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

Shiwei Lan∗

Department of Statistics, University of Illinois Urbana-Champaign

Abstract

Spatiotemporal processes are ubiquitous in our life and have been a trending topic in the scientific community, e.g., the dynamic brain connectivity study in neuroscience. There is usually complicated dependence among spatial locations and such relationship does not necessarily stay static over time. Spatiotemporal Gaussian process (STGP) is a popular nonparametric method to model this type of data. However, the classic STGP has a covariance kernel with space and time separated, failed to characterize the temporal evolution of spatial dependence (TESD). Even for some recent work on non-separable STGP, location and time are treated with no difference, which is unnecessarily inefficient. This paper generalizes STGP by introducing the time-dependence to the spatial kernel and varying its eigenvalues over time. A novel STGP model with the covariance kernel having a Kronecker sum structure is proposed and proved to be superior to the popular kernel with a Kronecker product structure. A simulation study on the spatiotemporal process and a longitudinal neuroimaging analysis of Alzheimer’s patients demonstrate the advantage of the proposed methodology in effectively and efficiently characterizing TESD.

Keywords—Spatiotemporal Gaussian process (STGP), Temporal Evolution of Spatial Dependence (TESD), Non-separable Kernel, Kronecker Sum Structure, Nonparametric Bayesian Spatiotemporal Model

1 Introduction

Spatiotemporal data are ubiquitous nowadays 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. They can be viewed as either multiple time series observed across various locations, or geographic data recorded at different time points.

∗shiwei@illinois.edu, http://shiwei.stat.illinois.edu
There is usually intricate relationship between space and time in this type of data. The dependence among spatial locations does not necessarily stay static over time. For example, in the study of dynamic brain connectivity (association among brain regions) during certain cognitive processes (Cribben et al., 2012; Fiecas and Ombao, 2016; Lan et al., 2019), multi-site brain signals are correlated to each other and such spatial dependence varies along the time. 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. In general, the temporal evolution of spatial dependence (TESD) is an important subject to understand the mechanism of some natural phenomena, e.g. disease development, to predict their progress and to extrapolate to unknown territory. In this paper, a fully Bayesian nonparametric model based on spatiotemporal Gaussian process (STGP) is proposed to characterize the spatial correlation and its evolution in time. In particular, the proposed method, generalized STGP, is used to analyze brain images of Alzheimer’s patients to uncover TESD of their brain regions.

STGP is a special type of Gaussian process that can model both spatial and temporal information simultaneously. There is a rich literature on STGP including, but not limited to, Singh et al. (2010); Hartikainen et al. (2011); Luttinen and Illin (2012); Soh et al. (2012); Sarkka and Hartikainen (2012); Sarkka et al. (2013); Lindström et al. (2013); Liu (2013); Marco et al. (2015); Niu et al. (2015); Datta et al. (2016); Hyun et al. (2016); Senanayake et al. (2016); Todescato et al. (2017); Nabarro et al. (2018); Kuzin et al. (2018). However, most of them are based on the separability between space and time, and not all of them can handle non-stationary processes. As detailed below, under such separability assumption, the spatial correlation of the spatiotemporal process conditioned on anytime is independent of time. Therefore, these models doom to fail in describing the temporal evolution of spatial dependence (TESD). To effectively characterize the space-time interaction, several models (Singh et al., 2010; Marco et al., 2015; Datta et al., 2016; Hyun et al., 2016; Kuzin et al., 2018) are built on non-separability conditions. However, they treat spatial and temporal variables without difference in the joint (full-sized space-time) covariance kernel, which may not be efficient in learning TESD (See more details in Section 2.5). More specifically, Marco et al. (2015) propose STGP based on kernel convolutions of a white noise GP with a full-sized kernel, which is similar to our model I with a Kronecker product structure (See more details in Section 2). Datta et al. (2016) sparsify such full-sized space-time kernel using nearest neighbors. Hyun et al. (2016) consider a functional PCA model with again the full-sized kernel for the random component. Their kernel admits a spectral decomposition however with eigenfunctions in the extended coordinates of both space and time, which is computationally inefficient. Additionally, almost all the existing work on STGP focuses on modeling and predicting the mean functions, but none of them models or predicts the covariance kernel in order to fully describe TESD. In general, there is a lack of flexible and efficient Bayesian non-parametric models that could effectively characterize TESD. This paper aims to fill the blank in the literature.

Starting from the classic separable STGP, we generalize it by introducing a time-
dependent spatial kernel. To learn TESD, we compare two model structures based on the
Kronecker product (the most popular one) and the Kronecker sum and discover that the
latter is more efficient and effective. The second model does not use a full-sized space-time
kernel thus avoids modeling unnecessary blocks off the main diagonal in learning TESD
(See more details in Section 2.5). To introduce the time-dependence to the spatial kernel,
we model the eigenvalues in the Mercer’s representation of the spatial kernel dynamically
using another independent GP. The resulting time-dependent spatial kernel, rigorously de-
defined under given conditions, is essential to reveal TESD of spatiotemporal processes. Due
to the construction, it can be used to capture the non-stationarity of the processes. We
find that Senanayake et al. (2016); Nabarro et al. (2018) have a similar additive structure
of the time, space, and space-time covariances, however missing the time-dependence in
the description of spatial correlation. Perhaps the most relevant work is the “coregion-
alization” model (Banerjee 2015) for measurements that covary jointly over a region. It
models a time series of spatial processes using a factor model with the resulting covariance
expressed as a weighted sum of “coregionalization matrices”, similar to a finite truncation
of our series representation. However, coregionalization matrices have finite dimensions
and our model is more general than such a semi-parametric approach (See more details in
Section 2.3). Kuzin et al. (2018) is also a related work that gives a time-evolving representa-
tion of the interdependencies between the signal components and they use a spike-slab
prior to induce sparsity. Our second model also generalizes the semi-parametric scheme of
dynamic covariance modeling (Wilson and Ghahramani 2011; Fox and Dunson, 2015; Lan
et al., 2019) with a covariance matrix of fixed size for the spatial domain to have a kernel
(infinite-dimensional spatial covariance) in order to enable extrapolation to new locations
while modeling the temporal evolution (See more details in Section 3.2.2).

The contributions of this work include: (1) The separable STGP is generalized effi-
ciently by introducing the time-dependence to the spatial kernel via Mercer’s representation
and the theoretic properties of the new kernel is systematically investigated for the first
time; (2) Two model structures based on the Kronecker product and the Kronecker sum
are compared and the latter is shown to be more effective in learning TESD both theoret-
ically and numerically; (3) TESD is modeled and predicted effectively and efficiently. The
impact of this work is not limited to the included application of longitudinal neuroimag-
ing analysis. The proposed methodology, generalized STGP, is generic to learn TESD for
all spatiotemporal processes, with potential applications to genome-wide association study
(GWAS) in evolution, climate change study, investment portfolio maintenance etc..

The rest of the paper is organized as follows. Section 2 reviews the separable spa-
tiotemporal Gaussian Process (STGP) and generalizes it to be non-separable by intro-
ducing the time-dependence to the spatial kernel. Two model structures, the Kronecker
product (model I) and the Kronecker sum (model II), in the generalized STGP are com-
pared for learning the temporal evolution of spatial dependence (TESD). Section 3 dis-
cusses the posterior inference and the prediction of both mean function and covariance
kernel. Model II with the Kronecker sum structure is shown to have computational ad-
vantage over model I based on the Kronecker product. Section 4 contains a simulation study to illustrate the effectiveness of the proposed generalized STGP (mainly model II) in modeling and predicting TESD of spatiotemporal processes. In Section 5, the proposed methodology is applied to analyze a series of positron emission tomography (PET) brain images of Alzheimer’s patients obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI). TESD of their brain regions is effectively characterized and predicted during the progression of AD. Finally, we conclude in Section 6 with a few comments on the methodology and some discussions of future directions. All the numerical codes are available at https://github.com/lanzithinking/TESD_gSTGP.

2 Generalized 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 the Kronecker product and the Kronecker sum are compared and the latter is more effective and efficient.

We fix some notations first. 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 $z := (x, t)$ as the joint variable. The spatiotemporal data $\{y_{ij} | i = 1, \cdots, I; j = 1, \cdots, J\}$ are taken on a grid of points $\{z_{ij} = (x_i, t_j) | x_i \in \mathcal{X}, t_j \in \mathcal{T}\}$ with the spatial discrete size $I$ and the temporal discrete size $J$. A (centered) STGP is uniquely determined by its covariance kernel $C_z : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, a bilinear symmetric positive-definite function. $C_z$ could be defined without differentiating $x$ and $t$ (full-size covariance) or by exploring structures in space and time, to be detailed below.

2.1 Spatiotemporal Gaussian Process

The spatiotemporal data $\{y_{ij}\}$ are usually modeled using the standard (separable) STGP model:

$$
\begin{align*}
y_{ij} &= f(x_i, t_j) + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \\
f(z) &\sim \mathcal{GP}(0, C_z)
\end{align*}
$$

where the joint spatiotemporal kernel $C_z$ often has the following separability condition assumed

$$
C_z = C_x \otimes C_t, \quad C_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad C_t : \mathcal{T} \times \mathcal{T} \to \mathbb{R}
$$

With these notations, we can define the (separable) STGP through the matrix normal distribution as follows.

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

\[
F = f(X, t) = [f(x_i, t_j)]_{I \times J} \sim \mathcal{MN}_{I \times J}(M, C_x, C_t)
\]

where \( M = m(X, t) = [m(x_i, t_j)]_{I \times J}, \) \( C_x = C_x(X, X) = [C(x_i, x_j)]_{I \times I}, \) and \( C_t = C_t(t, t) = [C(t_j, t_j)]_{J \times J}. \)

**Remark 1.** If we vectorize the matrix \( F_{I \times J}, \) then we have

\[
\text{vec}(F) \sim \mathcal{N}(\text{vec}(M), C_t \otimes C_x), \quad \text{vec}(F^T) \sim \mathcal{N}(\text{vec}(M^T), C_x \otimes C_t)
\]

Therefore, we often denote a sample of the (centered) STGP as \( f \sim \mathcal{GP}(0, C_x \otimes C_t). \)

Note that conditioned on any fixed time \( t \in T, \) the covariance of the process \( f(x, t) \) in the space domain is reduced to

\[
\text{Cov}[f(x, t), f(x', t)] \propto C_x(x, x'), \quad \forall t \in T
\]

which is static in time, if a stationary \( C_t \) is adopted. Such drawback of the separable kernel makes the corresponding STGP fail to characterize TESD in the field. Neither can it capture the spatial variation of the temporal correlation (SVTC) among the processes because conditioned on any fixed location \( x \in X, \) the covariance in the time domain becomes constant in space with a stationary \( C_x \)

\[
\text{Cov}[f(x, t), f(x', t')] \propto C_t(t, t'), \quad \forall x \in X
\]

We now formalize the definition of the temporal evolution of spatial dependence (TESD).

**Definition 2** (TESD). The temporal evolution of spatial dependence (TESD) of a spatiotemporal process \( y(x, t) \) is the spatial covariance conditioned on a common time

\[
C_{y|t} := \text{Cov}[y(x, t), y(x', t)], \quad \forall t \in T
\]

This paper focuses on the time progression of the spatial kernel (TESD), which motivates the following generalization of STGP to include the time-dependence.

### 2.2 Generalized Spatiotemporal Gaussian Process

In this subsection, we generalize the standard separable STGP to introduce the time-dependence to the spatial kernel \( C_x \) while keeping the desirable structure in the joint kernel \( C_z. \) We present two generalizations of the spatiotemporal kernel \( C_z \) based on two different ideas. One is to replace the spatial kernel \( C_x \) in (2) with a time-dependent analogy \( C_x|t, \) resulting in a popular Kronecker product structure; the other stems from vectorizing the
standard (scalar) GP regression model \cite{Lan2019} which yields a Kronecker sum structure.

First, the observations in the model (1) can be viewed as taken from the following process

\[ y(z) = f(z) + \varepsilon, \quad \varepsilon \sim \mathcal{GP}(0, \sigma^2 \mathcal{I}_z) \]  

(7)

The mean function is given a STGP prior \( f(z) \sim \mathcal{GP}(0, C_z) \) with

\[ C_z = C_{x|t} \otimes C_t \]  

(8)

Then the marginal covariance for \( y \) becomes

\[ C^1_y = C_{x|t} \otimes C_t + \sigma^2 \mathcal{I}_x \otimes \mathcal{I}_t \]  

(9)

where \( \mathcal{I}_x(x, x') = \delta(x = x') \), and \( \mathcal{I}_t(t, t') = \delta(t = t') \) with \( \delta(\cdot) \) being the Dirac function. The most informative part \( C_{x|t} \otimes C_t \) bares a Kronecker product structure.

On the other hand, \cite{Lan2019} vectorize the standard GP regression model

\[ y(t_j) = f(t_j) + \varepsilon_j, \quad \varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \Sigma_{\varepsilon}) \]  

\[ f_i(t) \overset{iid}{\sim} \mathcal{GP}(0, \mathcal{C}(t, t')), \quad i = 1, \ldots, D. \]  

(10)

and introduce the time-dependence in \( \Sigma_{\varepsilon} \) by replacing it with \( \Sigma_t \). In this case, the spatial dependence is encoded in the parametric covariance matrix \( \Sigma_t \) which itself evolves along time thus the noise becomes independent but not identical (i.n.i.d); while the temporal evolution is modeled by various non-parametric GPs. The marginal covariance is

\[ \text{Cov}[y_i(t), y_{i'}(t')] = \delta_{ii'} C(t, t') + \Sigma_{ii'}(t) \delta_{t=t'} \]  

(11)

A fully non-parametric generalization could be done by replacing \( \Sigma_t \) with a time-dependent spatial kernel \( C_{x|t} \)

\[ y(t) | m_{x|t}, C_{x|t} \sim \mathcal{GP}_x(m_{x|t}, C_{x|t}, \mathcal{I}_t) \]  

\[ m_{x|t} \sim \mathcal{GP}_t(0, \mathcal{I}_x, C_t) \]  

(12)

which has the following marginal covariance for \( y \) in a form of Kronecker sum:

\[ C^\text{II}_y = \mathcal{I}_x \otimes C_t + C_{x|t} \otimes \mathcal{I}_t =: C_{x|t} \oplus C_t \]  

(13)

The generalized STGP models can be summarized in the following unified form

\[ y(z) | m, C_{y|m} \sim \mathcal{GP}(m, C_{y|m}) \]  

\[ m(z) \sim \mathcal{GP}(0, C_m) \]  

(14)

model I : \( C_{y|m} = \sigma^2 \mathcal{I}_x \otimes \mathcal{I}_t, \quad C_m = C_{x|t} \otimes C_t \)

model II : \( C_{y|m} = C_{x|t} \otimes \mathcal{I}_t, \quad C_m = \mathcal{I}_x \otimes C_t \)
Figure 1: Joint kernels $\mathcal{C}_y$ specified by model 0 with a separable structure $\mathcal{C}_x \otimes \mathcal{C}_t$ (left), model I with a Kronecker product structure $\mathcal{C}_x|_t \otimes \mathcal{C}_t$ (middle) and model II with a Kronecker sum structure $\mathcal{C}_x|_t \oplus \mathcal{C}_t$ (right). Blocks in red frame illustrate the temporal evolution of spatial kernel.

