{D}|\mathbf{M})d\mathbf{M} = \int p(\mathbf{M}, m(\mathbf{z}_*))p(\mathcal{D}|\mathbf{M})d\mathbf{M}$$

Completing the square to integrate out $\mathbf{M}$ and completing the square for $m(\mathbf{z}_*)$ we have

$$m(\mathbf{z}_*)|\mathcal{D} \sim \mathcal{N}(m', C')$$
$$(C')^{-1} = C_{m_*|\mathbf{M}}^{-1} - C_{m_*}^{-1}c_*^{\mathsf{T}}C_{\mathbf{M}|m_*}^{-1}\mathbf{C}_{\text{post}}C_{\mathbf{M}|m_*}^{-1}c_*C_{m_*}^{-1}, \ m' = C'C_{m_*}^{-1}c_*^{\mathsf{T}}C_{\mathbf{M}|m_*}^{-1}\mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{Y}|\mathbf{M}}^{-1}K\overline{\mathbf{Y}}$$
$$C_{m_*|\mathbf{M}} := C_{m_*} - c_*^{\mathsf{T}}\mathbf{C}_{\mathbf{M}}^{-1}c_*, \quad C_{\mathbf{M}|m_*} := \mathbf{C}_{\mathbf{M}} - c_*C_{m_*}^{-1}c_*^{\mathsf{T}}, \quad \mathbf{C}_{\text{post}}^{-1} = C_{\mathbf{M}|m_*}^{-1} + K\mathbf{C}_{\mathbf{Y}|\mathbf{M}}^{-1}$$

By Sherman-Morrison-Woodbury formula, we further have

$$\begin{aligned}(C')^{-1} =& C_{m_*|\mathbf{M}}^{-1} - C_{m_*}^{-1}c_*^{\mathsf{T}}C_{\mathbf{M}|m_*}^{-1}(C_{\mathbf{M}|m_*}^{-1} + K\mathbf{C}_{\mathbf{Y}|\mathbf{M}}^{-1})^{-1}C_{\mathbf{M}|m_*}^{-1}c_*C_{m_*}^{-1} \\ =& C_{m_*}^{-1} + C_{m_*}^{-1}c_*^{\mathsf{T}}(\mathbf{C}_{\mathbf{M}} - c_*C_{m_*}^{-1}c_*^{\mathsf{T}})^{-1}c_*mC_{m_*}^{-1} \\ &- C_{m_*}^{-1}c_*^{\mathsf{T}}[C_{\mathbf{M}|m_*}^{-1} - (C_{\mathbf{M}|m_*} + K^{-1}\mathbf{C}_{\mathbf{Y}|\mathbf{M}})^{-1}]c_*C_{m_*}^{-1} \\ =& C_{m_*}^{-1} + C_{m_*}^{-1}c_*^{\mathsf{T}}(C_{\mathbf{M}|m_*} + K^{-1}\mathbf{C}_{\mathbf{Y}|\mathbf{M}})^{-1}c_*C_{m_*}^{-1} \\ =& C_{m_*}^{-1} + C_{m_*}^{-1}c_*^{\mathsf{T}}[(\mathbf{C}_{\mathbf{M}} + K^{-1}\mathbf{C}_{\mathbf{Y}|\mathbf{M}}) - c_*C_{m_*}^{-1}c_*^{\mathsf{T}}]^{-1}c_*C_{m_*}^{-1} \\ =& [C_{m_*} - c_*^{\mathsf{T}}(\mathbf{C}_{\mathbf{M}} + K^{-1}\mathbf{C}_{\mathbf{Y}|\mathbf{M}})^{-1}c_*]^{-1}\end{aligned}$$

and

$$\begin{aligned}m' =& C'C_{m_*}^{-1}c_*^{\mathsf{T}}C_{\mathbf{M}|m_*}^{-1}(C_{\mathbf{M}|m_*}^{-1} + K\mathbf{C}_{\mathbf{Y}|\mathbf{M}}^{-1})^{-1}\mathbf{C}_{\mathbf{Y}|\mathbf{M}}^{-1}K\overline{\mathbf{Y}} \\ =& C'C_{m_*}^{-1}c_*^{\mathsf{T}}(C_{\mathbf{M}|m_*} + K^{-1}\mathbf{C}_{\mathbf{Y}|\mathbf{M}})^{-1}\overline{\mathbf{Y}} \\ =& c_*^{\mathsf{T}}[I_{m_*} - (\mathbf{C}_{\mathbf{M}} + K^{-1}\mathbf{C}_{\mathbf{Y}|\mathbf{M}})^{-1}c_*C_{m_*}^{-1}c_*^{\mathsf{T}}](C_{\mathbf{M}|m_*} + K^{-1}\mathbf{C}_{\mathbf{Y}|\mathbf{M}})^{-1}\overline{\mathbf{Y}} \\ =& c_*^{\mathsf{T}}(\mathbf{C}_{\mathbf{M}} + K^{-1}\mathbf{C}_{\mathbf{Y}|\mathbf{M}})^{-1}\overline{\mathbf{Y}}\end{aligned}$$

∎

Appendix B. Posterior Inference

Discretize the spatial $\mathcal{X}$ and time $\mathcal{T}$ domains with I and J points respectively. Denote the observations on the discrete domain as $I \times J$ matrices $\mathbf{Y}_k$ for $k = 1, \cdots, K$ trials, and thus

$\mathbf{Y}_{I \times J \times K} = \{\mathbf{Y}_1, \cdots, \mathbf{Y}_K\}$. We summarize model I as follows

$$
\begin{aligned}
\mathbf{Y}_k | \mathbf{M}, \sigma_\varepsilon^2 &\sim \mathcal{MN}(\mathbf{M}, \sigma_\varepsilon^2 \mathbf{I}_\mathbf{x}, \mathbf{I}_t), \quad \mathbf{M}_{I \times J} = m(\mathbf{X}, \mathbf{t}) \\
m(\mathbf{x}, t) &\sim \mathcal{GP}(0, \mathcal{C}_{\mathbf{x}|t} \dot{\otimes} \mathcal{C}_t), \quad \mathcal{C}_t(t, t') = \sigma_t^2 \exp(-0.5 \|t - t'\|^s / \rho_t^s) \\
\mathcal{C}_\mathbf{x}(\mathbf{x}, \mathbf{x}') = \sigma_\mathbf{x}^2 \exp(-0.5 \|\mathbf{x} - \mathbf{x}'\|^s / \rho_\mathbf{x}^s), \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}} \mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}} &\dot{\otimes} \mathcal{C}_t(\mathbf{z}, \mathbf{z}') = \sum_{\ell=1}^{\infty} \lambda_\ell(t) \lambda_\ell(t') \phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}') \mathcal{C}_t(t, t') \\
\lambda_\ell(t) = \gamma_\ell u_\ell(t), \ u_\ell(\cdot) &\overset{iid}{\sim} \mathcal{GP}(0, \mathcal{C}_u), \quad \mathcal{C}_u(t, t') = \sigma_u^2 \exp(-0.5 \|t - t'\|^s / \rho_u^s) \\
\sigma_*^2 &\sim \Gamma^{-1}(a_*, b_*), \quad \log \rho_* \sim \mathcal{N}(m_*, V_*), \quad * = \varepsilon, \mathbf{x}, t, \text{ or } u
\end{aligned}
$$

$$\text{(B.1)}$$

and model II in the following

$$
\begin{aligned}
\mathrm{vec}(\mathbf{Y}_k) | \mathbf{M}, \mathbf{C}_{\mathbf{x}|t} &\sim \mathcal{N}(\mathrm{vec}(\mathbf{M}), \mathbf{C}_{\mathbf{x}|t}), \quad \mathbf{M}_{I \times J} = m(\mathbf{X}, \mathbf{t}), \ \mathbf{C}_{\mathbf{x}|t} = \mathcal{C}_{\mathbf{x}|t}(\mathbf{X}, \mathbf{X}; \mathbf{t}) \\
m(\mathbf{x}, t) &\sim \mathcal{GP}(0, \mathcal{I}_\mathbf{x} \otimes \mathcal{C}_t), \quad \mathcal{C}_t(t, t') = \sigma_t^2 \exp(-0.5 \|t - t'\|^s / \rho_t^s) \\
\mathcal{C}_\mathbf{x}(\mathbf{x}, \mathbf{x}') = \sigma_\mathbf{x}^2 \exp(-0.5 \|\mathbf{x} - \mathbf{x}'\|^s / \rho_\mathbf{x}^s), \quad \mathcal{C}_{\mathbf{x}|t}(\mathbf{x}, \mathbf{x}'; t) &= \sum_{\ell=1}^{\infty} \lambda_\ell^2(t) \phi_\ell(\mathbf{x}) \phi_\ell(\mathbf{x}') \\
\lambda_\ell(t) = \gamma_\ell u_\ell(t), \ u_\ell(\cdot) &\overset{iid}{\sim} \mathcal{GP}(0, \mathcal{C}_u), \quad \mathcal{C}_u(t, t') = \sigma_u^2 \exp(-0.5 \|t - t'\|^s / \rho_u^s) \\
\sigma_*^2 &\sim \Gamma^{-1}(a_*, b_*), \quad \log \rho_* \sim \mathcal{N}(m_*, V_*), \quad * = \mathbf{x}, t, \text{ or } u
\end{aligned}
$$