Note the above two models specify different structures for the marginal covariance $\mathcal{C}_y = \mathcal{C}_y|_m + \mathcal{C}_m$. As both models can incorporate the time-dependence for the spatial kernel through $\mathcal{C}_x|_t$, $\mathcal{C}_y$ in model I is in general non-sparse but more expressive, while $\mathcal{C}_y$ in model II is sparse but less expensive compared with model I. However as detailed in Section 2.5 when modeling TESD, it will be wasteful for model I to impose complex and unnecessary structure on the blocks off the main diagonal. See Figure 1 for an illustration of these structures in $\mathcal{C}_y$ where each of the blocks in red frame indicates a spatial kernel and together they describe TESD.

2.3 Construction of Time-Dependent Spatial Kernel

In both models [14], a time-dependent spatial kernel $\mathcal{C}_x|_t$ is needed to effectively characterize TESD. In this subsection, we construct $\mathcal{C}_x|_t$ through dynamically varying eigenvalues of the spatial kernel $\mathcal{C}_x$ in Mercer’s theorem.

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

$$T_{C_x} : L^2(\mathcal{X}) \rightarrow L^2(\mathcal{X}), \quad \phi(x') \mapsto \int \mathcal{C}_x(\cdot, x') \phi(x') dx'$$ (15)

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

$$C_x(x, x') = \sum_{\ell=1}^{\infty} \lambda^2_{\ell} \phi_{\ell}(x) \phi_{\ell}(x')$$  \hfill (16)

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

To introduce the time-dependence to the spatial kernel, therefore denoted as $C_{x|t}$, we let eigenvalues $\{\lambda^2_{\ell}\}$ change with time and denote them as $\{\lambda^2_{\ell}(t)\}$. Without imposing any parametric constraints, we model $\lambda(\cdot)$ in $L^2(T)$ using another Gaussian process for each $\ell \in \mathbb{N}$ and treat it as a random draw:

$$\lambda^2_{\ell}(t) := \langle \phi_{\ell}(x), C_{x|t}(x) \phi_{\ell}(x) \rangle, \quad \lambda(\cdot) \sim \mathcal{GP}(0, \mathcal{C}_{\lambda,\ell})$$  \hfill (17)

To ensure the well-posedness of the generalization, we make the following assumption

**Assumption 1.** Denote $\lambda(t) := \{\lambda_{\ell}(t)\}_{\ell=1}^{\infty}$ for $\forall t \in T$. We assume

$$\lambda \in L^2(L^2(T)), \quad \text{i.e.} \quad \|\lambda\|_{2,2}^2 := \sum_{\ell=1}^{\infty} \|\lambda_{\ell}(\cdot)\|^2_2 < +\infty \quad \text{(18)}$$

Therefore we consider $\lambda$ in the probability space $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ with $\Omega = L^2(L^2(T))$, $\sigma$-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 (similarly as Theorem 29 in section A.2.1 of [Dashti and Stuart, 2017]). To fulfill Assumption 1, we require $\mathcal{C}_{\lambda,\ell}$ to have magnitude decay in $\ell$ with certain rate such that

$$\mathcal{C}_{\lambda,\ell} = \gamma^2_{\ell} C_u, \quad \sum_{\ell=1}^{\infty} \gamma^2_{\ell} < \infty \quad \text{(19)}$$

We choose $\gamma_{\ell} = \lambda^0_{\ell}$ being the eigenvalues of $\mathcal{C}_x$ (as a trace-class operator), or more directly $\gamma_{\ell} = \ell^{-\kappa/2}$ for some $\kappa > 1$. That is, we essentially model

$$\lambda_{\ell}(t) = \gamma_{\ell} u_{\ell}(t), \quad u_{\ell}(\cdot) \overset{iid}{\sim} \mathcal{GP}(0, \mathcal{C}_u) \text{ for } \ell \in \mathbb{N} \quad \text{(20)}$$

To see this, we compute $\mathbb{E}[\|\lambda\|_{2,2}^2] = \sum_{\ell=1}^{\infty} \gamma^2_{\ell} \mathbb{E}[\|u_{\ell}\|_2^2] = \text{tr}(\mathcal{C}_u) \sum_{\ell=1}^{\infty} \gamma^2_{\ell} < \infty$. Thus Assumption 1 holds in $\mathbb{P}$.

Under this assumption we can have the series representation of the time-dependent spatial kernel $C_{x|t}$ as in Mercer’s theorem (16):

$$C_{x|t}(x, x') = \sum_{\ell=1}^{\infty} \lambda^2_{\ell}(t) \phi_{\ell}(x) \phi_{\ell}(x')$$  \hfill (21)

With this representation, for any $t \in T$, $\{\lambda^2_{\ell}(t)\}$ can be interpreted as eigenvalues of the integral operator $T_{C_{x|t}}$ as in (15) with $C_x$ replaced by $C_{x|t}$. Assumption 1 essentially
requires that the trace of $C_x|_t$ over the spatial domain is finite in $L^1(T)$. Such construction has some similarity to the “coregionalization” model for which the spatial process $Y(x) = Aw(x)$ with the temporal covariance $T = AA^T$ and $w_j(\cdot) \sim GP(0, \rho_j)$ being independent. Denoting $T_j = a_ja_j^T$ with $a_j$ as the $j$-th column of $A$, the resulting covariance $\text{Cov}(Y(x), Y(x')) = \sum_{j=1}^p \rho_j(x - x')T_j$ is analogous to a finite truncation of (21) however $T = \sum_{j=1}^p T_j$ is finite-dimensional. Similarly as (21) we can define the joint kernel $C_z := C_x|_t \otimes \mathcal{C}_t$ in model I in a symmetric way:

$$C^I_m(z, z') := C^I_x(z, z') := \sum_{t=1}^\infty \lambda_t(t)\mathcal{C}_t(t, t')\phi_t(x)\phi_t(x')$$

(22)

And we can also define the likelihood kernel $C_{y|m} = C_x|_t \otimes \mathcal{I}_t$ in model II as follows:

$$C^I_{y|m}(z, z') = C^I_x|_t \otimes \mathcal{I}_t (z, z') = \sum_{t=1}^\infty \lambda^I_t(t)\mathcal{I}_t(t, t')\phi_t(x)\phi_t(x')$$

(23)

With these definitions, it is straightforward to verify the format of TESD in the generalized STGP model (14), mainly characterized by the time-dependent kernel $C_x|_t$.

**Proposition 2.1.** If the process $y$ is according to the generalized STGP model (14) with kernel definitions (22) (23), then we have the following conditional GPs in the space domain

$$y(\cdot, t)|C_x|_t \sim GP(0, C_{y|t})$$

$$C^I_{y|t} = C_x|_t c + \sigma^2\mathcal{I}_t, \quad C^I_{y|t} = \mathcal{I}_t c + C_x|_t$$

(24)

where $c := C_t(t, t)$.

In this paper, we assume the locations $\{x_i\}$ are fixed. When the locations change with time, e.g. moving players in a game, they become a (vector) function of time, denoted $x(t)$. Then it is more natural to define the time-dependent spatial kernel by simply substituting $x(t)$ in $C_x$ to define $C_x|_t(x, x') = C_x(x(t), x'(t))$ [Soh et al., 2012]. We will treat this case in another paper.

### 2.4 Theoretic Properties

In this subsection, we systematically study the theoretic properties of the time-dependent spatial kernel $C_x|_t$. First we prove the well-posedness of kernels (22) and (23) in the Mercer’s representation in the following theorem.

**Theorem 2.1.** [Wellposedness of Mercer’s Kernels] Under Assumption 1, both $C^I_m = C^I_x|_t \otimes \mathcal{C}_t$ and $C^I_{y|m} = C_x|_t \otimes \mathcal{I}_t$ are well defined non-negative definite kernels on $Z = X \times T$.

**Proof.** See Appendix A. \hfill \square
With the Mercer’s kernels \(^{(22)}\) and \(^{(23)}\), we could represent STGP in spatial basis with random time-varying coefficients similarly as in Karhunen-Loéve theorem.

**Theorem 2.2.** [Karhunen-Loéve Expansion] Under Assumption \(^{1}\), STGP \(f(x, t) \sim \mathcal{GP}(0, C_z)\) has the following representation of series expansion:

\[
f(x, t) = \sum_{\ell=1}^{\infty} f_\ell(t) \phi_\ell(x), \quad f_\ell(t) = \int_X f(x, t) \phi_\ell(x) dx
\]

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

- if \(C_z = C_{x|t}^2 \otimes C_t\), then \(E[f_\ell(t)f_\ell'(t')]=\lambda_\ell(t)C_t(t, t')\lambda_\ell(t')\delta_{t,t'}\).
- if \(C_z = C_{x|t} \otimes I_t\), then \(E[f_\ell(t)f_\ell'(t')]=\lambda_\ell^2(t)\delta(t=t')\delta_{t,t'}\).

*Proof.* See Appendix \([A]\) \(\square\)

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 \(0 < p \leq \infty\).

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

And we denote the space \(\ell^{k,s}(L^p(T)) := \{|\lambda|\|\lambda\|_{k, s, p} < +\infty\}\). Note, the norm in Assumption \([1]\) corresponds to the special case \(k = 2, s = 0, p = 2\). For a given spatial basis \(\{\phi_\ell(x)\}_{\ell=1}^{\infty}\), there is one-one correspondence \(f(x, t) \leftrightarrow \{f_\ell(t)\}_{\ell=1}^{\infty}\) in \((25)\). Therefore, we could also define \((k, s, p)\)-norm \((26)\) for \(f \in \ell^{k,s}(L^p(T))\). Note, when \(p = 2, f \in \ell^{k,s}(L^2(T))\) with a fixed spatial basis \(\{\phi_\ell(x)\}_{\ell=1}^{\infty}\) also implies \(f \in \ell^{k,s}(L^2(\mathbb{Z}))\) regardless of spatial basis (normalized in \(L^2(X)\)) because \(\left(\sum_{\ell=1}^{\infty} \ell^{ks}\|f_\ell(t)\phi_\ell(x)\|_2^2\right)^{\frac{1}{2}} = \left(\sum_{\ell=1}^{\infty} \ell^{ks}\|f_\ell(t)\|_2^k\|\phi_\ell(x)\|_2^k\right)^{\frac{1}{2}} = \|f\|_{k, s, 2}\).

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 \(\lambda\) decay in order \(\kappa > 1\), the following proposition states that they fall in a subset of \(\ell^2(L^2(T))\).

**Proposition 2.2.** Assume \(\gamma_\ell = O(\ell^{-\kappa/2})\) for some \(\kappa > 1\) in \((20)\). Then \(\lambda \in \ell^{2,s}(L^2(T))\) in \(\mathbb{P}\) for \(s < (\kappa - 1)/2\).

*Proof.* It is straightforward to verify that

\[
E[\|\lambda\|_{2, s, 2}^2] = \sum_{\ell=1}^{\infty} \ell^{2s}E[\|\lambda_\ell\|_2^2] = \sum_{\ell=1}^{\infty} \ell^{2s-\kappa}E[\|u_\ell\|_2^2] \lesssim tr(C_u) \sum_{\ell=1}^{\infty} \ell^{2s-\kappa} < \infty
\]

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

---

\(^{1}\)When \(k = 2\), this is related to Sobolev norm in the frequency domain and Hilbert scales.
To discuss the regularity of the random functions, we need the following assumptions on the spatial basis \( \{ \phi_\ell(x) \}_{\ell=1}^\infty \) and the dynamic eigenvalues \( \lambda \).

**Assumption 2.** We assume the spatial basis \( \{ \phi_\ell(x) \}_{\ell=1}^\infty \) are bounded in \( L^\infty(\mathcal{X}) \) and are Lipschitz with controlled growth rate in the Lipschitz constants \( \text{Lip}(\phi_\ell) \):

\[
\sup_{\ell \in \mathbb{N}} \| \phi_\ell \|_\infty + \ell^{-1} \text{Lip}(\phi_\ell) \leq C, \quad \text{for some } C > 0 \tag{27}
\]

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

\[
\lambda \in \ell^{2, s}(L^\infty(\mathcal{T})), \quad \sup_{\ell \in \mathbb{N}} \ell^{-2} \sup_{t, t' \in \mathcal{T}} \frac{Q_{\lambda, C}(t, t')}{\| \lambda \|_\infty^2 |t - t'|^2} \leq C, \quad \text{for some } C > 0 \tag{28}
\]

The following theorem by Kolmogorovs continuity test (Theorem 3.42 of \cite{Hairer}) and (Theorem 30 in section A.2.5 of \cite{Dashti and Stuart}) states that the regularity of the random functions in \( (\mathcal{X}) \) depends on the decay rate of dynamic eigenvalues.

**Theorem 2.3.** [Regularity of Random Functions] Assume \( \lambda \in \ell^{2, s}(L^2(\mathcal{T})) \). If \( f(x, t) \sim \mathcal{G}(0, C_x) \) as in Theorem 2.2, then \( f = \sum_{\ell=1}^\infty f_\ell(x)\phi_\ell(x) \in \ell^{2, s}(L^2(\mathcal{Z})) \) in probability.

Moreover, under Assumption 2-(27), there is a version \( \hat{f}(x) \) of \( f(x) := \int_\mathcal{T} f(x, t)dt \in C^{0, s'}(\mathcal{X}) \) for \( s' < s \). If further \( C_x = C_{\chi_t}^1C_{\chi_t}^1 \otimes C_t \) and \( \{ Q_{\lambda, C_t} \} \) satisfies Assumption 2-(28), then there is a version \( \hat{f}(z) \) of \( f(z) \) in \( C^{0, s'}(\mathcal{Z}) \) for \( s' < s \).

**Proof.** See Appendix A

\( \square \)

**Remark 2.** If we consider the model \( (20) \) for random \( \lambda \) and require \( \gamma_\ell = O(\ell^{-\kappa/2}) \) for some \( \kappa > 1 \), the above results still hold for \( s < (\kappa - 1)/2 \) by Proposition 2.2.

**Corollary 2.1.** If \( f(x, t) \sim \mathcal{G}(0, C_x) \) has a continuous version, then \( \{ f_\ell \}_{\ell=1}^\infty \) as in Theorem 2.2 are \( \mathcal{G} \)'s defined on \( \mathcal{T} \).