$$\text{(B.2)}$$

Truncate the kernel expansion (16) or (17) at some L terms. We now focus on obtaining the posterior probability of $\mathbf{M}_{I \times J}, \mathbf{\Lambda}_{J \times L}, \boldsymbol{\sigma}^2 := (\sigma_\varepsilon^2, \sigma_\mathbf{x}^2, \sigma_t^2, \sigma_u^2)$ and $\boldsymbol{\rho} := (\rho_\mathbf{x}, \rho_t, \rho_u)$ in the models (B.1) (B.2). Transform the parameters $\boldsymbol{\eta} := \log(\boldsymbol{\rho})$ for the convenience of calculation. Denote $\sigma_\mathbf{z}^2 = (\sigma_\mathbf{x}^2, \sigma_t^2)$, and $\eta_\mathbf{z}^2 = (\eta_\mathbf{x}^2, \eta_t^2)$. Denote $\mathbf{C}_t = \mathcal{C}_t(\mathbf{t}, \mathbf{t})$, and $\mathbf{C}_u = \mathcal{C}_u(\mathbf{t}, \mathbf{t})$. Let $\mathbf{C}_\mathbf{x} := \mathcal{C}_\mathbf{x}(\mathbf{X}, \mathbf{X}) = \boldsymbol{\Phi} \mathbf{\Lambda}_0^2 \boldsymbol{\Phi}^\mathsf{T}$ where $\mathbf{\Lambda}_0 = \mathrm{diag}(\{\lambda_\ell^0\})$. Then $\mathbf{C}_{\mathbf{x}|t}^j := \mathcal{C}_{\mathbf{x}|t_j}(\mathbf{X}, \mathbf{X}) = \boldsymbol{\Phi} \mathrm{diag}(\mathbf{\Lambda}_j^2) \boldsymbol{\Phi}^\mathsf{T}$ where $\mathbf{\Lambda}_j = \{\lambda_{j\ell}\}$ is the j-th row of $\mathbf{\Lambda}$. Denote $\mathbf{C}_*(\sigma_*^2, \eta_*) = \sigma_*^2 \mathbf{C}_{0*}(\eta_*)$ where $* = \mathbf{x}, t, \mathbf{z}$, or u. Once the spatial eigen-basis $\boldsymbol{\Phi}$ has been calculated, it will be shared across all the following calculation. Since only normalized eigen-basis $\boldsymbol{\Phi}(\eta_\mathbf{x})$ is used, we can set $\sigma_\mathbf{x}^2 \equiv 1$ and exclude it from posterior distributions.

Notice that $\mathbf{C}_\mathbf{z} = \mathcal{C}_t \dot{\otimes} \mathcal{C}_{\mathbf{x}|t}^{\frac{1}{2}} \mathcal{C}_{\mathbf{x}|t'}^{\frac{1}{2}}(\mathbf{Z}, \mathbf{Z})$ for model I is a full $IJ \times IJ$ matrix; while $\mathbf{C}_{\mathbf{x}|t} = \mathrm{diag}(\{\mathbf{C}_{\mathbf{x}|t}^j\}_{j=1}^J)$ for model II is a block diagonal matrix formed by J blocks of $I \times I$ matrices. Both $\mathbf{C}_\mathbf{z}(\mathbf{\Lambda})$ and $\mathbf{C}_{\mathbf{x}|t}(\mathbf{\Lambda})$ are defined through the Mercer's expansions with fixed eigen-basis $\boldsymbol{\Phi}$ and newly modeled eigenvalues $\mathbf{\Lambda}$. We make some simplifications before proceeding the calculation of posteriors. Due to the linear independence requirement for $\boldsymbol{\Phi}$, we have $L \leq I$. Therefore $\mathbf{C}_\mathbf{z}$ is in general degenerate, and so is $\mathbf{C}_{\mathbf{x}|t}^j$ if $L < I$.

B.1 Model I

First, for model I (B.1) we have

$$\log p(\mathbf{M}, \boldsymbol{\Lambda}, \boldsymbol{\sigma}^2, \boldsymbol{\eta}|\mathbf{Y})$$
$$= \log p(\mathbf{Y}|\mathbf{M}, \sigma_\varepsilon^2) + \log p(\mathbf{M}|\boldsymbol{\Lambda}, \sigma_t^2, \eta_{\mathbf{z}}) + \log p(\boldsymbol{\Lambda}|\sigma_u^2, \eta_u) + \sum_{*=\varepsilon,t,u} \log p(\sigma_*^2) + \sum_{*=\mathbf{x},t,u} \log p(\eta_*)$$

$$= -\frac{1}{2} IJK \log \sigma_\varepsilon^2 - \frac{\sigma_\varepsilon^{-2}}{2} \sum_{k=1}^K \text{tr}((\mathbf{Y}_k - \mathbf{M})^\mathsf{T}(\mathbf{Y}_k - \mathbf{M}))$$

$$- \frac{1}{2} \log |\mathbf{C}_{\mathbf{z}}(\boldsymbol{\Lambda}, \sigma_t^2, \eta_{\mathbf{z}})| - \frac{1}{2} \text{vec}(\mathbf{M})^\mathsf{T} \mathbf{C}_{\mathbf{z}}^{-1} \text{vec}(\mathbf{M})$$

$$- J\mathbf{1}^\mathsf{T} \log |\boldsymbol{\gamma}(\eta_{\mathbf{x}})| - \frac{L}{2} \log |\mathbf{C}_u(\sigma_u^2, \eta_u)| - \frac{1}{2} \text{tr}(\mathbf{U}^\mathsf{T} \mathbf{C}_u^{-1} \mathbf{U})$$

$$- \sum_{*=\varepsilon,t,u} (a_* + 1) \log \sigma_*^2 + b_* \sigma_*^{-2} - \sum_{*=\mathbf{x},t,u} \frac{1}{2}(\eta_* - m_*)^2 / V_*$$

where $\boldsymbol{\gamma}(\eta_{\mathbf{x}})$ may (chosen as eigenvalues of $\mathcal{C}_{\mathbf{x}}$) or may not (chosen as in (20)) depend on $\eta_{\mathbf{x}}$.

$(\boldsymbol{\sigma}^2)$.   Note the prior for $\boldsymbol{\sigma}^2$ is conditionally conjugate. For $* = \varepsilon, t,$ or u,

$$\sigma_*^2|\cdot \sim \Gamma^{-1}(a_*', b_*'), \quad a_*' = a_* + \Delta_*^a, \quad b_*' = b_* + \Delta_*^b$$
$$\Delta_\varepsilon^a = \frac{1}{2} IJK, \quad \Delta_t^a = \frac{1}{2} IJ, \quad \Delta_u^a = \frac{1}{2} JL$$
$$\Delta_\varepsilon^b = \frac{1}{2} \sum_{k=1}^K \text{tr}((\mathbf{Y}_k - \mathbf{M})^\mathsf{T}(\mathbf{Y}_k - \mathbf{M})), \quad \Delta_t^b = \frac{1}{2} \text{vec}(\mathbf{M})^\mathsf{T} \mathbf{C}_{0\mathbf{z}}^{-1} \text{vec}(\mathbf{M}), \quad \Delta_u^b = \frac{1}{2} \text{tr}(\mathbf{U}^\mathsf{T} \mathbf{C}_{0u}^{-1} \mathbf{U})$$

$(\boldsymbol{\eta})$.   Given $* = \mathbf{x}, t,$ or u, we could sample η_* using the slice sampler (Neal, 2003), which only requires log-posterior density and works well for scalar parameters,

$$\log p(\eta_{\mathbf{x}}|\cdot) = -\frac{1}{2} \log |\mathbf{C}_{0\mathbf{z}}(\boldsymbol{\Lambda}, \eta_{\mathbf{z}})| - \frac{1}{2} \text{vec}(\mathbf{M})^\mathsf{T} \mathbf{C}_{0\mathbf{z}}^{-1} \text{vec}(\mathbf{M}) \sigma_t^{-2} - J\mathbf{1}^\mathsf{T} \log |\boldsymbol{\gamma}(\eta_{\mathbf{x}})| - \frac{1}{2}(\eta_{\mathbf{x}} - m_{\mathbf{x}})^2 / V_{\mathbf{x}}$$
$$\log p(\eta_t|\cdot) = -\frac{1}{2} \log |\mathbf{C}_{0\mathbf{z}}(\boldsymbol{\Lambda}, \eta_{\mathbf{z}})| - \frac{1}{2} \text{vec}(\mathbf{M})^\mathsf{T} \mathbf{C}_{0\mathbf{z}}^{-1} \text{vec}(\mathbf{M}) \sigma_t^{-2} - \frac{1}{2}(\eta_t - m_t)^2 / V_t$$
$$\log p(\eta_u|\cdot) = -\frac{L}{2} \log |\mathbf{C}_{0u}(\eta_u)| - \frac{1}{2} \text{tr}(\boldsymbol{\Lambda}^\mathsf{T} \mathbf{C}_{0u}^{-1} \boldsymbol{\Lambda}) \sigma_u^{-2} - \frac{1}{2}(\eta_u - m_u)^2 / V_u$$

$(\mathbf{M})$.   By the definition of STGP prior, we have $\text{vec}(\mathbf{M})|\boldsymbol{\Lambda}, \sigma_t^2, \eta_{\mathbf{z}} \sim \mathcal{N}_{IJ}(\mathbf{0}, \mathbf{C}_{\mathbf{z}}(\boldsymbol{\Lambda}, \sigma_t^2, \eta_{\mathbf{z}}))$. On the other hand, one can write the log-likelihood function as

$$\log p(\mathbf{M}; \mathbf{Y}) = -\frac{\sigma_\varepsilon^{-2}}{2} \sum_{k=1}^K \text{tr}((\mathbf{Y}_k - \mathbf{M})^\mathsf{T}(\mathbf{Y}_k - \mathbf{M}))$$
$$= -\frac{1}{2} \sum_{k=1}^K (\text{vec}(\mathbf{Y}_k) - \text{vec}(\mathbf{M}))^\mathsf{T} (\sigma_\varepsilon^2 \mathbf{I})^{-1} (\text{vec}(\mathbf{Y}_k) - \text{vec}(\mathbf{M}))$$

Therefore we have the analytic posterior

$$\text{vec}(\mathbf{M})|\cdot \sim \mathcal{N}_{ND}(\mathbf{M}', \mathbf{C}'), \quad \mathbf{M}' = \mathbf{C}'\sigma_\varepsilon^{-2}\sum_{k=1}^{K}\text{vec}(\mathbf{Y}_k),$$

$$\mathbf{C}' = \left(\mathbf{C}_\mathbf{z}^{-1} + K\sigma_\varepsilon^{-2}\mathbf{I}\right)^{-1} = \mathbf{C}_\mathbf{z}\left(\mathbf{C}_\mathbf{z} + K^{-1}\sigma_\varepsilon^2\mathbf{I}\right)^{-1}K^{-1}\sigma_\varepsilon^2\mathbf{I}$$

($\mathbf{\Lambda}$). Using a similar argument by matrix Normal prior for $\mathbf{\Lambda}$, we have $\mathbf{\Lambda}|\sigma_u^2, \eta_u \sim \mathcal{MN}(\mathbf{0}, \mathbf{C}_u(\sigma_u^2, \eta_u), \text{diag}(\boldsymbol{\gamma}^2))$. Therefore, we could use the elliptic slice sampler (ESS, Murray et al., 2010), which only requires the log-likelihood

$$\log p(\mathbf{\Lambda}; \mathbf{M}) = -\frac{1}{2}\log|\mathbf{C}_\mathbf{z}(\mathbf{\Lambda}, \sigma_t^2, \eta_\mathbf{z})| - \frac{1}{2}\text{vec}(\mathbf{M})^\mathsf{T}\mathbf{C}_\mathbf{z}^{-1}\text{vec}(\mathbf{M})$$

B.2 Model II

Now we consider model II (B.2)

$$\log p(\mathbf{M}, \mathbf{\Lambda}, \boldsymbol{\sigma}^2, \boldsymbol{\eta}|\mathbf{Y})$$

$$= \log p(\mathbf{Y}|\mathbf{M}, \mathbf{C}_{\mathbf{x}|t}(\mathbf{\Lambda}, \eta_\mathbf{x})) + \log p(\mathbf{M}|\sigma_t^2, \eta_t) + \log p(\mathbf{\Lambda}|\sigma_u^2, \eta_u) + \sum_{*=t,u}\log p(\sigma_*^2) + \sum_{*=\mathbf{x},t,u}\log p(\eta_*)$$

$$= -\frac{K}{2}\log|\mathbf{C}_{\mathbf{x}|t}(\mathbf{\Lambda}, \eta_\mathbf{x})| - \frac{1}{2}\sum_{k=1}^{K}\text{vec}(\mathbf{Y}_k - \mathbf{M})^\mathsf{T}\mathbf{C}_{\mathbf{x}|t}^{-1}\text{vec}(\mathbf{Y}_k - \mathbf{M})$$

$$- \frac{I}{2}\log|\mathbf{C}_t(\sigma_t^2, \eta_t)| - \frac{1}{2}\text{tr}(\mathbf{C}_t^{-1}\mathbf{M}^\mathsf{T}\mathbf{M})$$

$$- J\mathbf{1}^\mathsf{T}\log|\boldsymbol{\gamma}(\eta_\mathbf{x})| - \frac{L}{2}\log|\mathbf{C}_u(\sigma_u^2, \eta_u)| - \frac{1}{2}\text{tr}(\mathbf{U}^\mathsf{T}\mathbf{C}_u^{-1}\mathbf{U})$$

$$- \sum_{*=t,u}(a_* + 1)\log\sigma_*^2 + b_*\sigma_*^{-2} - \sum_{*=\mathbf{x},t,u}\frac{1}{2}(\eta_* - m_*)^2/V_*$$

($\boldsymbol{\sigma}^2$). Note the prior for $\boldsymbol{\sigma}^2$ is conditionally conjugate. For $* = t$ or u,

$$\sigma_*^2|\cdot \sim \Gamma^{-1}(a_*', b_*'), \quad a_*' = a_* + \Delta_*^a, \quad b_*' = b_* + \Delta_*^b$$

$$\Delta_t^a = \frac{1}{2}IJ, \quad \Delta_u^a = \frac{1}{2}JL, \quad \Delta_t^b = \frac{1}{2}\text{tr}(\mathbf{C}_{0t}^{-1}\mathbf{M}^\mathsf{T}\mathbf{M}), \quad \Delta_u^b = \frac{1}{2}\text{tr}(\mathbf{U}^\mathsf{T}\mathbf{C}_{0u}^{-1}\mathbf{U})$$

($\boldsymbol{\eta}$). Given $* = \mathbf{x}, t$, or u, we could sample η_* using the slice sampler (Neal, 2003), which only requires log-posterior density and works well for scalar parameters,

$$\log p(\eta_\mathbf{x}|\cdot) = -\frac{K}{2}\log|\mathbf{C}_{\mathbf{x}|t}(\mathbf{\Lambda}, \eta_\mathbf{x})| - \frac{1}{2}\sum_{k=1}^{K}\text{vec}(\mathbf{Y}_k - \mathbf{M})^\mathsf{T}\mathbf{C}_{\mathbf{x}|t}^{-1}\text{vec}(\mathbf{Y}_k - \mathbf{M})$$

$$- J\mathbf{1}^\mathsf{T}\log|\boldsymbol{\gamma}(\eta_\mathbf{x})| - \frac{1}{2}(\eta_\mathbf{x} - m_\mathbf{x})^2/V_\mathbf{x}$$

$$\log p(\eta_t|\cdot) = -\frac{I}{2}\log|\mathbf{C}_{0t}(\eta_t)| - \frac{1}{2}\text{tr}(\mathbf{C}_{0t}^{-1}\mathbf{M}^\mathsf{T}\mathbf{M})\sigma_t^{-2} - \frac{1}{2}(\eta_t - m_t)^2/V_t$$

$$\log p(\eta_u|\cdot) = -\frac{L}{2}\log|\mathbf{C}_{0u}(\eta_u)| - \frac{1}{2}\text{tr}(\mathbf{\Lambda}^\mathsf{T}\mathbf{C}_{0u}^{-1}\mathbf{\Lambda})\sigma_u^{-2} - \frac{1}{2}(\eta_u - m_u)^2/V_u$$

(**M**). By the definition of STGP prior, we have $\mathrm{vec}(\mathbf{M})|\sigma_t^2, \eta_t \sim \mathcal{N}_{IJ}(\mathbf{0}, \mathbf{C}_t(\sigma_t^2, \eta_t) \otimes \mathbf{I_x})$. On the other hand, one can write the log-likelihood function as

$$\log p(\mathbf{M}; \mathbf{Y}) = -\frac{1}{2}\sum_{k=1}^{K} \mathrm{vec}(\mathbf{Y}_k - \mathbf{M})^\mathsf{T} \mathbf{C}_{\mathbf{x}|t}^{-1} \mathrm{vec}(\mathbf{Y}_k - \mathbf{M})$$

$$= -\frac{1}{2}\sum_{k=1}^{K} (\mathrm{vec}(\mathbf{M}) - \mathrm{vec}(\mathbf{Y}_k))^\mathsf{T} \mathbf{C}_{\mathbf{x}|t}^{-1} (\mathrm{vec}(\mathbf{M}) - \mathrm{vec}(\mathbf{Y}_k))$$

Therefore we have the analytic posterior

$$\mathrm{vec}(\mathbf{M})|\cdot \sim \mathcal{N}_{ND}(\mathbf{M}', \mathbf{C}'), \quad \mathbf{M}' = \mathbf{C}' \mathbf{C}_{\mathbf{x}|t}^{-1} \sum_{k=1}^{K} \mathrm{vec}(\mathbf{Y}_k),$$

$$\mathbf{C}' = \left(\mathbf{C}_t^{-1} \otimes \mathbf{I_x} + K \mathbf{C}_{\mathbf{x}|t}^{-1}\right)^{-1} = (\mathbf{C}_t \otimes \mathbf{I_x})\left(\mathbf{C}_t \otimes \mathbf{I_x} + K^{-1}\mathbf{C}_{\mathbf{x}|t}\right)^{-1} K^{-1}\mathbf{C}_{\mathbf{x}|t}$$

(**Λ**). Using a similar argument by matrix Normal prior for $\mathbf{\Lambda}$, we have $\mathbf{\Lambda}|\sigma_u^2, \eta_u \sim \mathcal{MN}(\mathbf{0}, \mathbf{C}_u(\sigma_u^2, \eta_u), \mathrm{diag}(\boldsymbol{\gamma}^2))$. Therefore, we could use the elliptic slice sampler (ESS, Murray et al., 2010), which only requires the log-likelihood

$$\log p(\mathbf{\Lambda}; \mathbf{Y}) = -\frac{K}{2}\log|\mathbf{C}_{\mathbf{x}|t}(\mathbf{\Lambda}, \eta_{\mathbf{x}})| - \frac{1}{2}\sum_{k=1}^{K} \mathrm{vec}(\mathbf{Y}_k - \mathbf{M})^\mathsf{T} \mathbf{C}_{\mathbf{x}|t}^{-1} \mathrm{vec}(\mathbf{Y}_k - \mathbf{M})$$

B.3 Computational Advantage of Model II

The most intensive computation as above involves the inverse and determinant of the posterior covariance kernel $\mathbf{C}^*$ for two models:

$$\mathbf{C}_\mathrm{I}^* := \mathbf{C}_\mathbf{z}(\mathbf{\Lambda}, \sigma_t^2, \eta_\mathbf{z}) + K^{-1}\sigma_\varepsilon^2 \mathbf{I}, \qquad \mathbf{C}_\mathrm{II}^* := \mathbf{C}_t(\sigma_t^2, \eta_t) \otimes \mathbf{I_x} + K^{-1}\mathbf{C}_{\mathbf{x}|t}(\mathbf{\Lambda}, \eta_\mathbf{x})$$

Their structure dictates different amount of computation required. Actually, we can show that the kernel of model II, $\mathbf{C}_\mathrm{II}^*$, has computational advantage over that for model I.