**Proof.** See Appendix A

\( \square \)

Now we consider the posterior properties. On the separable Banach space \( \mathcal{B} = \ell^2(L^2(\mathcal{T})), \| \cdot \|_{2, 2} \), we consider a Gaussian random element \( \lambda \) and denote its associated reproducing kernel Hilbert space (RKHS) as \( (\mathcal{H}, \| \cdot \|_H) \). We assume \( \gamma_\ell = O(\ell^{-\kappa/2}) \) for some \( \kappa > 1 \) and \( \text{tr}(C_u) = 1 \) by rescaling in \( (20) \). Then RKHS is \( \mathcal{H} = \ell^{2, \kappa/2}(L^2(\mathcal{T})) \) with the following inner product and norm

\[
\langle h, h' \rangle_\mathcal{H} = \sum_{\ell=1}^\infty \langle \gamma_\ell^{-1}h_\ell, \gamma_\ell^{-1}h'_\ell \rangle, \forall h, h' \in \mathcal{H}, \quad \| \cdot \|_\mathcal{H} = \langle \cdot, \cdot \rangle_\mathcal{H}^{1/2} \tag{29}
\]

\( ^2 \)A version/modification of stochastic process \( \tilde{f}(x) \) of \( f(x) \) means \( \mathbb{P}[\tilde{f}(x) = f(x)] = 1 \) for \( \forall x \in \mathcal{X} \).
Define the contraction rate of $\lambda$ at $\lambda_0$ as follows

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

(30)

Let $p$ be a centered (assume $m \equiv 0$ for simplicity) Gaussian model, which is uniquely determined by its covariance $C_{X|t} = \sum_{j=1}^{\infty} \lambda_j^2(t) \phi_j \otimes \phi_j$. For a fixed spatial basis $\{\phi_j\}$, the model density is parametrized by $\lambda$, hence denoted as $p_\lambda$. Denote $P_{\lambda}^{(n)} := \otimes_{j=1}^{n} P_{\lambda,j}$ as the product measure on $\otimes_{j=1}^{n} (\mathcal{X}_j, \mathcal{B}_j, \mu_j)$. Each $P_{\lambda,j}$ has a density $p_{\lambda,j}$ with respect to the finite measure $\mu_j$. 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 $\lambda$ and its true value $\lambda_0$, we make the following assumption.

**Assumption 3.** Let $\lambda \in \ell^{1,s}(L^\infty(\mathcal{T}))$ with some $s > 0$. Assume $\lambda$ satisfy the following bounds

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

(31)

Denote the observations $Y^{(n)} = \{Y_j\}_{j=1}^{n}$ with $Y_j = y(X, t_j)$. Note they are iid in model I and independent but not identically distributed (inid) in model II. Let $n = I \wedge J$ be the minimum of the sizes of discretized spatial domain ($I$) and temporal domain ($J$). One can refer to Section 3 for the complete model details. We follow Ghosal and van der Vaart (2007); van der Vaart and van Zanten (2008) to prove the following posterior contraction about $C_{X|t}$ in model II, which generalizes Theorem 2.2 of Lan et al. (2019). 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 2.4.** [Posterior Contraction of $C_{X|t}$ in model II] Let $\lambda$ be a Borel measurable, zero-mean, tight Gaussian random element in $\mathcal{H} = \ell^2(L^2(\mathcal{T}))$ satisfying Assumption 3 and $P_{\lambda}^{(n)} = \otimes_{j=1}^{n} P_{\lambda,j}$ be the product measure of $Y^{(n)}$ parametrized by $\lambda$. If the true value $\lambda_0 \in \Theta$ is in the support of $\lambda$, and $\varepsilon_n$ 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 $\Pi_n(\lambda \in \Theta_n) : d_{n,H}(\lambda, \lambda_{n,0}) > M_n \varepsilon_n |Y^{(n)}| \rightarrow 0$ in $P_{\lambda_{n,0}}^{(n)}$ probability for every $M_n \rightarrow \infty$.

Proof. See Appendix A.

**Remark 3.** The observations $Y^{(n)}$ in model I are iid however the likelihood model is determined by $C_{y|m} = \sigma^2 \mathcal{I}_X \otimes \mathcal{I}_t$ which does not contain $\lambda$, thus the posterior of $\lambda$ in model I cannot contract to the correct distribution. Therefore, model I cannot be used to learn TESD, mainly conveyed in $C_{X|t}$. See more details in the next subsection.

Although the above theorem dictates that the posterior of $\lambda$ contracts to the true value $\lambda_0$ at certain rate $\varepsilon_n$, it does not provide the details of $\varepsilon_n$. The following theorem specifies the contraction rate, which depends on the regularity of both the truth and the prior used. The proof mainly follows Section 11.4 of Ghosal and van der Vaart (2017).
Theorem 2.5. [Posterior Contraction Rate of $C_{X|T}$ in model II] Let $\lambda$ be a Gaussian random element defined in (20) with $\gamma_t = \Theta(\ell^{-\kappa/2})$ for some $\kappa > 1$ and $\text{tr}(C_u) = 1$. The rest settings are the same as in Theorem 2.4. If the true value $\lambda_0 \in \ell^2, s(L^2(T))$, then we have the rate of posterior contraction $\varepsilon_n = \Theta(n^{-(\kappa+1)/2})$.

Proof. See Appendix A.

Remark 4. Posterior contraction rate is the minimal solution to the rate equation $\varphi_{\lambda_0}(\varepsilon_n) \leq n^{-\kappa/2}$. 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.

2.5 Comparison of Two Kernels

Before concluding this section, we compare these two spatiotemporal kernels with different structures as follows:

$$C_I = C_{X|T} \otimes C_T + \sigma^2 \otimes I_T$$ vs $$C_{II} = I_X \otimes C_T + C_{X|T} \otimes I_T$$

Note that, despite of the time-dependence in the spatial kernel, model I with the Kronecker product structure in the covariance kernel is the most popular model for spatiotemporal data (Hartikainen et al., 2011; Luttinen and Ilin, 2012; Sarkka and Hartikainen, 2012; Sarkka et al., 2013; Lindström et al., 2013; Liu, 2013; Niu et al., 2015; Senanayake et al., 2016; Todescato et al., 2017).

From the modeling perspective, it is not completely natural for model I to assume iid noise. Apart from the mean function sufficiently modeled with the Kronecker product kernel a priori, the noise processes could possibly be inter-correlated over time among the components but this is not correctly reflected in model I. It does not affect modeling the mean function with an expressive prior kernel, however model I suffers from modeling the covariance function due to the lack of structure in the likelihood kernel $\sigma^2 \otimes I_T$. 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. This is verified in the numerical simulation in Section 4.2. On the other hand, model II puts a balanced structure between the prior kernel and the likelihood kernel to achieve a good trade-off. What is more, the posterior concentrates on the likelihood with the kernel structure $C_{X|T} \otimes I_T$ that could correctly capture TESD (See Figure 1 and more numerical details in Section 4.2).

Considering the computational complexity of the two models, model II has significantly lower complexity than model I. As illustrated in Figure 1, if we are only interested in TESD as illustrated by the blocks in red frame – each block being a spatial covariance

13
matrix evolving with time, it is wasteful to model the blocks off the main diagonal (not in red frame) with complicated structures in model I. On the contrary, model II puts enough details in the main diagonal blocks (in red frame) for describing TESD and yet the resulting joint kernel is highly sparse. When discretizing the space-time domain, both $C_y^I$ and $C_y^H$ become matrices of size $IJ \times IJ$ with $C_{x|t}^I$, $C_{x|t'}^I$ and $C_{x|t}^H$ as the most computation-intensive part respectively. The latter has a spectral decomposition inherited from that of the spatial kernel thus many calculations can be simplified with the Sherman-Morrison-Woodbury formula and the matrix determinant lemma; however it does not apply to the former. See more details in Appendix B.3. In the following we discuss the posterior inference by MCMC and various predictions for the mean and covariance functions.

3 Posterior and Predictions

Suppose we are given the spatiotemporal data $D := \{Z, Y\}$, and there are $K$ independent trials in the data $Y := \{Y_k\}_{k=1}^K$, with each $Y_k = y(Z) = \{y(x_i, t_j)\}_{i \times j}$. Denote $M_{I \times J} := m(Z)$ and $C_{y|M} := C_{y|m}(Z, Z)$. Consider the model (14) with hyper-parameters which are in turn given priors respectively, summarized as follows

\[ Y_k|M, C_{Y|M} \sim \mathcal{N}(M, C_{Y|M}), \quad \text{i.i.d. for } k = 1, \ldots, K \]

\[ m(z) \sim \mathcal{GP}(0, C_m) \]

\[ C_{x|t}(z, z') = \sum_{\ell=1}^{\infty} \lambda_\ell(t) \lambda_\ell(t') \phi_\ell(x) \phi_\ell(x'), \quad \lambda_\ell(\cdot) \sim \mathcal{GP}(0, \gamma_\ell^2 C_u) \]  \hspace{1cm} (32)

\[ C_s = \sigma_s^2 \exp(-0.5 \| s - s' \|^2 / \rho_s^2) \]

\[ \sigma_s^2 \sim \Gamma^{-1}(a_s, b_s), \quad \log \rho_s \sim \mathcal{N}(m_s, \nu_s), \quad s = x, t, u \]

where the likelihood kernel $C_{y|m}$ and the prior kernel $C_m$ are specified in (14). Let $\{u_\ell(\cdot)\}$ be i.i.d draws from $\mathcal{GP}(0, C_u)$ a priori and we set $\lambda_\ell(t) = \gamma_\ell u_\ell(t)$. Denote $C_m := C_m(Z, Z)$, $C_u := C_u(t, t)$, $\Lambda := \{\lambda_\ell(t)\}_{\ell=1}^L$, $U := \{u_\ell(t)\}_{\ell=1}^L$ and $\gamma := \{\gamma_\ell\}_{\ell=1}^L$. Then the total probability, up to a constant, is as follows

\[ \log p(M, \Lambda, C^2, \eta|Y) \]

\[ = \log p(Y|M, C_{Y|M}) + \log p(M|C_M) + \log p(\Lambda) + \sum_{s=x,t,u} \log p(\sigma_s^2) + \sum_{s=x,t,u} \log p(\eta_s) \]

\[ = - \frac{K}{2} \log |C_{Y|M}| - \frac{1}{2} \sum_{k=1}^K \text{vec}(Y_k - M)^T C_{Y|M}^{-1} \text{vec}(Y_k - M) \]

\[ - \frac{1}{2} \log |C_M| - \frac{1}{2} \text{vec}(M)^T C_M^{-1} \text{vec}(M) - J T^T \log |\gamma| - \frac{L}{2} \log |C_u| - \frac{1}{2} \text{tr}(U^T C_u^{-1} U) \]

\[ - \sum_{s=x,t,u} (a_s + 1) \log \sigma_s^2 + b_s \sigma_s^{-2} - \sum_{s=x,t,u} \frac{1}{2} (\eta_s - m_s)^2 / \nu_s \]  \hspace{1cm} (33)
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 \((x_*, t_*)\), \((x, t_*)\) or \((x_*, t)\):

\[
m(x_*, t_*)|D, \quad m(x, t_*)|D, \quad m(x_*, t)|D, \quad C_{x|x'}(x, x')|D, \quad C_{x|x}(x, x_*)|D
\]  

(34)

### 3.1 Prediction of Mean

We only consider the prediction \(m(x_*, t_*)|D\) because the other two predictions \(m(x_*, t)|D, \ m(x, t_*)|D\) are sub-problems of it. The following proposition gives the predictive distribution of the mean function \(m(x, t)\).

**Proposition 3.1.** Fit the spatiotemporal data \(D = \{Z, Y\}\) with the model (32). Then given a new point \(z_* = (x_*, t_*)\) we have

\[
m(z_*)|D \sim N(m', C')
\]

where we denote

\[
m' = c_\star^T(c_M + K^{-1}C_{Y|M})^{-1}Y, \quad C' = c_{m*} - c_\star^T(c_M + K^{-1}C_{Y|M})^{-1}c_\star
\]

\[
Y := \frac{1}{K} \sum_{k=1}^K \text{vec}(Y_k), \quad c_{m_*} := c_m(z_*, z_*), \quad c_* := c_m(Z, z_*), \quad c_{\star}^T := c_m(z_*, Z)
\]

**Proof.** See Appendix A.

### 3.2 Prediction of Covariances

Now we consider the prediction of covariances. There are two types of prediction of particular interest. The first one, \(C_{x|x'}(x, x')|D\), evolves the spatial dependence among existing locations to other (future) times \(t_*\). The second one, \(C_{x|x}(x, x_*)|D\), extends the temporal evolution of spatial dependence to new (neighboring) locations \(x_*\). The latter is also exclusive to the proposed fully nonparametric model. The semi-parametric methods of dynamic covariance modeling (Wilson and Ghahramani 2011; Fox and Dunson 2015; Lan et al. 2019) with a covariance matrix (instead of a kernel) for the spatial dependence do not have this feature. 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.
3.2.1 Evolve Spatial Dependence to Future Time

Note from the definition [21], we know that \( C_{x|t_0} \) is a function of dynamic eigenvalues \( \{\lambda_\ell(t)\} \) with fixed spatial basis \( \{\phi_\ell(x)\} \). Therefore, the prediction of the kernel \( C_{x|t} \) in the time direction can be reduced to predicting \( \lambda_\ell(t_s)|D \) as follows. Denote \( \lambda_\ell := \lambda_\ell(t) \).

\[
p(\lambda_\ell(t_s)|D) = \int p(\lambda_\ell(t_s), \lambda_\ell|D)d\lambda_\ell = \int p(\lambda_\ell(t_s)|\lambda_\ell)p(\lambda_\ell|D)d\lambda_\ell
\]

where \( p(\lambda_\ell(t_s)|\lambda_\ell) \) is the standard GP predictive distribution. We can use the standard GP predictive mean and covariance to predict and quantify the associated uncertainty for \( (\lambda_\ell(t_s)|\lambda_\ell(s))^2 \) with \( \lambda_\ell(s) \sim p(\lambda_\ell|D) \), and then take average over all the posterior samples to get an approximation of \( \lambda_\ell^2(t_s)|D \). Therefore, \( C_{x|t_*}(x,x')|D \) can be obtained/approximated by substituting \( \lambda_\ell^2(t) \) with \( \lambda_\ell^2(t_s)|D \) in [21]

\[
C_{x|t_*}(x,x')|D = \sum_{\ell=1}^{\infty} (\lambda_\ell^2(t_s)|D) \phi_\ell(x)\phi_\ell(x')
\]

\[
\approx \frac{1}{S} \sum_{s=1}^{S} \sum_{\ell=1}^{L} (\lambda_\ell(t_s)|\lambda_\ell(s))^2 \phi_\ell(x;\eta_\ell(s))\phi_\ell(x';\eta_\ell(s)), \quad \lambda_\ell(s) \sim p(\lambda_\ell|D)
\]

where \( \lambda_\ell(t)|\lambda_\ell(s) = C_u(t_*,t)C_u(t,t)\lambda_\ell(s) \).