Note, according to the definition of the dynamic spatial kernel (16), we can rewrite

$$\mathbf{C}_{\mathbf{x}|t} = \mathrm{diag}(\{\mathbf{C}_{\mathbf{x}|t}^j\}_{j=1}^J) = \mathrm{diag}(\{\mathbf{\Phi}\,\mathrm{diag}(\mathbf{\Lambda}_j^2)\mathbf{\Phi}^\mathsf{T}\}) = (\mathbf{I}_t \otimes \mathbf{\Phi})\,\mathrm{diag}(\mathrm{vec}^\mathsf{T}(\mathbf{\Lambda}^2))(\mathbf{I}_t \otimes \mathbf{\Phi}^\mathsf{T})$$

where $\mathrm{vec}^\mathsf{T}(\cdot)$ is row-wise vectorization. Then by the Sherman-Morrison-Woodbury formula we have

$$(\mathbf{C}_\mathrm{II}^*)^{-1} = (\mathbf{C}_t^{-1} \otimes \mathbf{I_x}) - (\mathbf{C}_t^{-1} \otimes \mathbf{I_x})(\mathbf{I}_t \otimes \mathbf{\Phi})\cdot$$
$$[K\,\mathrm{diag}(\mathrm{vec}^\mathsf{T}(\mathbf{\Lambda}^{-2})) + (\mathbf{I}_t \otimes \mathbf{\Phi}^\mathsf{T})(\mathbf{C}_t^{-1} \otimes \mathbf{I_x})(\mathbf{I}_t \otimes \mathbf{\Phi})]^{-1}(\mathbf{I}_t \otimes \mathbf{\Phi}^\mathsf{T})(\mathbf{C}_t^{-1} \otimes \mathbf{I_x})$$
$$= (\mathbf{C}_t^{-1} \otimes \mathbf{I_x}) - (\mathbf{C}_t^{-1} \otimes \mathbf{\Phi})[K\,\mathrm{diag}(\mathrm{vec}^\mathsf{T}(\mathbf{\Lambda}^{-2})) + (\mathbf{C}_t^{-1} \otimes \mathbf{I}_L)]^{-1}(\mathbf{C}_t^{-1} \otimes \mathbf{\Phi}^\mathsf{T})$$
$$= \mathbf{C}_t^{-1} \otimes (\mathbf{I_x} - \mathbf{\Phi}\mathbf{\Phi}^\mathsf{T}) + (\mathbf{I}_t \otimes \mathbf{\Phi})[K^{-1}\,\mathrm{diag}(\mathrm{vec}^\mathsf{T}(\mathbf{\Lambda}^2)) + (\mathbf{C}_t \otimes \mathbf{I}_L)]^{-1}(\mathbf{I}_t \otimes \mathbf{\Phi}^\mathsf{T})$$

Similarly we have

$$\mathbf{C}' = \mathbf{C}_t \otimes (\mathbf{I_x} - \mathbf{\Phi}\mathbf{\Phi}^\mathsf{T}) + (\mathbf{I}_t \otimes \mathbf{\Phi})[K\,\mathrm{diag}(\mathrm{vec}^\mathsf{T}(\mathbf{\Lambda}^{-2})) + (\mathbf{C}_t^{-1} \otimes \mathbf{I}_L)]^{-1}(\mathbf{I}_t \otimes \mathbf{\Phi}^\mathsf{T})$$
$$(\mathbf{C}')^{\frac{1}{2}} = \mathbf{C}_t^{\frac{1}{2}} \otimes (\mathbf{I_x} - \mathbf{\Phi}\mathbf{\Phi}^\mathsf{T}) + (\mathbf{I}_t \otimes \mathbf{\Phi})[K\,\mathrm{diag}(\mathrm{vec}^\mathsf{T}(\mathbf{\Lambda}^{-2})) + (\mathbf{C}_t^{-1} \otimes \mathbf{I}_L)]^{-\frac{1}{2}}(\mathbf{I}_t \otimes \mathbf{\Phi}^\mathsf{T})$$

where we use the following calculation that is numerically more stable

$$[K \operatorname{diag}(\operatorname{vec}^\mathsf{T}(\mathbf{\Lambda}^{-2})) + (\mathbf{C}_t^{-1} \otimes \mathbf{I}_L)]^{-1} = K^{-1} \operatorname{diag}(\operatorname{vec}^\mathsf{T}(\mathbf{\Lambda}^2)$$
$$- K^{-1} \operatorname{diag}(\operatorname{vec}^\mathsf{T}(\mathbf{\Lambda}^2)[K^{-1} \operatorname{diag}(\operatorname{vec}^\mathsf{T}(\mathbf{\Lambda}^2)) + (\mathbf{C}_t \otimes \mathbf{I}_L)]^{-1} K^{-1} \operatorname{diag}(\operatorname{vec}^\mathsf{T}(\mathbf{\Lambda}^2)$$

Based on the matrix determinant lemma we can calculate

$$\begin{aligned}
\det(\mathbf{C}_{\mathrm{II}}^*) &= \det[K \operatorname{diag}(\operatorname{vec}^\mathsf{T}(\mathbf{\Lambda}^{-2})) + (\mathbf{I}_t \otimes \mathbf{\Phi}^\mathsf{T})(\mathbf{C}_t^{-1} \otimes \mathbf{I}_\mathbf{x})(\mathbf{I}_t \otimes \mathbf{\Phi})] \\
&\quad \det(K^{-1} \operatorname{diag}(\operatorname{vec}^\mathsf{T}(\mathbf{\Lambda}^2))) \det(\mathbf{C}_t \otimes \mathbf{I}_\mathbf{x}) \\
&= \det[\operatorname{diag}(\operatorname{vec}^\mathsf{T}(\mathbf{\Lambda}^{-2})) + K^{-1}(\mathbf{C}_t^{-1} \otimes \mathbf{I}_L)] \prod_{j,\ell} \lambda_{j\ell}^2 \det(\mathbf{C}_t)^I \\
&= \det[\mathbf{C}_t \otimes \mathbf{I}_L + K^{-1} \operatorname{diag}(\operatorname{vec}^\mathsf{T}(\mathbf{\Lambda}^2))] \det(\mathbf{C}_t)^{I-L}
\end{aligned}$$

However in model I, we note that $\mathbf{C_z} = \mathbf{C}_{\mathbf{x}|t}^{\frac{1}{2}} \mathbf{C}_{\mathbf{x}|t'}^{\frac{1}{2}} \odot (\mathbf{C}_t \otimes \mathbf{1}_{I \times I})$, where $\odot$ is element-wise multiplication, and $\mathbf{1}_{I \times I}$ is an $I \times I$ matrix with all elements 1. According to (17), we have

$$\mathbf{C}_{\mathbf{x}|t}^{\frac{1}{2}} \mathbf{C}_{\mathbf{x}|t'}^{\frac{1}{2}} = [\mathbf{\Phi} \operatorname{diag}(\mathbf{\Lambda}_j) \operatorname{diag}(\mathbf{\Lambda}_{j'}) \mathbf{\Phi}^\mathsf{T}] = (\mathbf{I}_t \otimes \mathbf{\Phi}) \operatorname{vec}\{\operatorname{diag}(\mathbf{\Lambda}_j)\} \operatorname{vec}^\mathsf{T}\{\operatorname{diag}(\mathbf{\Lambda}_{j'})\}(\mathbf{I}_t \otimes \mathbf{\Phi}^\mathsf{T})$$

where $\operatorname{vec}\{\cdot\}$ and $\operatorname{vec}^\mathsf{T}\{\cdot\}$ are column/row wise vectorization of block matrices. Applying the above equation to the inverse or determinant of $\mathbf{C}_{\mathrm{I}}^*$ does not simplify computation in general.

B.4 Spatial Kernel Based On Graph Laplacian

Graph Laplacian, also known as discrete Laplace operator, is a matrix representation of a graph. It is a popular tool for image processing, clustering and semi-supervised/unsupervised learning on graphs (Chung and Graham, 1997; Smola and Kondor, 2003). For a weighted graph $G = (Z, W)$ with Z being the vertices $\{x_i\}_{i=1}^n$ of the graph and W being the edge weight matrix, the graph Laplacian L is defined as follows

$$\mathrm{L} = D - W, \quad W = [w_{ij}], \ w_{ij} = \eta_\varepsilon(|x_i - x_j|), \quad D = \operatorname{diag}\{d_{ii}\}, \ d_{ii} = \sum_{x_j \sim x_i} w_{ij} \qquad \text{(B.3)}$$

where η_ε is some distance function, e.g. Euclidean distance, D is called degree matrix, and $x_i \sim x_j$ means two vertices x_i, x_j connected with an edge. When $w_{ij} \equiv 1$, W is also called adjacency matrix, denoted as A. If we assume $x_j \in \Omega$ are sampled i.i.d from a probability measure μ supported on the graph domain Ω with smooth Lebesgue density ρ bounded above and below by positive constants, then L can be viewed as an approximation of the Laplace operator $\mathcal{L}$ in the following PDE:

$$\mathcal{L}u = -\frac{1}{\rho}\nabla \cdot (\rho^2 \nabla u), \quad x \in \Omega, \qquad \frac{\partial u}{\partial n} = 0, \quad x \in \partial\Omega.$$

Based on the graph Laplacian, we can define the following discrete spatial kernel for the brain images (Dunlop et al., 2020)

$$C_\mathbf{x} = (s_n L + \tau^2 I)^{-s}, \quad s_n = o\left(\frac{1}{n^{1-2/d} \log^{\delta(d=2)/2+2/d} n}\right) \qquad \text{(B.4)}$$

(a) Model 0 (stat-sep)

(b) Model I (qKron-prod)

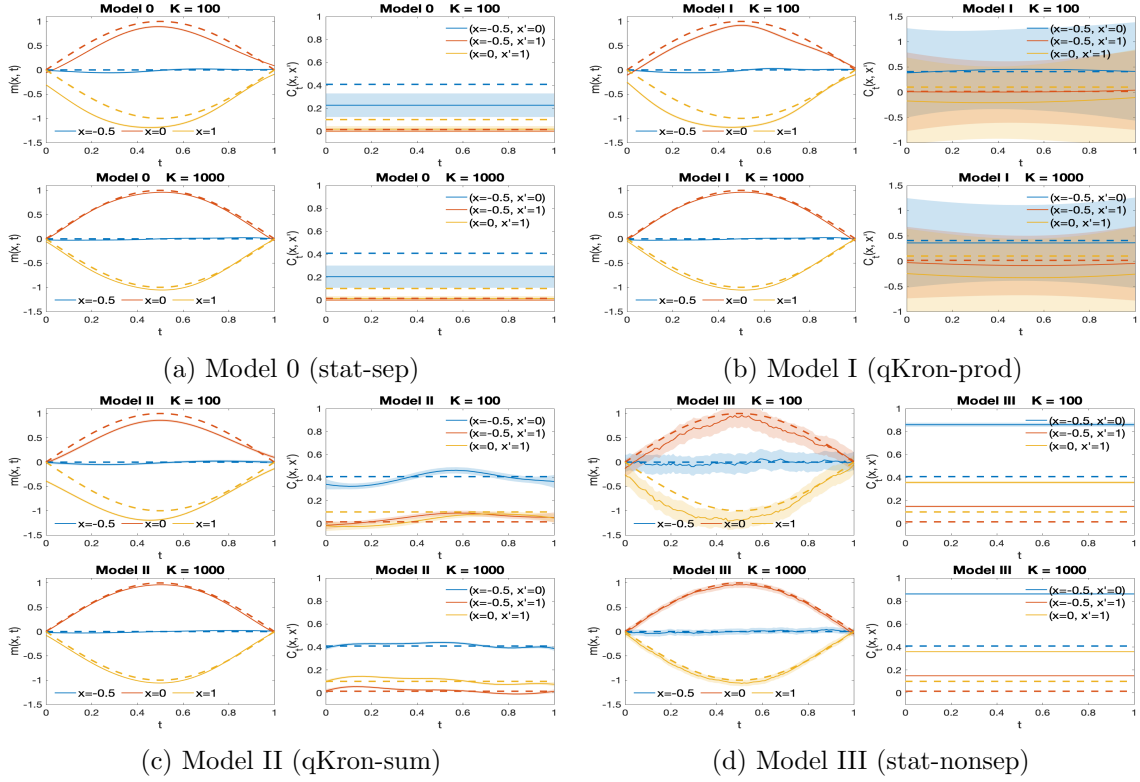(c) Model II (qKron-sum)

(d) Model III (stat-nonsep)

Figure B.1: Stationary Process: selective mean functions $m(x, t)$ (left column) and covariance functions $C^s_{y|t}(x, x')$ (right column) fitted by (a) model 0, (b) model I, (c) model II and (d) model III with $K = 100$ trials of data (upper row) and $K = 1000$ trials of data (lower row) on each panel. Dashed lines are true values, solid curves are estimates with shaded credible regions indicating their uncertainty.

where d is the spatial dimension, i.e. $d = 2$ for the chosen slice of brain images. We choose $s_n = \frac{1}{n^{1-2/d} \log^{1+2/d} n}$ in this experiment. Further assuming conditions (open, connected, and with smooth boundary) on the graph domain Ω, Dunlop et al. (2020) prove that for $s > d/2$ and $\tau \geq 0$, Gaussian measure $\mathcal{N}(0, \mathcal{C})$ with $\mathcal{C} = (\mathcal{L} + \tau^2 \mathcal{I})^{-s}$ is well-defined on the weighted Hilbert space L^2_μ. In another word, the graph-Laplacian based spatial kernel (B.4) is well-behaved for large graphs including the brain images we investigate with $n = 25600$ nodes.

To obtain the spatial kernel (B.4) for the analysis of PET scans, we first construct the graph Laplacian. On the 160×160 mesh grid, each node is connected to its $(2w+1)^2 - 1$ neighbors, where we can choose $w = 1$ for example. Depending on the location, some nodes may have $2w(w+1)+w$ neighbors (on the edge) or $w(w+1)+w$ neighbors (at the corner). The resulting graph Laplacian matrix L has the size 25600×25600 but is highly sparse (with the density of non-zero entries 3.4864×10^{-4}). We also assume a hyper-prior for $\tau^2 \sim \log -\mathcal{N}(m_{\mathbf{x}}, V_{\mathbf{x}})$ and fix $s = 2$ in this experiment. Then for given τ^2, we calculate the precision matrix $C^{-1}_{\mathbf{x}}$ based on (B.4). Hence the dense covariance matrix $C_{\mathbf{x}}$ of size 25600×25600 is not directly calculated in the inference procedure.
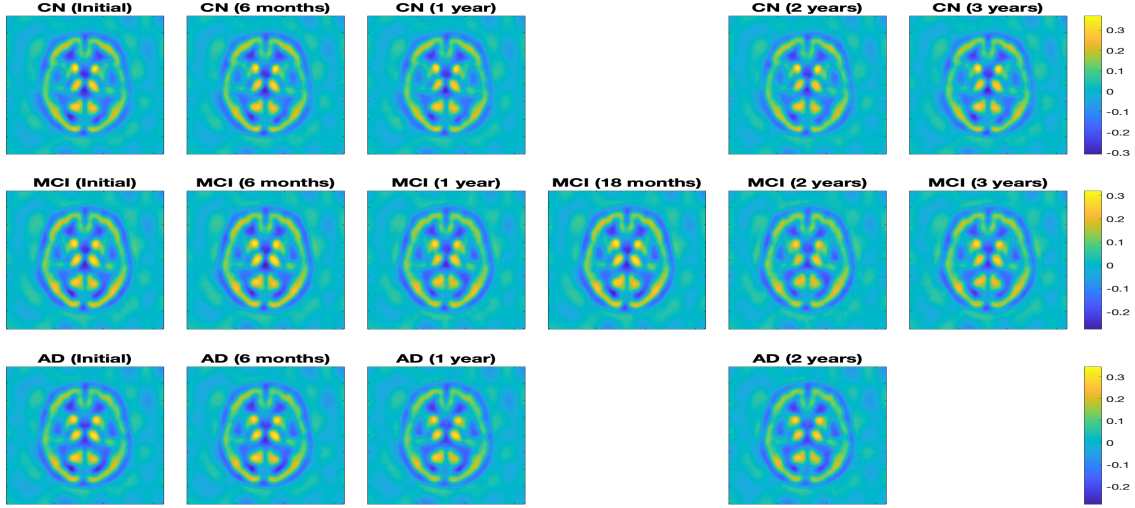
Figure C.1: Estimated brain images of ADNI patients as a function of time for CN (top row), MCI (middle row) and AD (bottom row) respectively by the proposed model II (qKron-sum).