3.2.2 Extend Evolution of Spatial Dependence to Neighbors

Recall that in the definition of \( C_{x|t} \), the fixed basis \( \{\phi_\ell(x)\} \) is taken from the eigenfunctions of the spatial kernel \( C_x \). To extend \( C_{x|t} \) 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 \( \phi_\ell(x_s) \), using its known values \( \phi_\ell := \phi_\ell(X) \) as in the conditional Gaussian:

\[
\phi_\ell(x_s)|\phi_\ell = C_X(x_s,X)C_X(X,X)^{-1}\phi_\ell = C_X(x_s,X)\lambda_\ell^{-2}\phi_\ell, \quad \forall \ell = 1, \ldots, L
\]

Then, \( C_{x|t_0}(x,x_s)|D \) can be predicted/approximated by substituting \( \phi_\ell(x') \) with \( \phi_\ell(x_s)|\phi_\ell \) in [21] as follows

\[
C_{x|t_0}(x,x_s)|D = \sum_{\ell=1}^{\infty} \lambda_\ell^2(t)|D(T)
\]

\[
\approx \frac{1}{S} \sum_{s=1}^{S} \sum_{\ell=1}^{L} (\lambda_\ell(s)^2) \phi_\ell(x;\eta_\ell(s))\phi_\ell(x_s)|\phi_\ell), \quad \lambda_\ell(s)(t) \sim p(\lambda_\ell(t)|D)
\]
Figure 2: Simulated spatiotemporal data over $[-1, 1] \times [0, 1]$, viewed in 3d (left) and projected in space-time domains (right).

4 Simulation

In this section, we focus on the study of a simulated example of spatiotemporal processes. We compare two models with different kernel formats in modeling and predicting mean and covariance functions. In particular we illustrate that model II with the Kronecker sum structure is more effective and efficient in characterizing TESD compared with model I with the Kronecker product structure.

4.1 Model Setup

Now we consider the following simulated spatiotemporal process with the spatial dimension $d = 1$.

$$y(x, t) \sim \mathcal{GP}(m, \mathcal{C}_y), \quad x \in \mathcal{X} = [-1, 1], \quad t \in \mathcal{T} = [0, 1]$$

$$m(x, t) = \cos(\pi x) \sin(2\pi t)$$

$$\mathcal{C}_y(z, z') = \exp\left(-\frac{|x - x'|^2}{2\ell_x} - \frac{|t - t'|^2}{2\ell_t} - \frac{|xt - x't'|}{2\ell_{xt}}\right) + \sigma^2 \delta(z = z') \quad (35)$$

Note that conditioned on time, we have the following true spatial covariance function (TESD)

$$C_t(x, x') := \text{Cov}[y(x, t), y(x', t)] = \exp\left(-\frac{|x - x'|^2}{2\ell_x} - \frac{|x - x'|t}{2\ell_{xt}}\right) + \sigma^2 \delta(x = x') \quad (36)$$

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 $\sigma^2 = 10^{-2}$, we generate 20301 data points $\{y_{ij}\}$ over the mesh grid. Such random process can be repeated for $K$ trials and one of them is plotted in Figure 2.
Figure 3: Selective mean functions $m(x, t)$ and covariance functions $C_t(x, x')$ fitted by model I (left two columns) and model II (right two columns) with $K = 100$ trials of data (upper row) and $K = 1000$ trials of data (lower row). Dashed lines are true values, solid curves are estimates with shaded credible regions indicating their uncertainty.

4.2 Model Fit

For simplicity we use a subset of these 20301 data points taken on an equally spaced sub-mesh with $I = 5$ and $J = 101$. Now we fit the resulting data with model I and model II in (14) respectively. In the discretized model (32), we set $\gamma_\ell = \ell^{-\kappa/2}$ with $\kappa = 2$, $a = [1, 1, 1]$, $m = [0, 0, 0]$ for both models; $b = [5, 10, 10]$, $V = [0.1, 0.1, 0.01]$ for model 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 $L = I = 5$. We run MCMC to collect $2.4 \times 10^4$ samples, burn in the first 4000, and subsample every other. The resulting 10^4 posterior samples are used to estimate the mean function $m(\{y_{ij}\})$ and the covariance function $C_{y_{ij}}(\{y_{ij}\})$. (See more details in Appendix B). Their estimates at selective locations are plotted in Figure 3. With growing data information (increasing trial number $K$) 

most posterior estimates contract towards some values (with decreasing width of credible bands). Both models can ‘recover’ the true mean functions as their MCMC estimates get closer to the truth, however, model I fails to properly represent the true covariance functions, the underlying TESD. Due to the separability assumption (9) for both location and time in the likelihood (lack of structure), the posterior of covariance functions (conditioned on time) in model I, dominated by the likelihood, is forced to concentrate on ‘flat’ functions of $t$, as seen in the second column of Figure 3. However model II, by contrast, correctly captures TESD, as conveyed in the

---

3More repeated trials ($K$) can be viewed as increasing data ($n$) though they are stacked over the same discrete time points $\{t_j\}_{j=1}$. See more contraction results by increasing $J$ in [Lan et al., 2019] for $N$ in Figure 5.)

18
4.3 Model Prediction

Now we consider the predictions for both mean and covariance. The prediction of mean functions has been well studied in the literature. For simplicity we only illustrate the mean prediction in the time direction, i.e. $m(x, t_*)|D$. We hold out 15% data for testing as indicated by the black dash-dot ticks in subplots of Figure 4. Then we train the models (solid lines with light shaded regions) based on the rest 85% data (shorter gray ticks) and use them to predict the mean functions on the testing data set. The thick dash-dot lines are the predicted values of mean with dark shaded regions as the corresponding credible bands. In general, with more data (trials) we get better fit and prediction as they are closer to the truth, illustrated by the dash lines. In this example, model II gives similar
Figure 5: Prediction of covariance functions at future times $t_*$ (left column), $C_{t*}(x, x')$, and at neighboring locations $x_*$ (right column), $C_t(x, x_*)$, by model II based on $K = 100$ trials (upper row) and $K = 1000$ trials (lower row) of data.

Lastly, we consider the prediction of covariance functions. Since model I cannot learn the covariance function well, we only test model II for the covariance prediction. Here we consider two types of prediction, namely TESD to future (Section 3.2.1) and TESD to neighbor (Section 3.2.2). The left column of Figure 5 shows the prediction of covariance functions in the time direction, i.e. $C_{x|t_*}(x, x)$. Again we hold out 15% testing data (black dash-dot ticks) and train the generalized STGP model II based on the rest 85% data (shorter gray ticks). In general, the predicted values of the covariance in this case follow the trend of the fitted values with higher certainty (narrower credible bands) when there are still nearby training points (interpolation) but become more uncertain (wider credible bands) while entering the ‘no-data’ region (extrapolation). Now we consider the second type of covariance prediction, TESD to a new neighbor (e.g. $x_* = 0.1$), $C_{x|t_*}(x, x_*)$. Note, we do not have any data at this location $x_* = 0.1$, nor can we fit model II for covariances involving this spatial point. The right column of Figure 5 shows that model II can extend TESD (thick dash-dot lines) to new locations with decent precision, in the reference to the truth, indicated by the dash lines. Again, such prediction becomes more credible (narrower
bands) with increasing data information (trials). This prediction is of particular interest since we can ‘lend’ TESD information to new neighbors based on existing knowledge learnt from the data. It can enable us to discover new mechanism of some underlying process (e.g. global warming) at uninformed locations (e.g. unobserved territories).

We will apply the generalized STGP models to Alzheimer’s neuroimaging data to study the association of brain regions in the progression of the disease in the next section. Since model I fails to characterize TESD and it involves much heavier computation, we will only focus on model II in the following.

5 Longitudinal Analysis of Alzheimer’s Brain Images

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. It is one of the globally prevalent diseases that cost billions or even trillions of dollars every year. 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 (ADNI) project (ADN). A distinctive feature of these neuroimaging data is that they contain both spatial and temporal information that is not separable from each other. We will use the proposed generalized STGP: (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.

5.1 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 till 2 years. These subjects are chosen because they participate the study for the longest time available so that we can investigate the longitudinal change of their brains over a sufficiently long period. Each subject takes fludeoxyglucose (FDG), an analogue of glucose, as the biologically active tracer that could indicate the tissue metabolic activity. PET brain image scans are obtained 30-60 minutes post-injection and then processed by co-regsitering 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.
Though our proposed method can be applied to the whole 3d images, we focus on a (48-th) slice in the middle (horizontal section) and model the images of size 160 \times 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 \times 25600, which is enormous if it is a full matrix. In the following we introduce a spatial kernel based on the graph Laplacian. The resulting precision matrix is highly sparse and thus amenable to an efficient learning of TESD. Graph Laplacian has been a popular tool in the analysis of brain images (Shen et al., 2010; Ng et al., 2012; Hu et al., 2015; Huang et al., 2018).

5.2 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 et al., 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

$$L = D - W, \quad W = [w_{ij}], \quad w_{ij} = \eta_e(|x_i - x_j|), \quad D = \text{diag}\{d_{ii}\}, \quad d_{ii} = \sum_{x_j \sim x_i} w_{ij}$$ (37)
where \( \eta_\epsilon \) 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 \( \mu \) supported on the graph domain \( \Omega \) with smooth Lebesgue density \( \rho \) bounded above and below by positive constants, then \( L \) can be viewed as an approximation of the Laplace operator \( L \) in the following PDE:

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

Based on the graph Laplacian, we can define the following discrete spatial kernel for the brain images (Dunlop et al., 2018)

\[
C_x = (s_n L + \tau^2 I)^{-s}, \quad s_n = o \left( \frac{1}{n^{1-2/d} \log^{\delta(d-2)/2+2/d} n} \right)
\]  

(38)

where \( d \) is the spatial dimension, i.e. \( d = 2 \) for the chosen slice of brain images. We choose \( s_n = \frac{1}{n^{1-2/d} \log^{\delta(d-2)/2+2/d} n} \) in this experiment. Further assuming conditions (open, connected, and with smooth boundary) on the graph domain \( \Omega \), Dunlop et al. (2018) prove that for \( s > d/2 \) and \( \tau \geq 0 \), Gaussian measure \( \mathcal{N}(0, C) \) with \( C = (L + \tau^2 I)^{-s} \) is well-defined on the weighted Hilbert space \( L^2_\mu \). In another word, the graph-Laplacian based spatial kernel (38) is well-behaved for large graphs including the brain images we investigate with \( n = 25600 \) nodes.

To obtain the spatial kernel (38) for the analysis of PET scans, we first construct the graph Laplacian. On the 160 \( \times \) 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 \( 25600 \times 25600 \) but is highly...
Figure 8: Estimated correlation between the brain region of interest and a selected point of interest (red cross) for CN (top row), MCI (middle row) and AD (bottom row) respectively.

spare (with the density of non-zero entries $3.4864 \times 10^{-4}$). We also assume a hyper-prior for $\tau^2 \sim \log -N(m_X, V_X)$ and fix $s = 2$ in this experiment. Then for given $\tau^2$, we calculate the precision matrix $C_x^{-1}$ based on (38). Hence the full sized covariance matrix $C_x$ is not directly calculated in the inference procedure.

5.3 Model Fit

Now we fit model II with the graph-Laplacian based spatial kernel (38) 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)-(21) for $C_x(t)$. It results in an improper prior, however regularized by the likelihood (See more details in Figure 7). 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 $2.4 \times 10^4$ samples, burn in the first 4000, and subsample every other. The resulting $10^4$ samples are used to obtain posterior estimates of mean functions $M(t)$ and covariance functions $C_y(t)$. Figure 6 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 Appendix C. Figure 7 compares the dynamic
Figure 9: Estimated connection of the brain images for CN (top row), MCI (middle row) and AD (bottom row) respectively.

eigenvalues $\lambda^2_\ell(t)$ for different groups. Interestingly, they do not decrease monotonically in $\ell$ (on y-axis) but rather damp out as $\ell$ becomes larger. When $\ell$ gets close to $L = 100$, the magnitude of $\lambda^2_\ell(t)$ becomes small enough to be negligible. To investigate TESD in the PET brain images, we select a point of interest (POI) (marked as red cross) in the occipital lobe and plot the correlation between the brain region of interest (ROI) (chosen based on the pixel values above the 83.5% quantile) and the selected POI in Figure 8. The spatial correlation evolves with time. We can see that 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.

TESD of the brain images in the discretize field is a (covariance) matrix valued function of time $t$. At each time the covariance matrix is of size $25600 \times 25600$. To better summarize it, we threshold the $(25600^2)$ absolute correlation values at the top 10%, then we obtain the adjacency matrix $(25600 \times 25600)$ based on the nonzero values of the correlation. Finally we define the connection graph as the diagonal of the degree matrix (row sums of the adjacency) projected back to $160 \times 160$ mesh. Therefore, the value of each point on the connection graph indicates how many nodes are connected to it. Figure 9 plots the connection graphs of the brain ROI for different groups. For each of these connection graphs in Figure 9, the truncation at any value yields a network of connected nodes that are the most active. As seen from Figure 9, 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
5.4 Model Prediction

Now we hold out the data at the last time point for testing. For each group, the generalized STGP model II 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 10 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 11 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 8. 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 12 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 8. Such extension provides more fine details of TESD at new locations without data. There are more numerical results the neuroimaging analysis in Appendix C.
Figure 11: Predicted correlation between the brain region of interest and a selected point of interest (red cross) for CN (left), MCI (middle) and AD (right) respectively.

6 Conclusion

In this paper, we generalize the separable spatiotemporal Gaussian process (STGP) to model the temporal evolution of spatial dependence (TESD) in given data. Instead of treating the space variable $x$ and the time variable $t$ as an extended variable $z = (x, t)$ in the traditional non-separable STGP [Marco et al., 2015; Datta et al., 2016; Hyun et al., 2016], we generalize STGP by varying the eigenvalues of the spatial kernel in the Mercer’s representation. Theoretic properties of such time-dependent spatial kernel, including the convergence, the regularity of random draws from the prior and the posterior contraction, have been thoroughly investigated. Two Bayesian nonparametric models are introduced based on the generalized STGP with the joint kernels structured as the Kronecker product and the Kronecker sum respectively. The Kronecker sum structure is proved to be superior to the Kronecker product in characterizing TESD, from both modeling and computing perspectives. The advantage of the Kronecker sum kernel is demonstrated by a simulation study of spatiotemporal process. The generalized STGP model based on this structure is then applied to analyze PET brain images of Alzheimer’s patients recorded for up to 3 years to describe 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 evidence verifies the effectiveness and efficiency of the proposed model in characterizing TESD.

There is room to improve the current method. 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 $b_\ell / a_\ell = O(\ell^{-\kappa / 2})$ 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 12: Extended correlation between the brain region of interest and a selected point of interest from the coarse mesh (upper) to the fine mesh (lower).

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 (ADN) 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, we hope they can facilitate more accurate description of TESD of AD brain images to aid the exploration 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 as AD, which could shed more light on the reason of such disease.

Acknowledgements

Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf
References

The alzheimer’s disease neuroimaging initiative. http://adni.loni.usc.edu.

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

Chung, F., Graham, F., Society, A. M., on Recent Advances in Spectral Graph Theory, C. C., and of the Mathematical Sciences, C. B. (1997). *Spectral Graph Theory*. Number no. 92 in CBMS Regional Conference Series. American Mathematical Society.

Cribben, I., Haraldsdottir, R., Atlas, L. Y., Wager, T. D., and Lindquist, M. A. (2012). Dynamic connectivity regression: Determining state-related changes in brain connectivity. *NeuroImage*, 61(4):907 – 920.

Dashti, M. and Stuart, A. M. (2017). The bayesian approach to inverse problems. *Handbook of Uncertainty Quantification*, pages 311–428.

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A. S., and Schaap, M. (2016). 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.

Dunlop, M. M., Slepčev, D., Stuart, A. M., and Thorpe, M. (2018). Large data and zero noise limits of graph-based semi-supervised learning algorithms. *arXiv:1805.09450*.

Fiecas, M. and Ombao, H. (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association*, 111(516):1440–1453.

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

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

Ghosal, S. and van der Vaart, A. (2017). Fundamentals of nonparametric bayesian inference.

Hairer, M. (2009). An introduction to stochastic pdes. *arXiv:0907.4178*.

Hartikainen, J., Riihimäki, J., and Särkkä, S. (2011). Sparse spatio-temporal gaussian processes with general likelihoods. In Honkela, T., Duch, W., Girolami, M., and Kaski, S., editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 193–200, Berlin, Heidelberg. Springer Berlin Heidelberg.
Hu, C., Cheng, L., Sepulcre, J., Johnson, K. A., Fakhri, G. E., Lu, Y. M., and Li, Q. (2015). A spectral graph regression model for learning brain connectivity of alzheimer’s disease. *PLOS ONE*, 10(5):e0128136.

Huang, W., Bolton, T. A. W., Medaglia, J. D., Bassett, D. S., Ribeiro, A., and Ville, D. V. D. (2018). A graph signal processing perspective on functional brain imaging. *Proceedings of the IEEE*, 106(5):868–885.

Hyun, J. W., Li, Y., Huang, C., Styner, M., Lin, W., and Zhu, H. (2016). Stgp: Spatio-temporal gaussian process models for longitudinal neuroimaging data. *NeuroImage*, 134:550–562.

Kuzin, D., Isupova, O., and Mihaylova, L. S. (2018). Spatio-temporal structured sparse regression with hierarchical gaussian process priors. *IEEE Transactions on Signal Processing*, 66:4598–4611.

Lan, S., Holbrook, A., Elias, G. A., Fortin, N. J., Ombao, H., and Shahbaba, B. (2019). Flexible bayesian dynamic modeling of correlation and covariance matrices. *Bayesian Analysis (to appear).*

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53.

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

Lindström, J., Szpiro, A., Sampson, P. D., Bergen, S., and Sheppard, L. (2013). Spatiotemporal : An r package for spatio-temporal modelling of air-pollution.

Liu, J. (2013). Gaussian process regression using spatial-temporal covariance : Research on sea level prediction.

Luttinen, J. and Ilin, A. (2012). Efficient gaussian process inference for short-scale spatio-temporal modeling. In Lawrence, N. D. and Girolami, M., 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. PMLR.

Marco, L., Ziegler, G., Alexander, D. C., and Ourselin, S. (2015). Modelling non-stationary and non-separable spatio-temporal changes in neurodegeneration via gaussian process convolution. In Bhatia, K. and Lombaert, H., editors, *Machine Learning Meets Medical Imaging*, pages 35–44, Cham. Springer International Publishing.

Murray, I., Adams, R. P., and MacKay, D. J. (2010). Elliptical slice sampling. *JMLR: W&CP*, 9:541–548.
Nabarro, S., Fletcher, T., and Shawe-Taylor, J. (2018). Spatiotemporal prediction of ambulance demand using gaussian process regression. CoRR, abs/1806.10873.

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

Ng, B., Siless, V., Varoquaux, G., Poline, J., Thirion, B., and Abgharbieh, R. (2012). Connectivity-informed sparse classifiers for fmri brain decoding. In 2012 Second International Workshop on Pattern Recognition in NeuroImaging, pages 101–104.

Niu, M., Dai, Z., Lawrence, N., and Becker, K. (2015). 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.

Report, W. A. (2018). The state of the art of dementia research: New frontiers. https://www.alz.co.uk/research/world-report-2018.

Sarkka, S. and Hartikainen, J. (2012). Infinite-dimensional kalman filtering approach to spatio-temporal gaussian process regression. In Lawrence, N. D. and Girolami, M., 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. PMLR.

Sarkka, S., Solin, A., and Hartikainen, J. (2013). 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.

Senanayake, R., O’Callaghan, S. T., and Ramos, F. T. (2016). Predicting spatio-temporal propagation of seasonal influenza using variational gaussian process regression. In AAAI.

Shen, X., Papademetris, X., and Constable, R. (2010). Graph-theory based parcellation of functional subunits in the brain from resting-state fmri data. NeuroImage, 50(3):1027–1035.

Singh, A., Ramos, F., Whyte, H. D., and Kaiser, W. J. (2010). Modeling and decision making in spatio-temporal processes for environmental surveillance. In 2010 IEEE International Conference on Robotics and Automation, pages 5490–5497.

Smola, A. J. and Kondor, R. (2003). Kernels and regularization on graphs. In Schölkopf, B. and Warmuth, M. K., editors, Learning Theory and Kernel Machines, pages 144–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Soh, H., Su, Y., and Demiris, Y. (2012). Online spatio-temporal gaussian process experts with application to tactile classification. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 4489–4496.
Todescato, M., Carron, A., Carli, R., Pillonetto, G., and Schenato, L. (2017). Efficient spatio-temporal gaussian regression via kalman filtering.

van der Vaart, A. and van Zanten, H. (2011). Information rates of nonparametric gaussian process methods. *J. Mach. Learn. Res.*, 12:2095–2119.

van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675.

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

**Theorem 2.1.** [Wellposedness of Mercer’s Kernels] Under Assumption 1, both \( C^I_m = C^I_x \otimes C^I_t \) and \( C^{II}_m = C^I_x \otimes I_t \) are well defined non-negative definite kernels on \( Z = \mathcal{X} \times \mathcal{T} \).

**Proof of Theorem 2.1.** We first prove both series (22) and (23) converge in \( L^1(\mathcal{Z} \times \mathcal{Z}) \).

For (22) we have

\[
\sum_{\ell=1}^{\infty} \left| \int_{\mathcal{Z}} \int_{\mathcal{Z}} \lambda_\ell(t) C_\ell(t, t') \lambda_\ell(t') \phi_\ell(x) \phi_\ell(x') dz \right| dt \leq \sum_{\ell=1}^{\infty} \left( \int_{\mathcal{T}} \int_{\mathcal{T}} \lambda_\ell(t) C_\ell(t, t') dt dt' \right) \left( \int_{\mathcal{X}} \int_{\mathcal{X}} |\phi_\ell(x) \phi_\ell(x')| dx dx' \right) \\
\leq \sum_{\ell=1}^{\infty} \left| \langle \lambda_\ell, C_\ell \lambda_\ell \rangle \right| \left( \| \phi_\ell(x) \|_2^2 + \| \phi_\ell(x') \|_2^2 \right) / 2 \\
\leq \| C_t \| \| \lambda_\ell \|_2^2 = \| C_t \| \| \lambda_\ell \|_2^2 < +\infty
\]

And for (23) we can bound

\[
\sum_{\ell=1}^{\infty} \left| \int_{\mathcal{Z}} \int_{\mathcal{Z}} \lambda_\ell^2(t) \delta_\ell(t') \phi_\ell(x) \phi_\ell(x') dz \right| dt \leq \sum_{\ell=1}^{\infty} \lambda_\ell^2(t) dt \left( \int_{\mathcal{X}} \int_{\mathcal{X}} |\phi_\ell(x) \phi_\ell(x')| dx dx' \right) \leq \| \lambda_\ell \|_2^2 < +\infty
\]

The convergence of series (22) and (23) follows by the dominated convergence theorem.

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

Then we have

\[
\langle f(z), C^I_m f(z') \rangle = \langle f(z), \sum_{\ell=1}^{\infty} \lambda_\ell(t) C_\ell(t, t') \lambda_\ell(t') \phi_\ell(x) \phi_\ell(x') f(z') dz' \rangle \\
\quad = \langle f(z), \sum_{\ell=1}^{\infty} \int_{\mathcal{T}} \lambda_\ell(t) C_\ell(t, t') \lambda_\ell(t') \phi_\ell(x) f_\ell(t') dt' \rangle \\
\quad = \sum_{\ell=1}^{\infty} \int_{\mathcal{T}} \int_{\mathcal{T}} f_\ell(t) \lambda_\ell(t) C_\ell(t, t') \lambda_\ell(t') f_\ell(t') dt dt' \\
\quad = \sum_{\ell=1}^{\infty} \langle f_\ell \lambda_\ell, C_\ell \lambda_\ell f_\ell \rangle \geq 0
\]

33
where the convergence can be shown as above. Similarly we have

\[
\langle f(z), C^{II}_{y|m} f(z') \rangle = \langle f(z), \sum_{t=1}^{\infty} \lambda_t^2(t) \delta(t = t') \phi_t(x) \phi_t(x') f(z') dz' \rangle
\]

\[
= \langle f(z), \sum_{t=1}^{\infty} \int_{T} \lambda_t^2(t) \delta(t = t') \phi_t(x) f_t(t') dt' \rangle = \sum_{t=1}^{\infty} \int_{T} \lambda_t^2(t) f_t^2(t) dt \geq 0
\]

Therefore we complete the proof.

\[\square\]

**Theorem 2.2.** [Karhunen-Loève Expansion] Under Assumption 1, STGP \(f(x, t) \sim \mathcal{GP}(0, C_z)\) has the following representation of series expansion:

\[
f(x, t) = \sum_{t=1}^{\infty} f_t(t) \phi_t(x), \quad f_t(t) = \int_{\mathcal{X}} f(x, t) \phi_t(x) dx \tag{25}
\]

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

- if \(C_z = C_{x|t}^{1/2} C_{x|t'}^{1/2} \otimes C_t\), then \(E[f_t(t) f_t(t')] = \lambda_t(t) C_t(t, t') \lambda_t(t') \delta_{tt'}\).
- if \(C_z = C_{x|t} \otimes I_t\), then \(E[f_t(t) f_t(t')] = \lambda_t^2(t) \delta(t = t') \delta_{tt'}\).

**Proof of Theorem 2.2.** Note that \(\{\phi_t(x)\}_{t=1}^{\infty}\) is an orthonormal basis for \(L^2(\mathcal{X})\), therefore we have the series representation (25) of \(f(\cdot, t)\) for each \(t \in T\). Then we can calculate for \(C_{x|t}^{1/2} C_{x|t'}^{1/2} \otimes C_t\) as in (22):

\[
E[f_t(t)] = E \left[ \int_{\mathcal{X}} f(x, t) \phi_t(x) dx \right] = \int_{\mathcal{X}} E[f(x, t)] \phi_t(x) dx = 0
\]

\[
E[f_t(t) f_t(t')] = E \left[ \int_{\mathcal{X}} f(x, t) \phi_t(x) dx \int_{\mathcal{X}} f(x', t') \phi_t(x') dx' \right]
\]

\[
= \int_{\mathcal{X}} \int_{\mathcal{X}} E[f(x, t) f(x', t')] \phi_t(x) \phi_t(x') dx dx'
\]

\[
= \int_{\mathcal{X}} \int_{\mathcal{X}} C_{x|t}^{1/2} C_{x|t'}^{1/2} \otimes C_t(z, z') \phi_t(z) \phi_t(z') dx dz'
\]

\[
= \sum_{t=1}^{\infty} \lambda_t^2(t) C_t(t, t') \lambda_t(t') \int_{\mathcal{X}} \phi_t(x) \phi_t(x') \phi_t(x) \phi_t(x') dx dx'
\]

Similarly we have \(E[f_t(t) f_t(t')] = \lambda_t^2(t) \delta(t = t') \delta_{tt'}\) for \(C_z = C_{x|t} \otimes I_t\) using the definition (23).
Lastly, we prove the convergence of the infinite sum. Denote \( F_L(x, t) = \sum_{\ell=1}^{L} f_\ell(t) \phi_\ell(x) \).

We have for \( \mathbb{C}_{\mathbb{X}|t}^{1/2} \mathbb{C}_{\mathbb{X}|t'}^{1/2} \otimes \mathbb{C}_t \)

\[
\mathbb{E}[|f - F_L|^2] = \mathbb{E}[f^2] + \mathbb{E}[F_L^2] - 2\mathbb{E}[f F_L]
\]
\[
= \mathbb{C}_{\mathbb{X}|t}^{1/2} \mathbb{C}_{\mathbb{X}|t'}^{1/2} \otimes \mathbb{C}_t(z, z) + \mathbb{E} \left[ \sum_{\ell=1}^{L} \sum_{\ell'=1}^{L} f_\ell(t) f_{\ell'}(t) \phi_\ell(x) \phi_{\ell'}(x) \right] - 2F_L \left[ \sum_{\ell=1}^{L} f_\ell(t) \phi_\ell(x) \right]
\]
\[
= \mathbb{C}_{\mathbb{X}|t}^{1/2} \mathbb{C}_{\mathbb{X}|t'}^{1/2} \otimes \mathbb{C}_t(z, z) + \sum_{\ell=1}^{L} \lambda_\ell^2(t) C_t(t, t) \phi_\ell^2(x) - 2 \sum_{\ell=1}^{L} \int_{\mathcal{X}} \mathbb{E}[f(x, t) f(x', t)] \phi_\ell(x) \phi_{\ell'}(x') dx'
\]
\[
= \mathbb{C}_{\mathbb{X}|t}^{1/2} \mathbb{C}_{\mathbb{X}|t'}^{1/2} \otimes \mathbb{C}_t(z, z) - \sum_{\ell=1}^{L} \lambda_\ell^2(t) C_t(t, t) \phi_\ell^2(x) = \sum_{\ell=L+1}^{\infty} \lambda_\ell^2(t) C_t(t, t) \phi_\ell^2(x) \rightarrow 0, \text{ as } L \rightarrow \infty
\]

The same argument (by replacing \( C(t, t') \) with \( \delta(t = t') \)) yields the \( L^2_B \) convergence of the expansion \([25]\) for \( C_z = \mathbb{C}_{\mathbb{X}|t} \otimes \mathbb{T}_t \).

\[\square\]

**Theorem 2.3.** [Regularity of Random Functions] Assume \( \lambda \in \ell^{2,s}(L^2(\mathcal{T})) \). If \( f(x, t) \sim \mathcal{G}\mathcal{P}(0, \mathbb{C}_z) \) as in Theorem 2.2, then \( f = \sum_{\ell=1}^{\infty} f_\ell(t) \phi_\ell(x) \in \ell^{2,s}(L^2(\mathcal{Z})) \) in probability.