Appendix C. More Numerical Results

C.1 Simulated Stationary Non-separable Process

For the covariance $\mathcal{C}_y$ in the model (34), we specify the following stationary and non-separable kernel:

$$\mathcal{C}_y^{\mathrm{s}}(\mathbf{z}, \mathbf{z}') = \exp\left(-\frac{|x - x'|^2}{2\ell_x} - \frac{|t - t'|^2}{2\ell_t} - \frac{|x - x'|}{2\ell_{xt}(|t - t'| + 1)}\right) \frac{1}{|t - t'| + 1} + \sigma_\varepsilon^2 \delta(\mathbf{z} = \mathbf{z}') \tag{C.1}$$

$$\tag{C.2}$$

where the stationary non-separable covariance is modified from Gneiting (2002). We have the following true TESD constant over time:

$$C_{y|t}^{\mathrm{s}}(x, x') := \mathrm{Cov}[y(x, t), y(x', t)] = \exp\left(-\frac{|x - x'|^2}{2\ell_x} - \frac{|x - x'|}{2\ell_{xt}}\right) + \sigma_\varepsilon^2 \delta(x = x') \tag{C.3}$$

Estimates generated by MCMC samples are plotted at selective locations in Figure B.1. All models produce estimates mean functions contracting to the truth. However, only model II gives faithful estimates of covariance functions (TESD). Note, despite of the first two separable terms in (C.1), the stationary data are generated mostly according to stat-nonsep model (31) with $\sigma^2 = 1$ and $c = \frac{1}{2\ell_{xt}}$, yet stat-nonsep model is still not flexible enough to correctly estimate constant TESD.
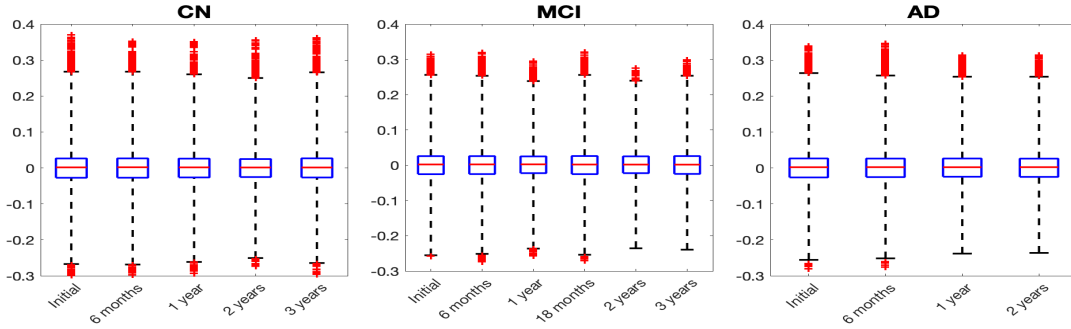
43

Figure C.2: Pixel values of the estimated brain images for CN (left), MCI (middle) and AD (right) respectively indicate more hollow area increased with time in the latter two groups.
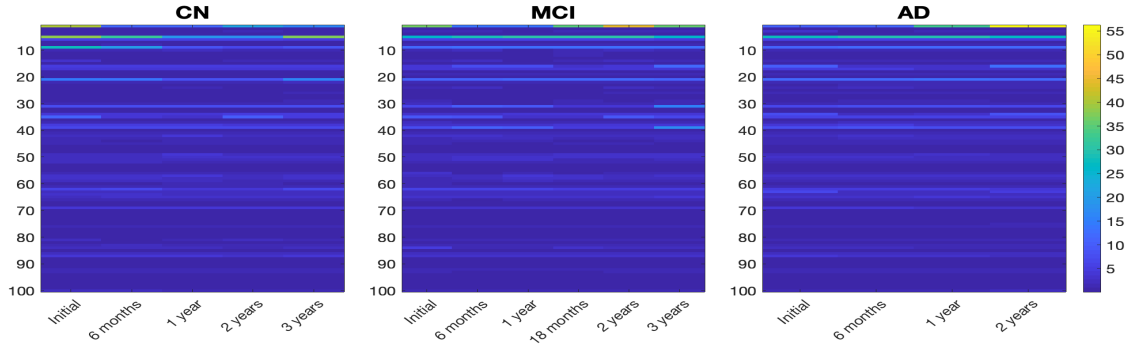

Figure C.3: Dynamic eigenvalues in the generalized STGP model analyzing PET brain images for CN (left), MCI (middle) and AD (right) respectively by the proposed model II (qKron-sum).

C.2 Longitudinal Analysis of Brain Images

Figure C.2 shows that the highest quantiles (horizontal bars) of AD patients decrease with time. This means there are increasing 'hollow' area in these brain images (especially in the MCI and AD groups) as time goes by, indicating the brain shrinkage.

Figure C.3 compares the dynamic eigenvalues $\lambda_\ell^2(t)$ for different groups. Interestingly, they do not decrease monotonically in ℓ (on y-axis) but rather damp out as ℓ becomes larger. When ℓ gets close to $L = 100$, the magnitude of $\lambda_\ell^2(t)$ becomes small enough to be negligible.
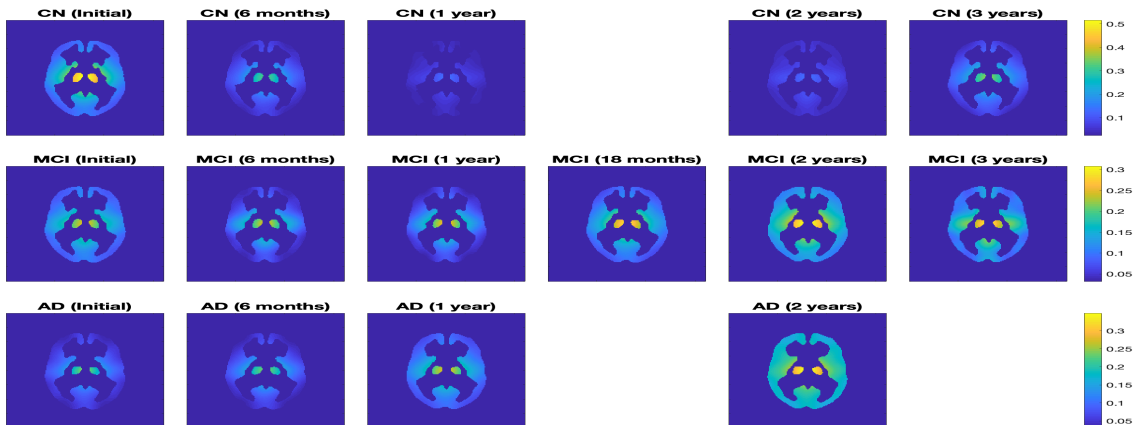
Figure C.4: Estimated variance of the brain images for CN (top row), MCI (middle row) and AD (bottom row) respectively by the proposed model II (qKron-sum).

In Section 5.2.2 we summarize the correlation between the brain ROI and POI. In fact, we have more results regarding TESD presented in different forms. Figure C.4 shows the estimated variances of the brain images as functions of time. They are all small across different groups with small variation along the time. Comparatively, the thalamus and some part of the temporal lobe are more active than the rest of the brain.

In Figure C.5, we compare the estimated ROI-POI correlation on a 40×40 mesh by various spatiotemporal models. All fail to capture the time evolution of such correlations except the proposed qKron-sum model.