Moreover, under Assumption \([27]\), there is a version \( \hat{f}(x) \) of \( f(x) := \int_{\mathcal{T}} f(x, t) dt \) in \( C^{0,s'}(\mathcal{X}) \) for \( s' < s \). If further \( \mathbb{C}_z = \mathbb{C}_{\mathbb{X}|t}^{1/2} \mathbb{C}_{\mathbb{X}|t'}^{1/2} \otimes \mathbb{C}_t \) and \( \{Q_{\lambda, \mathcal{C}_t}\} \) satisfies Assumption \([28]\), then there is a version \( \hat{f}(z) \) of \( f(z) \) in \( C^{0,s'}(\mathcal{Z}) \) for \( s' < s \).

**Proof of Theorem 2.3.** We compute the expectation of the \((2, s, 2)\)-norm of \( f \)

\[
\mathbb{E}[\|f\|_{2,s,2}^2] = \sum_{\ell=1}^{\infty} \ell^{2s} \mathbb{E}[\|f_\ell\|_2^2] = \sum_{\ell=1}^{\infty} \ell^{2s} \int_{\mathcal{T}} \mathbb{E}[f_\ell^2(t)] dt
\]
\[
= \left\{ \begin{array}{ll}
\sum_{\ell=1}^{\infty} \ell^{2s} \int_{\mathcal{T}} C_t(t, t) \lambda_\ell^2(t) dt & \leq \sum_{\ell=1}^{\infty} \ell^{2s} \|\lambda_\ell\|_2^2, \\
\sum_{\ell=1}^{\infty} \ell^{2s} \|\lambda_\ell\|_2^2, & \text{if } C_z = \mathbb{C}_{\mathbb{X}|t}^{1/2} \mathbb{C}_{\mathbb{X}|t'}^{1/2} \otimes \mathbb{C}_t
\end{array} \right.
\]
\[
= \|\lambda\|_{2,s,2}^2 < +\infty
\]

where the equality on the second line follows from Theorem 2.2. This implies \( f \in \ell^{2,s}(L^2(\mathcal{T})) \) in probability.

\footnote{A version/modification of stochastic process \( \hat{f}(x) \) of \( f(x) \) means \( \mathbb{P}[\hat{f}(x) = f(x)] = 1 \) for \( \forall x \in \mathcal{X} \).}
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 of Dashiti and Stuart 2017). First, we consider the marginal function. By Jensen’s inequality

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

\[
= \sum_{\ell} \|\lambda_\ell\|_2^2 \phi_\ell(x) - \phi_\ell(x')|^2 \leq \sum_{\ell} \|\lambda_\ell\|_2^2 \min\{2\|\phi_\ell\|_\infty^2, \text{Lip}(\phi_\ell)^2\}|x - x'|^2
\]

\[
\leq 2 \sum_{\ell} \|\lambda_\ell\|_2^2 \|\phi_\ell\|_\infty^{-\delta} \text{Lip}(\phi_\ell)^{\delta/2} |x - x'|^\delta \leq \sum_{\ell} \ell^\delta \|\lambda_\ell\|_2^2 |x - x'|^\delta
\]

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

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

Lastly, we consider the full function \( f(x, t) \).

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

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

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

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

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

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

where \( \Delta f := f(t) - f(t') \), \( \Delta \phi := \phi(x) - \phi(x') \), 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 2.1.** If \( f(x, t) \sim GP(0, C_x) \) has a continuous version, then \( \{f_\ell\}_{\ell = 1}^\infty \) as in Theorem 2.2 are GP’s defined on \( T \).
Proof of Corollary 2.1: \( f_\ell(t) \) can be viewed as infinite weighted sum of GP \( f(x, t) \) thus becomes another GP. This can be made rigorous by approximating \( \phi_\ell(x) \) with a sequence of simple functions \( \phi_{n, \ell} = \sum_{i=1}^n a_i 1_{A_i} \) with disjoint \( \{A_i\} \):

\[
f_\ell(t) = \int_X f(x, t) \phi_\ell(x) dx = \lim_{n \to +\infty} \int_X f(x, t) \phi_{n, \ell}(x) dx = \sum_{i=1}^\infty a_i \int_{A_i} f(x, t) dx
\]

Note for \( \forall t \in \mathcal{T} \), \( f_{A_i}(t) := \int_{A_i} f(x, t) dx \) 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 \( t = (t_1, \cdots, t_k) \) replacing \( t \). Thus it concludes the proof.

For the dynamic spatial kernels \( C_i = \sum_{\ell=1}^{\infty} \lambda_{i, \ell}^2(t) \phi_\ell \otimes \phi_\ell \), we consider the Gaussian likelihood models \( p_i \sim \mathcal{N}_n(m_i(t), C_i(t)) \), with \( C_i = \sum_{\ell=1}^n \lambda_{i, \ell}^2(t) \phi_\ell(x) \otimes \phi_\ell(x') = \Phi A_i \Phi^T \), 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 \( (K(p_0, p_1) := E_0(\log(p_0/p_1)) \) and K-L variation \( (V(p_0, p_1) := E_0(\log(p_0/p_1))^2) \) between two models with their difference in eigenvalues measured by \( || \cdot ||_{1,s,\infty} \) in the following lemma.

**Lemma A.1.** Let \( p_i \sim \mathcal{N}_n(0, C_i(t)) \) be Gaussian models for \( i = 0, 1 \), with \( \{\lambda_{i, \ell}^2(t)\} \) being the eigenvalues of \( C_i = \Phi A_i \Phi^T \) satisfying Assumption 3. Then we have

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

**Proof.** First we calculate the Kullback-Leibler divergence

\[
K(p_0, p_1) = \frac{1}{2} \left\{ \text{tr}(C_0^{-1}C_0 - I) + (m_1 - m_0)^T C_1^{-1}(m_1 - m_0) + \log \frac{|C_0|}{|C_1|} \right\}
\]

Consider \( m_i \equiv 0 \). By the non-negativity of K-L divergence we have for general \( C_i > 0 \),

\[
\log \frac{|C_0|}{|C_1|} \leq \text{tr}(C_1^{-1}C_0 - I) \tag{A.1}
\]

Therefore we can bound K-L divergence

\[
K(p_0, p_1) \leq \frac{1}{2} \left\{ \text{tr}(C_1^{-1}C_0 - I) + \text{tr}(C_0^{-1}C_1 - I) \right\} \leq 2C ||\lambda_0 - \lambda_1||_{1,s,\infty}
\]

where we use

\[
\text{tr}(C_1^{-1}C_0 - I) = \sum_{\ell=1}^n \lambda_{0, \ell}^{-2}(t)(\lambda_{0, \ell}^2(t) - \lambda_{1, \ell}^2(t)) \leq 2C \sum_\ell c_\ell^{-2}||\lambda_0 - \lambda_1||_{\infty} \leq 2C ||\lambda_0 - \lambda_1||_{1,s,\infty} \tag{A.2}
\]
Now we calculate the following K-L variation

\[ V(p_0, p_1) = \frac{1}{2} \text{tr}((C_1^{-1}C_0 - I)^2) + (m_1 - m_0)^T C_1^{-1}C_0 C_1^{-1}(m_1 - m_0) + K^2(p_0, p_1) \]

Consider \( m_i \equiv 0 \) and we can bound it by the similar argument as (A.2)

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

It is easy to see that the centered K-L variation \( V_0(p_0, p_1) = \text{Var}_0(\log(p_0/p_1)) = E_0(\log(p_0/p_1) - K(p_0, p_1))^2 \) 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

\[ h^2(p_0, p_1) = 1 - \frac{|C_0 C_1|^{1/4}}{|C_0 + C_1|^{1/4}} \exp \left\{ -\frac{1}{8} (m_0 - m_1)^T \left( \frac{C_0 + C_1}{2} \right)^{-1} (m_0 - m_1) \right\} \]

Consider \( m_i \equiv 0 \). 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)

\[ h^2(p_0, p_1) \leq \frac{1}{4} \{ \text{tr}(C_1^{-1/2} C_0^{-1/2} - I) + \text{tr}(C_0^{-1/2} C_1^{-1/2} - I) \} \leq \frac{1}{2} \| \lambda_0 - \lambda_1 \|_{1,s,\infty}^2 \]

\[ \square \]

**Theorem 2.4.** [Posterior Contraction of \( C_{\lambda} \) in model II] Let \( \lambda \) be a Borel measurable, zero-mean, tight Gaussian random element in \( \Theta = L^2(L^2(T)) \) satisfying Assumption 3 and \( P_{\lambda}^{(n)} = \otimes_{j=1}^n P_{\lambda,j} \) be the product measure of \( Y^{(n)} \) parametrized by \( \lambda \). If the true value \( \lambda_0 \in \Theta \) is in the support of \( \lambda \), and \( \varepsilon_n \) satisfies the rate equation \( \varphi_{\lambda_0}(\varepsilon_n) \leq n\varepsilon_n^2 \) with \( \varepsilon_n \geq n^{-1/2} \), 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 of Theorem 2.4.** 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 \( k \in \mathbb{N} \):

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

\[ \frac{\Pi_n(\lambda \in \Theta_n : k\varepsilon_n < d_{n,H}(\lambda, \lambda_{n,0}) < 2k\varepsilon_n)}{\Pi_n(B_{\varepsilon_n}(\lambda_{n,0}, \varepsilon_n))} \leq e^{Kn\varepsilon_n^2k^2/2} \tag{A.4} \]
where the left side of (A.3) is called Le Cam dimension \cite{LeCam1973, LeCam1975}, logarithm of the minimal number of \( d_{n,H} \)-balls of radius \( \xi \varepsilon /2 \) needed to cover a ball of radius \( \varepsilon \) around the true value \( \lambda_{n,0} : B_{n}(\lambda_{n,0}, \varepsilon) := \{ \lambda \in \Theta : \frac{1}{n} \sum_{j=1}^{n} K_j(\lambda_{n,0}, \lambda) \leq \frac{\varepsilon^2}{2} \} \). Now by Lemma A.1 and (A.8) the following global entropy bound because \( N \) and \( \varepsilon \) are easier to bound (Lemma B.1). Note we do not have the complementary assertion as in Lemma 1 of \cite{Ghosal2007} such that

\[
\log N(3\varepsilon_{n,\ell}, B_{n,\ell}, \|\cdot\|_{\infty}) \leq 6Cn\varepsilon_{n,\ell}\varepsilon^2
\]

\[
\Pi_{\ell}(\|\lambda_{\ell} - \lambda_{0,\ell}\|_{\infty} < 2\varepsilon_{n,\ell}) \geq e^{-n\varepsilon_{n,\ell}^2}
\]

Now let \( \varepsilon_{n,\ell} = 2^{-\ell}\ell^{-s}\varepsilon_n^2 \) for \( \ell = 1, \ldots, n \). Set \( \Theta_{n \ell} = \{ \lambda \in \Theta \cap \ell^{1,s}(L^{\infty}(T)) : \lambda_{\ell} \in B_{n,\ell} \} \subset \Theta \), and \( N(\varepsilon_{n,\ell}, \Theta_{n \ell}, d_{n,H}) = \max_{1 \leq \ell \leq n} N(3\varepsilon_{n,\ell}, B_{n,\ell}, \|\cdot\|_{\infty}) \). 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 \( \forall \lambda, \lambda' \in \Theta_{n \ell} \).

\[
\log N(\varepsilon_{n,\ell}, \Theta_{n \ell}, d_{n,H}) \leq 6C(2^{-\ell}\ell^{-s}\varepsilon_n^2)^2 \leq Cn\varepsilon_n^2 \leq n\varepsilon_n^2
\]

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

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

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

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

\textbf{Remark 5.} This theorem generalizes Theorem 2.2 of \cite{Lan2019} where the spatial domain has fixed dimension \( D \). 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 \cite{Ghosal2007} such that the resulting contraction is only on \( \Theta_{n \ell} \), weaker than that in Theorem 2.2 of \cite{Lan2019}.

39
Theorem 2.5. [Posterior Contraction Rate of $C_{\mathcal{X};\ell}$ in model II] Let $\lambda$ be a Gaussian random element defined in (20) with $\gamma_\ell = \Theta(\ell^{-\kappa/2})$ for some $\kappa > 1$ and $\text{tr}(C_u) = 1$. The rest settings are the same as in Theorem 2.4. If the true value $\lambda_0 \in \ell^{2,s}(L^2(T))$, then we have the rate of posterior contraction $\varepsilon_n = \Theta(n^{-\frac{\kappa+2}{\kappa+1}/s})$.

Proof of Theorem 2.5. First, we prove that the negative logarithm of small ball probability $\varphi(\varepsilon) = -\log \Pi(\|\lambda\|_{2,2} < \varepsilon) = O(\varepsilon^{-\frac{2}{(\kappa-1)}})$. Apply Karhunen-Loève theorem to $u_\ell$ in model (20) to get $u_\ell(t) = \sum_{i=1}^{\infty} Z_{\ell,i} \xi_i(t)$ with $Z_{\ell,i} \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\{\xi^2_i, \phi_i\}$ being the eigen-pairs of $C_u$. Note $\mathbb{E}[\|u_\ell\|_{2,2}^2] = \sum_{i=1}^{\infty} \xi^2_i = \text{tr}(C_u) = 1$. 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

$$
P\left(\sum_{\ell \leq L} \gamma^2_\ell \|u_\ell\|_{2,2}^2 < \varepsilon^2\right) = P\left(\sum_{\ell \leq L} \gamma^2_\ell \sum_{i=1}^{\infty} Z^2_{\ell,i} \xi^2_i < \varepsilon^2\right) = \int_{\sum_{\ell \leq L} \sum_{i \in \mathbb{N}} \xi^2_i} \prod_{\ell \leq L, i \in \mathbb{N}} \phi_{\xi_i}(z_{\ell,i}) dz_{\ell,i}
$$

$$
\geq \prod_{\ell=1}^{L} \frac{\gamma_\ell}{\gamma_L} P\left(\sum_{\ell \leq L} \sum_{i=1}^{\infty} \gamma^2_\ell Z^2_{\ell,i} \xi^2_i < \varepsilon^2\right) \geq \left(\frac{L!}{L^L}\right)^{\frac{1}{2}} P\left(\sum_{\ell=1}^{L} \|u_\ell\|_{2,2}^2 < \varepsilon^2\right) \geq e^{-L\kappa/2} \frac{1}{2}
$$

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

$$
P\left(\sum_{\ell \geq L} \gamma^2_\ell \|u_\ell\|_{2,2}^2 < \varepsilon^2\right) \geq 1 - \frac{1}{\varepsilon^2} \sum_{\ell \geq L} \mathbb{E}[\gamma^2_\ell \|u_\ell\|_{2,2}^2] \geq 1 - \frac{1}{\varepsilon^2} \int_{L}^{\infty} x^{-\kappa} dx = 1 - \frac{1}{(\kappa-1)L^{\kappa-1} \varepsilon^2} \geq \frac{1}{2}
$$

for $L$ large enough such that $(\kappa-1)L^{\kappa-1} \varepsilon^2 \geq 2$, i.e., $L \geq e^{-2/(\kappa-1)} \left(\frac{2}{\kappa-1}\right)^{\frac{1}{\kappa-1}} \geq e^{-2/(\kappa-1)} e^{1/(2\varepsilon)}$. Thus the best upper bound for $\varphi(\varepsilon) \lesssim L\kappa/2 \lesssim e^{-2/(\kappa-1)}$.