Finally, we consider the problem of extending TESD to new locations, infeasible in dynamic covariance models. We coarsen the mesh by using every other pixel and build our model based on the resulted 80×80 images. Figure C.7 compares the estimated ROI-POI correlation on the coarse mesh (upper row) and the prediction to the original 160×160 mesh (lower row) which is consistent with the estimation result in Figure 7. Such extension provides more fine details of TESD at new locations without data. They all illustrate the benefit of a fully nonparametric approach in modeling TESD in the spatiotemporal data.
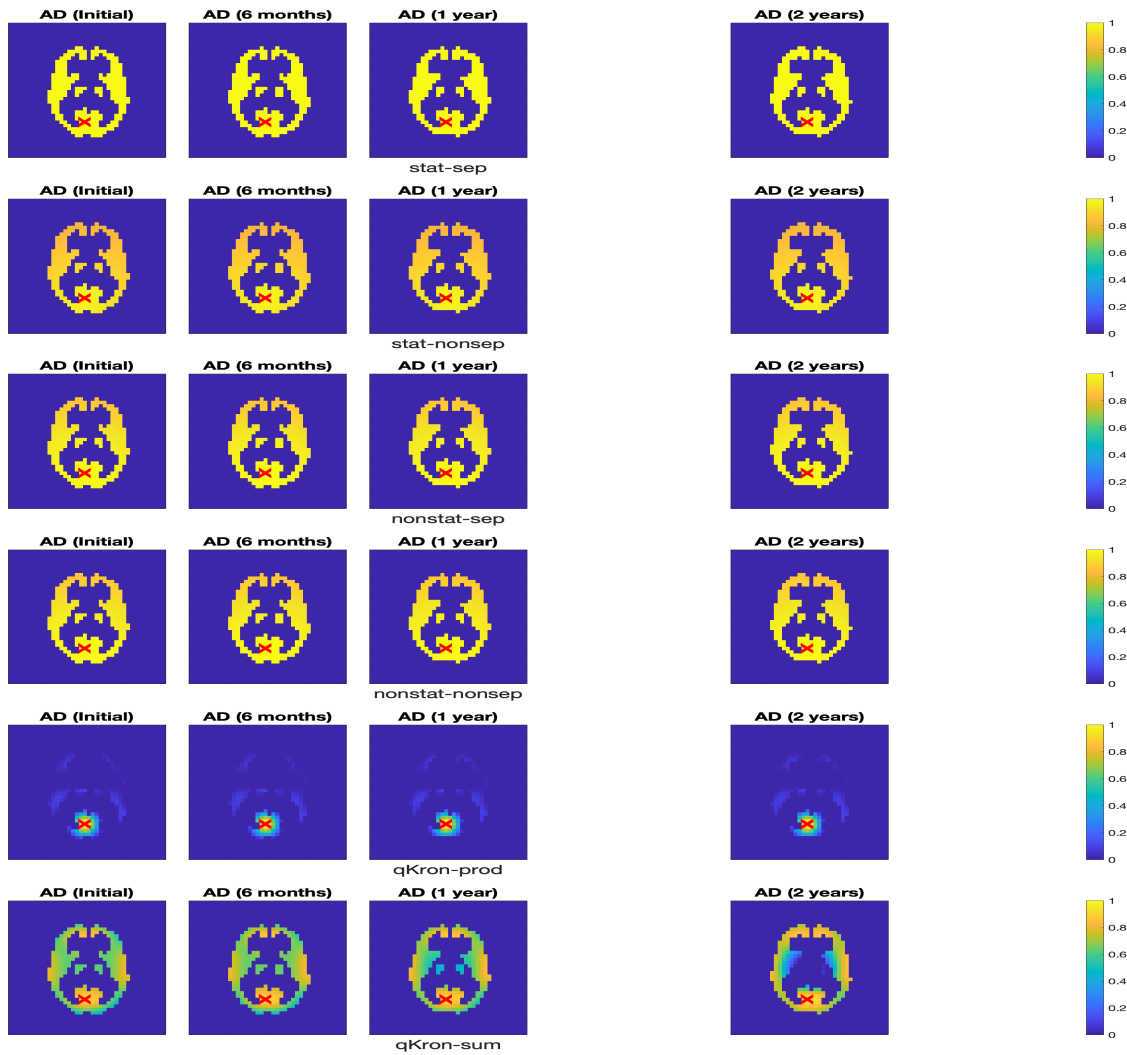
Figure C.5: Estimated correlation between the brain region of interest and a point of interest of the brain images for AD by various models.
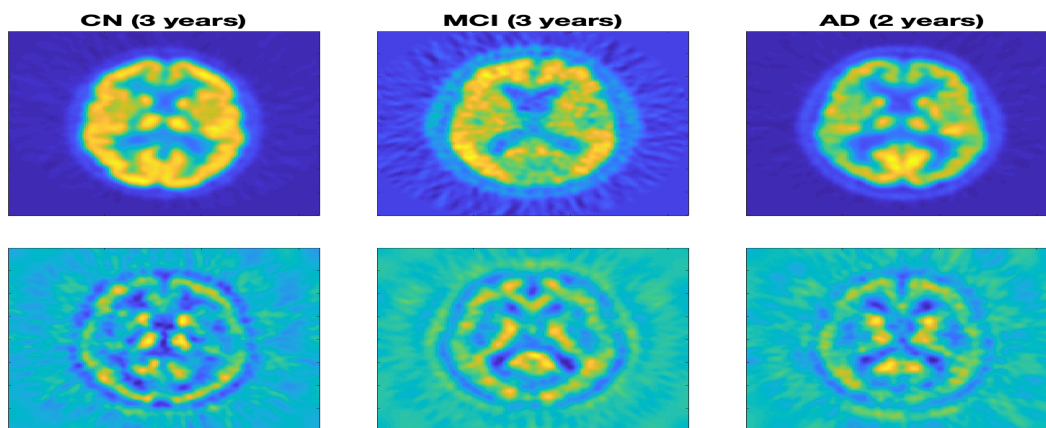
Figure C.6: Prediction of the brain images at the last time point for CN (left column), MCI (middle column) and AD (right column) respectively by the proposed model II (qKron-sum). The upper row shows individuals' brain images; the model outputs are displayed in the lower row.
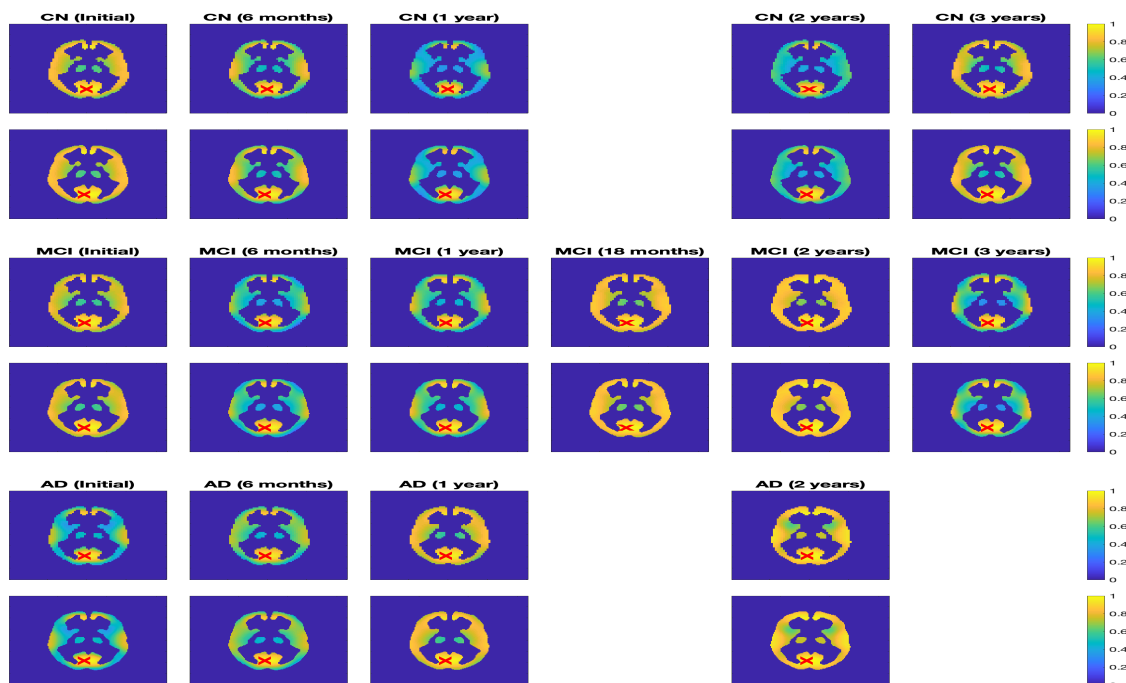

Figure C.7: Extended correlation between the brain region of interest and a point of interest from the coarse mesh (upper) to the fine mesh (lower) in each of CN, MCI and AD groups by the proposed model II (qKron-sum).

References

ADNI. The alzheimer's disease neuroimaging initiative. http://adni.loni.usc.edu, 2003.

Sudipto P Banerjee. *Hierarchical modeling and analysis for spatial data.* Chapman & Hall/CRC, Boca Raton (Fla.); London; New York [etc], 2nd edition. edition, 2015. ISBN 9781439819173.

F.R.K. Chung and F.C. Graham. *Spectral Graph Theory.* Number no. 92 in CBMS Regional Conference Series. American Mathematical Society, 1997. ISBN 9780821803158. URL `https://books.google.com/books?id=4IK8DgAAQBAJ`.

N. Cressie and C.K. Wikle. *Statistics for Spatio-Temporal Data.* CourseSmart Series. Wiley, 2011. ISBN 9780471692744. URL `https://books.google.com/books?id=-kOC6D0DiNYC`.

Noel Cressie and Hsin-Cheng Huang. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1339, 1999. doi: 10.1080/01621459.1999.10473885. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473885`.

Ivor Cribben, Ragnheidur Haraldsdottir, Lauren Y. Atlas, Tor D. Wager, and Martin A. Lindquist. Dynamic connectivity regression: Determining state-related changes in brain connectivity. *NeuroImage*, 61(4):907 – 920, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2012.03.070. URL `http://www.sciencedirect.com/science/article/pii/S1053811912003515`.

Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL `https://proceedings.mlr.press/v31/damianou13a.html`.

Moumita Das and Sourabh Bhattacharya. Nonstationary, nonparametric, nonseparable bayesian spatio-temporal modeling using kernel convolution of order based dependent dirichlet process, 2020.

Masoumeh Dashti and Andrew M. Stuart. The bayesian approach to inverse problems. *Handbook of Uncertainty Quantification*, pages 311–428, 2017. doi: 10.1007/978-3-319-12385-1_7. URL `http://dx.doi.org/10.1007/978-3-319-12385-1_7`.

Abhirup Datta, Sudipto Banerjee, Andrew O. Finley, Nicholas A. S. Hamm, and Martijn Schaap. Nonseparable dynamic nearest neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The Annals of Applied Statistics*, 10(3):1286–1316, Sep 2016. ISSN 1932-6157. doi: 10.1214/16-aoas931. URL `http://dx.doi.org/10.1214/16-AOAS931`.

D. Dong, P. Fang, Y. Bock, F. Webb, L. Prawirodirdjo, S. Kedar, and P. Jamason. Spatiotemporal filtering using principal component analysis and karhunen-loeve expansion approaches for regional gps network analysis. *Journal of Geophysical Research:*

Solid Earth, 111(B3), 2006. doi: https://doi.org/10.1029/2005JB003806. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JB003806`.

Matthew M. Dunlop, Mark A. Girolami, Andrew M. Stuart, and Aretha L. Teckentrup. How deep are deep gaussian processes? *J. Mach. Learn. Res.*, 19(1):2100–2145, jan 2018. ISSN 1532-4435.

Matthew M. Dunlop, Dejan Slepčev, Andrew M. Stuart, and Matthew Thorpe. Large data and zero noise limits of graph-based semi-supervised learning algorithms. *Applied and Computational Harmonic Analysis*, 49(2):655–697, 2020. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2019.03.005. URL `https://www.sciencedirect.com/science/article/pii/S1063520318301398`.

Mark Fiecas and Hernando Ombao. Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association*, 111(516):1440–1453, 2016. doi: 10.1080/01621459.2016.1165683. URL `http://dx.doi.org/10.1080/01621459.2016.1165683`.

Thaís C. O. Fonseca and Mark F. J. Steel. A general class of nonseparable space-time covariance models. *Environmetrics*, 22(2):224–242, mar 2011. doi: 10.1002/env.1047. URL `https://doi.org/10.1002%2Fenv.1047`.

Lara Fontanella and Luigi Ippoliti. Dynamic models for space-time prediction via karhunen-loéve expansion. *Statistical Methods & Applications*, 12(1):61–78, feb 2003. doi: 10.1007/bf02511584.

Emily B Fox and David B Dunson. Bayesian nonparametric covariance regression. *Journal of Machine Learning Research*, 16:2501–2542, 2015.

Montserrat Fuentes, Li Chen, and Jerry M. Davis. A class of nonseparable and nonstationary spatial temporal covariance functions. *Environmetrics*, 19(5):487–507, 2008. doi: https://doi.org/10.1002/env.891. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/env.891`.

Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Boston, second edition edition, 1990. ISBN 978-0-08-047865-4. doi: https://doi.org/10.1016/B978-0-08-047865-4.50007-7. URL `https://www.sciencedirect.com/science/article/pii/B9780080478654500077`.

Alan E Gelfand, Athanasios Kottas, and Steven N MacEachern. Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, sep 2005. doi: 10.1198/016214504000002078.

S. Ghosal and A.W. van der Vaart. Convergence rates of posterior distributions for non-i.i.d. observations. *Annals of Statistics*, 35(1):192–223, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001172. MR2332274.

Subhashis Ghosal and Aad van der Vaart. Fundamentals of nonparametric bayesian inference. 2017. doi: 10.1017/9781139029834. URL `http://dx.doi.org/10.1017/9781139029834`.

Tilmann Gneiting. Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600, 2002. doi: 10.1198/016214502760047113. URL `https://doi.org/10.1198/016214502760047113`.

Martin Hairer. An introduction to stochastic pdes. *arXiv:0907.4178*, 07 2009.

Jouni Hartikainen, Jaakko Riihimäki, and Simo Särkkä. Sparse spatio-temporal gaussian processes with general likelihoods. In Timo Honkela, Włodzisław Duch, Mark Girolami, and Samuel Kaski, editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 193–200, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-21735-7.

Chenhui Hu, Lin Cheng, Jorge Sepulcre, Keith A. Johnson, Georges E. Fakhri, Yue M. Lu, and Quanzheng Li. A spectral graph regression model for learning brain connectivity of alzheimer's disease. *PLOS ONE*, 10(5):e0128136, May 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0128136. URL `http://dx.doi.org/10.1371/journal.pone.0128136`.

W. Huang, T. A. W. Bolton, J. D. Medaglia, D. S. Bassett, A. Ribeiro, and D. Van De Ville. A graph signal processing perspective on functional brain imaging. *Proceedings of the IEEE*, 106(5):868–885, May 2018. ISSN 0018-9219. doi: 10.1109/JPROC.2018.2798928.

Jung Won Hyun, Yimei Li, Chao Huang, Martin Styner, Weili Lin, and Hongtu Zhu. Stgp: Spatio-temporal gaussian process models for longitudinal neuroimaging data. *NeuroImage*, 134:550–562, Jul 2016. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2016.04.023. URL `http://dx.doi.org/10.1016/j.neuroimage.2016.04.023`.

Danil Kuzin, Olga Isupova, and Lyudmila S. Mihaylova. Spatio-temporal structured sparse regression with hierarchical gaussian process priors. *IEEE Transactions on Signal Processing*, 66:4598–4611, 2018.

Shiwei Lan, Andrew Holbrook, Gabriel A. Elias, Norbert J. Fortin, Hernando Ombao, and Babak Shahbaba. Flexible Bayesian Dynamic Modeling of Correlation and Covariance Matrices. *Bayesian Analysis*, 15(4):1199 – 1228, 2020. doi: 10.1214/19-BA1173. URL `https://doi.org/10.1214/19-BA1173`.

L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, Jan 1973. ISSN 0090-5364. doi: 10.1214/aos/1193342380. URL `http://dx.doi.org/10.1214/aos/1193342380`.

L. LeCam. On local and global properties in the theory of asymptotic normality of experiments. *Stochastic Processes and Related Topics*, 1:13–54, 1975.

Jaakko Luttinen and Alexander Ilin. Efficient gaussian process inference for short-scale spatio-temporal modeling. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 741–750, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL `http://proceedings.mlr.press/v22/luttinen12.html`.

Lorenzi Marco, Gabriel Ziegler, Daniel C. Alexander, and Sebastien Ourselin. Modelling non-stationary and non-separable spatio-temporal changes in neurodegeneration via gaussian process convolution. In Kanwal Bhatia and Herve Lombaert, editors, *Machine Learning Meets Medical Imaging*, pages 35–44, Cham, 2015. Springer International Publishing. ISBN 978-3-319-27929-9.

Jeff Harrison Mike West. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, 2nd edition, 1997. doi: 10.1007/b98971. URL `https://doi.org/10.1007%2Fb98971`.

Iain Murray, Ryan Prescott Adams, and David J.C. MacKay. Elliptical slice sampling. *JMLR: W&CP*, 9:541–548, 2010.

Radford M. Neal. Slice sampling. *Annals of Statistics*, 31(3):705–767, 2003.

B. Ng, V. Siless, G. Varoquaux, J. Poline, B. Thirion, and R. Abugharbieh. Connectivity-informed sparse classifiers for fmri brain decoding. In *2012 Second International Workshop on Pattern Recognition in NeuroImaging*, pages 101–104, July 2012. doi: 10.1109/PRNI.2012.11.

Mu Niu, Zhenwen Dai, Neil Lawrence, and Kolja Becker. Spatio-temporal gaussian processes modeling of dynamical systems in systems biology. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 37. JMLR: W&CP, 10 2015. URL `https://arxiv.org/pdf/1610.05163`.

Christopher Paciorek and Mark Schervish. Nonstationary covariance functions for gaussian process regression. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL `https://proceedings.neurips.cc/paper/2003/file/326a8c055c0d04f5b06544665d8bb3ea-Paper.pdf`.

Christopher J. Paciorek and Mark J. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506, 2006. doi: 10.1002/env.785. URL `https://doi.org/10.1002%2Fenv.785`.

World Alzheimer Report. The state of the art of dementia research: New frontiers. https://www.alz.co.uk/research/world-report-2018, 2018.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/8208974663db80265e9bfe7b222dcb18-Paper.pdf`.

S. Sarkka, A. Solin, and J. Hartikainen. Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, July 2013. ISSN 1053-5888. doi: 10.1109/MSP.2013.2246292.

Simo Sarkka and Jouni Hartikainen. Infinite-dimensional kalman filtering approach to spatio-temporal gaussian process regression. In Neil D. Lawrence and Mark Girolami,

editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 993–1001, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL `http://proceedings.mlr.press/v22/sarkka12.html`.

Ransalu Senanayake, Simon Timothy O'Callaghan, and Fabio Tozeto Ramos. Predicting spatio-temporal propagation of seasonal influenza using variational gaussian process regression. In *AAAI*, 2016.

X. Shen, X. Papademetris, and R.T. Constable. Graph-theory based parcellation of functional subunits in the brain from resting-state fmri data. *NeuroImage*, 50(3): 1027–1035, Apr 2010. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2009.12.119. URL `http://dx.doi.org/10.1016/j.neuroimage.2009.12.119`.

A. Singh, F. Ramos, H. D. Whyte, and W. J. Kaiser. Modeling and decision making in spatio-temporal processes for environmental surveillance. In *2010 IEEE International Conference on Robotics and Automation*, pages 5490–5497, May 2010. doi: 10.1109/ROBOT.2010.5509934.

Alexander J. Smola and Risi Kondor. Kernels and regularization on graphs. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 144–158, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45167-9.

Marco Todescato, Andrea Carron, Ruggero Carli, Gianluigi Pillonetto, and Luca Schenato. Efficient spatio-temporal gaussian regression via kalman filtering. *Automatica*, 118: 109032, 2020. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2020.109032. URL `https://www.sciencedirect.com/science/article/pii/S0005109820302302`.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008. ISSN 00905364. URL `http://www.jstor.org/stable/25464673`.

A. W. van der Vaart and J. H. van Zanten. Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 10 2009. doi: 10.1214/08-AOS678. URL `https://doi.org/10.1214/08-AOS678`.

Aad van der Vaart and Harry van Zanten. Information rates of nonparametric gaussian process methods. *J. Mach. Learn. Res.*, 12:2095–2119, July 2011. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1953048.2021067`.

Kangrui Wang, Oliver Hamelijnck, Theodoros Damoulas, and Mark Steel. Non-separable non-stationary random fields. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9887–9897. PMLR, 13–18 Jul 2020. URL `http://proceedings.mlr.press/v119/wang20g.html`.

Christopher K Wikle. A kernel-based spectral model for non-gaussian spatio-temporal processes. *Statistical Modelling*, 2(4):299–314, 2002. doi: 10.1191/1471082x02st036oa. URL `https://doi.org/10.1191/1471082x02st036oa`.

CK Wikle and N Cressie. A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, 86(4):815–829, 12 1999. ISSN 0006-3444. doi: 10.1093/biomet/86.4.815. URL https://doi.org/10.1093/biomet/86.4.815.

Andrew Wilson and Zoubin Ghahramani. Generalised wishart processes. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 736–744, Corvallis, Oregon, 2011. AUAI Press.

Bohai Zhang and Noel Cressie. Bayesian inference of spatio-temporal changes of arctic sea ice. *Bayesian Analysis*, 15(2):605–631, jun 2020. doi: 10.1214/20-ba1209.

Zheng Zhao, Muhammad Emzir, and Simo Särkkä. Deep state-space gaussian processes. *Statistics and Computing*, 31(6), sep 2021. doi: 10.1007/s11222-021-10050-6. URL https://doi.org/10.1007%2Fs11222-021-10050-6.