Next, we show the de-centering function $\inf_{h \in \mathbb{H} \setminus \lambda \mathbb{H}} \|h - \lambda \mathbb{H}\|_{2,2} \leq \|h\|_{2,2}^2 \leq \|\lambda\|_{2,s}^2 e^{-\kappa-2s}/s$ if $\lambda \in \ell^{2,s}(L^2(T))$ 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 $\lambda$ and square RHKS-norm satisfy

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

$$
\|\lambda^L\|_{2,2}^2 = \sum_{\ell=1}^{L} \gamma^{-2}_\ell \|\lambda_\ell\|_{2,2}^2 \leq \|\lambda\|_{2,s}^2 \max_{1 \leq \ell \leq L} \gamma^{-2}_\ell \ell^{-2s} \leq \|\lambda\|_{2,s}^2 \max_{1 \leq \ell \leq L} \ell^{-2s}
$$

Choosing the minimal integer $L$ such that $L \geq \|\lambda\|_{2,s}^{1/s} e^{-1/s}$ yields the result.
Finally, when the true parameter $\lambda_0 \in \ell^{2, s}(L^2(T))$, then we get the minimal solution to the rate equation $\varphi_{\lambda_0}(\varepsilon_n) \leq n\varepsilon_n^2$ by setting both $\varepsilon_n^{-2/(\kappa-1)} \leq n\varepsilon_n^2$ and $\varepsilon_n^{-(\kappa-2s)/s} \leq n\varepsilon_n^2$, which gives the rate of posterior contraction $n^{-\left(\frac{2}{\kappa-1}+\frac{s}{\kappa-2s}\right)/\kappa}$.

**Proposition 3.1.** Fit the spatiotemporal data $D = \{Z, Y\}$ with the model (32). Then given a new point $z_* = (x_*, t_*)$ we have

$$m(z_*)|D \sim N(m', C')$$

$$m' = c_1^T (C_M + K^{-1}C_{Y|M})^{-1}Y, \quad C' = C_{m_*} - c_s^T (C_M + K^{-1}C_{Y|M})^{-1}c_s$$

where we denote

$$\overline{Y} := \frac{1}{K} \sum_{k=1}^K \text{vec}(Y_k), \quad C_{m_*} := C_m(z_*, z_*), \quad c_s := C_m(Z, z_*), \quad c_1^T := C_m(z_*, Z)$$

**Proof of Proposition 3.1.** Compute using the following formula

$$p(m(z_*)|D) = \int p(m(z_*), M|D) dM = \int p(m(z_*)|M)p(M|D)dM \propto \int p(m(z_*)|M)p(D|M)dM = \int p(M, m(z_*))p(D|M)dM$$

Completing the square to integrate out $M$ and completing the square for $m(z_*)$ we have

$$m(z_*)|D \sim N(m', C')$$

$$C^{-1} = \frac{1}{C_{m_*}|M} - \frac{1}{C_{m_*}|M}c_s^T C_{M|m_*}^{-1}c_s C_{M|m_*}^{-1} - C_{M|m_*}^{-1}C_{Y|M}^{-1} K \overline{Y}$$

$$C_{m_*}|M := C_{m_*} - c_s^T C_{M|m_*}^{-1}c_s, \quad C_{M|m_*} := C_M - c_s C_{m_*}^{-1} c_s^T, \quad C_{post} = C_{M|m_*}^{-1} + K^2 C_{Y|M}^{-1}$$

By Sherman-Morrison-Woodbury formula, we further have

$$(C')^{-1} = C_{m_*}|M - C_{m_*}^{-1}c_s^T C_{M|m_*}^{-1}c_s C_{M|m_*}^{-1} - C_{M|m_*}^{-1}C_{Y|M}^{-1} K \overline{Y}$$

$$= C_{m_*}^{-1} - C_{m_*}^{-1} \left( C_M - c_s C_{m_*}^{-1} c_s^T \right)^{-1} c_s C_{m_*}^{-1}$$

$$- C_{m_*}^{-1} \left( C_M - c_s C_{m_*}^{-1} c_s^T \right)^{-1} c_s C_{m_*}^{-1}$$

$$= C_{m_*}^{-1} + C_{m_*}^{-1} \left( C_M - c_s C_{m_*}^{-1} c_s^T \right)^{-1} c_s C_{m_*}^{-1}$$

$$= C_{M|m_*}^{-1} C_{M|M}^{-1} \left[ (C_M + K^{-1}C_{Y|M}) - c_s C_{m_*}^{-1} c_s^T \right]^{-1} c_s C_{m_*}^{-1}$$

and

$$m' = C' C_{m_*}^{-1} c_s^T \left( C_{M|m_*}^{-1} C_{M|M}^{-1} \left[ (C_M + K^{-1}C_{Y|M}) - c_s C_{m_*}^{-1} c_s^T \right]^{-1} c_s C_{m_*}^{-1} \right) K \overline{Y}$$

$$= C' C_{m_*}^{-1} c_s^T \left( C_{M|M}^{-1} + K^{-1}C_{Y|M}^{-1} \right) \overline{Y}$$

$$= c_s^T \left[ C_{m_*}^{-1} \left( C_M + K^{-1}C_{Y|M}^{-1} \right) \right]^{-1} \overline{Y}$$

$$= c_s^T (C_M + K^{-1}C_{Y|M}^{-1})^{-1} \overline{Y}$$
B Posterior Inference

Discretize the spatial $X$ and time $T$ domains with $I$ and $J$ points respectively. Denote the observations on the discrete domain as $I \times J$ matrices $Y_k$ for $k = 1, \ldots, K$ trials, and thus $Y_{I \times J \times K} = \{Y_1, \ldots, Y_K\}$. We summarize model I as follows

$$
Y_k | M, \sigma_\varepsilon^2 \sim \mathcal{N}(M, \sigma_\varepsilon^2 I_x, I_t), \quad M_{I \times J} = m(X, t)
$$

$$
m(x, t) \sim \mathcal{GP}(0, C_{x|x'} \otimes C_t), \quad C_t(t, t') = \sigma_t^2 \exp(-0.5\|t - t'\|^2 / \rho_t^2)
$$

$$
C_x(x, x') = \sigma_x^2 \exp(-0.5\|x - x'|^2 / \rho_x^2), \quad C_{x|x'}(z, z') = \sum_{\ell=1}^{\infty} \lambda_\ell(t) \phi_\ell(z) \phi_\ell(z') C_t(t, t')
$$

$$
\lambda_\ell(t) = \gamma_\ell u_\ell(t), \quad u_\ell(\cdot) \sim \mathcal{GP}(0, \Sigma_u), \quad \Sigma_u(t, t') = \sigma_u^2 \exp(-0.5\|t - t'\|^2 / \rho_u^2)
$$

$$
\sigma_x^2 \sim \text{Gamma}(a_x, b_x), \quad \log \rho_x \sim \mathcal{N}(m_x, V_x), \quad * = \varepsilon, x, t, or u
$$

(B.1)

and model II in the following

$$
\text{vec}(Y_k) | M, C_{x|x'} \sim \mathcal{N}(\text{vec}(M), C_{x|x'}) , \quad M_{I \times J} = m(X, t), \quad C_{x|x'} = C_{x|x'}(X, X; t)
$$

$$
m(x, t) \sim \mathcal{GP}(0, I_x \otimes C_t), \quad C_t(t, t') = \sigma_t^2 \exp(-0.5\|t - t'\|^2 / \rho_t^2)
$$

$$
C_x(x, x') = \sigma_x^2 \exp(-0.5\|x - x'|^2 / \rho_x^2), \quad C_{x|x'}(x, x'; t) = \sum_{\ell=1}^{\infty} \lambda_\ell^2(t) \phi_\ell(x) \phi_\ell(x')
$$

(B.2)

$$
\lambda_\ell(t) = \gamma_\ell u_\ell(t), \quad u_\ell(\cdot) \sim \mathcal{GP}(0, \Sigma_u), \quad \Sigma_u(t, t') = \sigma_u^2 \exp(-0.5\|t - t'\|^2 / \rho_u^2)
$$

$$
\sigma_x^2 \sim \text{Gamma}(a_x, b_x), \quad \log \rho_x \sim \mathcal{N}(m_x, V_x), \quad * = \varepsilon, x, t, or u
$$

Truncate the kernel expansion (21) or (22) at some $L$ terms. We now focus on obtaining the posterior probability of $M_{I \times J}, \Lambda_{J \times L}$, $\sigma^2 := (\sigma_\varepsilon^2, \sigma_x^2, \sigma_t^2, \sigma_u^2)$ and $\rho := (\rho_x, \rho_t, \rho_u)$ in the models (B.1) (B.2). Transform the parameters $\eta := \log(\rho)$ for the convenience of calculation. Denote $\sigma_\varepsilon^2 = (\sigma_x^2, \sigma_t^2)$, and $\eta_\varepsilon^2 = (\eta_x^2, \eta_t^2)$. Denote $C_t = C_t(t, t)$, and $C_u = C_u(t, t)$. Let $C_x := C_x(X, X) = \Phi \Lambda_0^2 \Phi^T$ where $\Lambda_0 = \text{diag}(\{\lambda_\ell^2\})$. Then $C_{x|x'} := C_{x|x'}(X, X) = \Phi \text{diag}(\Lambda_j^2) \Phi^T$ where $\Lambda_j = \{\lambda_j^2\}$ is the $j$-th row of $\Lambda$. Denote $C_x(\sigma^2_x, \eta_\varepsilon) = \sigma_x^2 C_{o|x}(\eta_\varepsilon)$ where $* = x, t, z, or u$. Once the spatial eigen-basis $\Phi$ has been calculated, it will be shared across all the following calculation. Since only normalized eigen-basis $\Phi(\eta_\varepsilon)$ is used, we can set $\sigma_\varepsilon^2 \equiv 1$ and exclude it from posterior distributions.

Notice that $C_z = C_t \otimes C_{x|x'}(Z, Z)$ for model I is a full $IJ \times IJ$ matrix; while $C_{x|x'} = \text{diag}(\{C_{x|x'}^{\ell, j}\}_{j=1}^{J})$ for model II is a block diagonal matrix formed by $J$ blocks of $I \times I$ matrices. Both $C_x(\Lambda)$ and $C_{x|x'}(\Lambda)$ are defined through the Mercer’s expansions with fixed eigen-basis $\Phi$ and newly modeled eigenvalues $\Lambda$. We make some simplifications before proceeding.
the calculation of posteriors. Due to the linear independence requirement for \( \Phi \), we have \( L \leq I \). Therefore \( C_x \) is in general degenerate, and so is \( C_x^{j|t} \) if \( L < I \). To avoid inversion of degenerate matrices, we marginalize \( M \) out. In the following we calculate the posteriors for model I \([B.1]\) and model II \([B.2]\) respectively.

**B.1 Model I**

First, integrating out \( M \) in model I \([B.1]\) we have

\[
\log p(\Lambda, C^2, \eta|Y) = \log p(Y|\Lambda, \sigma^2, \sigma_t^2, \eta_x) + \log p(\Lambda|\sigma^2, \eta_u) + \sum_{* = \varepsilon, t, u} \log p(\sigma^2_*) + \sum_{* = \varepsilon, t, u} \log p(\eta_*)
\]

\[
= -\frac{IJ(K-1)}{2} \log \sigma^2_\varepsilon - \frac{1}{2} \log |C^*(\Lambda, \sigma^2_\varepsilon, \sigma_t^2, \eta_x)| - \frac{1}{2} \left[ K\sigma^2_\varepsilon (\overline{Y}^2 - \bar{Y}^2) + \bar{Y}^T (C^*)^{-1} \bar{Y} \right] - JI^T \log |\gamma(\eta_x)| - \frac{L}{2} \log |C_u(\sigma^2_\varepsilon, \eta_u)| - \frac{1}{2} \text{tr}(U^T C_u^{-1} U)
\]

\[
- \sum_{* = \varepsilon, t, u} (a_* + 1) \log \sigma^2_* + b_* \sigma_*^{-2} - \sum_{* = \varepsilon, t, u} \frac{1}{2} (\eta_* - m_*^2)/V_*
\]

where \( C^*(\Lambda, \sigma^2_\varepsilon, \sigma_t^2, \eta_x) := C_x(\Lambda, \sigma^2_\varepsilon, \sigma_t^2, \eta_x) + K^{-1}\sigma^2_\varepsilon I \), \( Y := \frac{1}{K} \sum_{k=1}^K \text{vec}(Y_k) \), \( \bar{Y}^2 := \bar{Y}^T \bar{Y} \), and \( \overline{Y}^2 := \frac{1}{K} \sum_{k=1}^K \text{tr}(Y_k^T Y_k) \).

\((\sigma^2)\). For \( * = \varepsilon \) or \( t \), we sample \( \sigma^2_* \) using the slice sampler \([\text{Neal} 2003]\), which only requires log-posterior density and works well for scalar parameters,

\[
\log p(\sigma^2_\varepsilon|\cdot) = -\frac{IJ(K-1)}{2} \log \sigma^2_\varepsilon - \frac{1}{2} \log |C^*(\Lambda, \sigma^2_\varepsilon, \sigma_t^2, \eta_x)| - \frac{1}{2} \left[ K\sigma^2_\varepsilon (\overline{Y}^2 - \bar{Y}^2) + \bar{Y}^T (C^*)^{-1} \bar{Y} \right] - (a_\varepsilon + 1) \log \sigma^2_\varepsilon - b_\varepsilon \sigma^2_\varepsilon
\]

\[
\log p(\sigma^2_t|\cdot) = -\frac{1}{2} \log |C^*(\Lambda, \sigma^2_\varepsilon, \sigma_t^2, \eta_x)| - \frac{1}{2} \bar{Y}^T (C^*)^{-1} \bar{Y} - (a_t + 1) \log \sigma^2_t - b_t \sigma^2_t
\]

Note the prior for \( \sigma^2_u \) is conditionally conjugate:

\[
\sigma^2_u|\cdot \sim \Gamma^{-1}(a'_u, b'_u), \quad a'_u = a_u + \frac{1}{2} JI, \quad b'_u = b_u + \frac{1}{2} \text{tr}(U^T C_u^{-1} U)
\]

\((\eta)\). Given \( * = x, t, \) or \( u \), we could sample \( \eta_* \) using the slice sampler \([\text{Neal} 2003]\), which only requires log-posterior density,

\[
\log p(\eta_x|\cdot) = -\frac{1}{2} \log |C^*(\Lambda, \sigma^2_\varepsilon, \sigma_t^2, \eta_x)| - \frac{1}{2} \bar{Y}^T (C^*)^{-1} \bar{Y} - JI^T \log |\gamma(\eta_x)| - \frac{1}{2} (\eta_x - m_x)^2/V_x
\]

\[
\log p(\eta_t|\cdot) = -\frac{1}{2} \log |C^*(\Lambda, \sigma^2_\varepsilon, \sigma_t^2, \eta_x)| - \frac{1}{2} \bar{Y}^T (C^*)^{-1} \bar{Y} - \frac{1}{2} (\eta_t - m_t)^2/V_t
\]

\[
\log p(\eta_u|\cdot) = -\frac{L}{2} \log |C_{0u}(\eta_u)| - \frac{1}{2} \text{tr}(U^T C_{0u}^{-1} U) \sigma_u^{-2} - \frac{1}{2} (\eta_u - m_u)^2/V_u
\]
(Λ). Using a similar argument by the matrix normal prior for Λ, we have \( \Lambda|\sigma^2_u, \eta_u \sim MN(0, C_u(\sigma^2_t, \eta_t), \text{diag}(\gamma^2)) \). Therefore, we could use the elliptic slice sampler (ESS, Murray et al. [2010]), which only requires the log-likelihood
\[
\log p(\Lambda; Y) = -\frac{1}{2} \log |C^*(\Lambda, \sigma^2_t, \sigma^2_t, \eta_u)| - \frac{1}{2} Y^T (C^*)^{-1} Y
\]

(M)*. By the definition of STGP prior, we have \( \text{vec}(M)|\Lambda, \sigma^2_t, \eta_u \sim N_{IJ}(0, C_z(\Lambda, \sigma^2_t, \eta_u)) \). On the other hand, one can write the log-likelihood function as
\[
-\frac{\sigma^2}{2} \sum_{k=1}^K \text{tr}((Y_k - M)(Y_k - M)) = -\frac{1}{2} \sum_{k=1}^K (\text{vec}(Y_k) - \text{vec}(M))^T (\sigma^2_k I)^{-1} (\text{vec}(Y_k) - \text{vec}(M))
\]
Therefore we have the analytic posterior
\[
\text{vec}(M)| \sim N_{ND}(M', C'), \quad M' = C'\sigma^2 \sum_{k=1}^K \text{vec}(Y_k) = C'K\sigma^2 Y,
\]
\[
C' = (C_z^{-1} + K\sigma^2 I)^{-1} = C_z(C^*)^{-1}K^{-1}\sigma^2 I
\]

B.2 Model II

Now we integrate out \( M \) in model II (B.2) to get
\[
\log p(\Lambda, C^2, \eta|Y) = \log p(Y|C_{x|t}(\Lambda, \eta_x), \sigma^2_t, \eta_t) + \log p(\Lambda|\sigma^2_u, \eta_u) + \sum_{s=t,u} \log p(\sigma^2_s) + \sum_{s=x,t,u} \log p(\eta_s)
\]
\[
= -\frac{K-1}{2} \log |C_{x|t}(\Lambda, \eta_x)| - \frac{1}{2} \log |C^*(\Lambda, \sigma^2_t, \eta_u)| - \frac{1}{2} \left[ \sum_{k=1}^K \text{vec}^*(Y_k)^T C_{x|t}^{-1} \text{vec}^*(Y_k) + Y^T (C^*)^{-1} Y \right]
\]
\[
- JL^T \log |\gamma(\eta_x)| - \frac{L}{2} \log |C_u(\sigma^2_u, \eta_u)| - \frac{1}{2} \text{tr}(U^T C_u^{-1} U)
\]
\[
- \sum_{s=t,u} (a_s + 1) \log \sigma^2_s + b_s \sigma^2_s - \sum_{s=x,t,u} \frac{1}{2} (\eta_s - m_s)^2/V_s
\]
where \( C^*(\Lambda, \sigma^2_t, \eta_u) := C_t(\sigma^2_t, \eta_t) \otimes I_x + K^{-1} C_{x|t}(\Lambda, \eta_x), \) and \( \text{vec}^*(Y_k) := \text{vec}(Y_k) - Y. (\sigma^2) \). We could sample \( \sigma^2_t \) using the slice sampler (Neal [2003]), which only requires log-posterior density and works well for scalar parameters,
\[
\log p(\sigma^2_t|\cdot) = -\frac{1}{2} \log |C^*(\Lambda, \sigma^2_t, \eta_u)| - \frac{1}{2} Y^T (C^*)^{-1} Y - (a_t + 1) \log \sigma^2_t - b_t \sigma^2_t
\]
Note the prior for \( \sigma^2_u \) is conditionally conjugate:
\[
\sigma^2_u| \sim \Gamma^{-1}(a'_u, b'_u), \quad a'_u = a_u + \frac{1}{2} JL, \quad b'_u = b_u + \frac{1}{2} \text{tr}(U^T C_u^{-1} U)
\]
(\(\eta\)). Given \(* = x, t,\) or \(u,\) we could sample \(\eta_x\) using the slice sampler (Neal, 2003), which only requires log-posterior density and works well for scalar parameters,

\[
\log p(\eta_x|::) = -\frac{K}{2} - \frac{1}{2} \log |C_{x|t}(\Lambda, \eta_x)| - \frac{1}{2} \log |C^*(\Lambda, \sigma_t^2, \eta_x)| \\
- \frac{1}{2} \left[ \sum_{k=1}^K \text{vec}^*(Y_k)^T C^{-1}_{x|t} \text{vec}^*(Y_k) + \bar{Y}^T (C^*)^{-1} \bar{Y} \right] - J1^T \log |\gamma(\eta_x)| - \frac{1}{2}(\eta_x - m_x)^2/V_x
\]

\[
\log p(\eta_t|::) = -\frac{1}{2} \log |C^*(\Lambda, \sigma_t^2, \eta_t)| - \frac{1}{2} \bar{Y}^T (C^*)^{-1} \bar{Y} - \frac{1}{2}(\eta_t - m_t)^2/V_t
\]

\[
\log p(\eta_u|::) = -\frac{L}{2} \log |C_{0u}(\eta_u)| - \frac{1}{2} \text{tr}(U^T C_{0u}^{-1} U) \sigma_u^{-2} - \frac{1}{2}(\eta_u - m_u)^2/V_u
\]

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

\[
\log p(\Lambda; Y) = -\frac{K}{2} - \frac{1}{2} \log |C_{x|t}(\Lambda, \eta_x)| - \frac{1}{2} \log |C^*(\Lambda, \sigma_t^2, \eta_x)| \\
- \frac{1}{2} \left[ \sum_{k=1}^K \text{vec}^*(Y_k)^T C^{-1}_{x|t} \text{vec}^*(Y_k) + \bar{Y}^T (C^*)^{-1} \bar{Y} \right]
\]

(M)*. By the definition of STGP prior, we have \(\text{vec}(M)|\sigma_t^2, \eta_t \sim N_{I,J}(0, C_t(\sigma_t^2, \eta_t) \otimes I_x).\) On the other hand, one can write the log-likelihood function as

\[
-\frac{1}{2} \sum_{k=1}^K \text{vec}(Y_k - M)^T C^{-1}_{x|t} \text{vec}(Y_k - M) = -\frac{1}{2} \sum_{k=1}^K (\text{vec}(Y_k) - \text{vec}(M))^T C^{-1}_{x|t} (\text{vec}(Y_k) - \text{vec}(M))
\]

Therefore we have the analytic posterior

\[
\text{vec}(M)|\sim N_{ND}(M', \Sigma), \quad M' = C' C^{-1}_{x|t} \sum_{k=1}^K \text{vec}(Y_k) = C' C^{-1}_{x|t} K \bar{Y},
\]

\[
C' = \left( C_t^{-1} \otimes I_x + K C^{-1}_{x|t} \right)^{-1} = (C_t \otimes I_x) (C^*)^{-1} K^{-1} C_{x|t}
\]

B.3 Computational Advantage of Model II

The most intensive computation as above involves the inverse and determinant of the posterior covariance kernel \(C^*\) for two models:

\[
C^*_t := C_t(\Lambda, \sigma_t^2, \eta_x) + K^{-1} \sigma_z^2 I, \quad C^*_t := C_t(\sigma_t^2, \eta_t) \otimes I_x + K^{-1} C_{x|t}(\Lambda, \eta_x)
\]

Their structure dictates different amount of computation required. Actually, we can show that the kernel of model II, \(C^*_t\), has computational advantage over that for model I.
where we have 

\[ \begin{align*}
\{ \Phi \} \text{diag}(\Lambda_j^2) = (I_t \otimes \Phi) \text{diag}(\text{vec}^T(\Lambda^2))(I_t \otimes \Phi^T)
\end{align*} \]

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

\[ \begin{align*}
(C'_t)^{-1} &= (C^{-1}_t \otimes I_x) - (C^{-1}_t \otimes I_x)(I_t \otimes \Phi)[K \text{diag}(\text{vec}^T(\Lambda^{-2})) + (I_t \otimes \Phi^T)(C^{-1}_t \otimes I_x)(I_t \otimes \Phi)]^{-1} \\
& = (C^{-1}_t \otimes I_x) - (C^{-1}_t \otimes \Phi)[K \text{diag}(\text{vec}^T(\Lambda^{-2})) + (C^{-1}_t \otimes I_L)]^{-1}(C^{-1}_t \otimes \Phi^T) \\
& = (C^{-1}_t \otimes I_x) - (C^{-1}_t \otimes \Phi)K^{-1} \text{diag}(\text{vec}^T(\Lambda^2))[K^{-1} \text{diag}(\text{vec}^T(\Lambda^2)) + (C_t \otimes I_L)]^{-1}(I_t \otimes \Phi^T) \\
& = C^{-1}_t \otimes (I_x - \Phi\Phi^T) + (I_t \otimes \Phi)[K^{-1} \text{diag}(\text{vec}^T(\Lambda^2)) + (C_t \otimes I_L)]^{-1}(I_t \otimes \Phi^T)
\end{align*} \]

Similarly we have

\[ \begin{align*}
C'_t = C_t \otimes (I_x - \Phi\Phi^T) + (I_t \otimes \Phi)[K \text{diag}(\text{vec}^T(\Lambda^{-2})) + (C^{-1}_t \otimes I_L)]^{-1}(I_t \otimes \Phi^T) \\
(C'_t)^{\frac{1}{2}} &= C^{-\frac{1}{2}}_t \otimes (I_x - \Phi\Phi^T) + (I_t \otimes \Phi)[K \text{diag}(\text{vec}^T(\Lambda^{-2})) + (C^{-1}_t \otimes I_L)]^{-\frac{1}{2}}(I_t \otimes \Phi^T)
\end{align*} \]

where we use the following calculation that is numerically more stable

\[ [K \text{diag}(\text{vec}^T(\Lambda^{-2})) + (C^{-1}_t \otimes I_L)]^{-\frac{1}{2}} = [K^{-1} \text{diag}(\text{vec}^T(\Lambda^2))(K^{-1} \text{diag}(\text{vec}^T(\Lambda^2)) + (C_t \otimes I_L)]^{-1}(C_t \otimes I_L)]^{\frac{1}{2}} \]

Based on the matrix determinant lemma we can calculate

\[ \begin{align*}
\det(C'_t) &= \det(K \text{diag}(\text{vec}^T(\Lambda^{-2})) + (I_t \otimes \Phi^T)(C^{-1}_t \otimes I_x)(I_t \otimes \Phi)) \\
& = \det(K^{-1} \text{diag}(\text{vec}^T(\Lambda^2))) \det(C_t \otimes I_x) \\
& = \det(\text{diag}(\text{vec}^T(\Lambda^{-2})) + K^{-1}(C^{-1}_t \otimes I_L)) \prod_{j, \ell} \lambda^{2}_{j, \ell} \det(C_t) \\
& = \det(C_t \otimes I_L + K^{-1} \text{diag}(\text{vec}^T(\Lambda^2))) \det(C_t)^{I-L}
\end{align*} \]

However in model I, we note that \( C_z = C_{x|t}^{\frac{1}{2}} C_{x|t'}^{\frac{1}{2}} \circ (C_t \otimes 1_{I \times I}), \) where \( \circ \) is element-wise multiplication, and \( 1_{I \times I} \) is an \( I \times I \) matrix with all elements 1. According to [22], we have

\[ C_{x|t}^{\frac{1}{2}} C_{x|t'}^{\frac{1}{2}} = [\Phi \text{diag}(\Lambda_j) \text{diag}(\Lambda_{j'}) \Phi^T] = (I_t \otimes \Phi) \text{vec}\{\text{diag}(\Lambda_j)\} \text{vec}^T\{\text{diag}(\Lambda_{j'})\}(I_t \otimes \Phi^T) \]

where \text{vec}\{\cdot\} and \text{vec}^T\{\cdot\} are column/row wise vectorization of block matrices. Applying the above equation to the inverse or determinant of \( C_t^* \) does not simplify computation in general.
Figure 13: 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.

C More Results of Longitudinal Analysis of Brain Images

Figure 13 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 14: Estimated variance of the brain images for CN (top row), MCI (middle row) and AD (bottom row) respectively.

In Section 5.3 we summarize the correlation between the brain ROI and POI. In fact, we have more results regarding TESD presented in different forms. Figure 14 shows the estimated variances of the brain images as functions of time. They are all small across
Figure 15: Extended correlation between the brain region of interest and a selected point of interest from the coarse mesh (upper) to the fine mesh (lower) in each of MCI and AD groups.

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.

Figure 15 include more results about extending TESD to new locations in MCI and AD groups. They all illustrate the benefit of a fully nonparametric approach in modeling TESD in the spatiotemporal data.