Random Surface Covariance Estimation by Shifted Partial Tracing

Tomas Masak and Victor M. Panaretos

Institut de Mathématiques, École polytechnique fédérale de Lausanne, Lausanne, Switzerland

ABSTRACT

The problem of covariance estimation for replicated surface-valued processes is examined from the functional data analysis perspective. Considerations of statistical and computational efficiency often compel the use of separability of the covariance, even though the assumption may fail in practice. We consider a setting where the covariance structure may fail to be separable locally—either due to noise contamination or due to the presence of a nonseparable short-range dependent signal component. That is, the covariance is an additive perturbation of a separable component by a nonseparable but banded component. We introduce nonparametric estimators hinging on the novel concept of shifted partial tracing, enabling computationally efficient estimation of the model under dense observation. Due to the denoising properties of shifted partial tracing, our methods are shown to yield consistent estimators even under noisy discrete observation, without the need for smoothing. Further to deriving the convergence rates and limit theorems, we also show that the implementation of our estimators, including prediction, comes at no computational overhead relative to a separable model. Finally, we demonstrate empirical performance and computational feasibility of our methods in an extensive simulation study and on a real dataset. Supplementary materials for this article are available online.

1. Introduction

Functional Data Analysis (FDA, Ramsay and Silverman 2002; Hsing and Eubank 2015) focuses on the problem of statistical inference on the law of a random process \( X(u) : [0, 1]^D \rightarrow \mathbb{R} \) given multiple realizations thereof. The process realizations are treated as elements of a separable Hilbert space \( \mathcal{H} \) of functions on \([0, 1]^D\) (e.g., \( L^2([0, 1]^D) \)). FDA covers the full gamut of statistical tasks, including regression, classification, and testing, to name a few. In any of these problems, the covariance operator \( C : \mathcal{H} \to \mathcal{H} \) of the random function \( X(u) \) is elemental. This trace-class integral operator with kernel \( c(u_1, u_2) = \text{cov}(X(u_1), X(u_2)) \), encodes the second-order characteristics of \( X(u) \) and its associated spectral decomposition is at the core of many (or even most) FDA inferential methods. Consequently, the efficient estimation of the covariance operator \( C \) (or equivalently its kernel \( c \)) associated with \( X \) is a fundamental task in FDA, on which further methodology can be based. This is to be done on the basis of \( N \) iid realizations of the random process \( X \), say \( \{ X_1, \ldots, X_N \} \). One wishes to do so nonparametrically, since the availability of replicated realizations should allow so. When \( D = 1 \), it is fair to say that this is entirely feasible and well understood, under a broad range of observation regimes (see Wang, Chiou, and Müller 2016 for a comprehensive overview).

Though conceptually similar, things are much less straightforward in the case of random surfaces, that is, when \( D = 2 \), which is the case we focus on in this article. In this case, one faces additional challenging limitations to statistical and computational efficiency when attempting to nonparametrically estimate \( c : [0, 1]^4 \to \mathbb{R} \) on the basis of \( N \) replications (see Aston, Pigoli, and Tavakoli 2017, sec. 1 for a detailed discussion). The number of grid points on which \( c \) is measured may even exceed \( N \), especially in densely observed functional data scenarios. Worse still, one may not be able to even store the empirical covariance, much less invert it. To appreciate this, assume that each of the \( N \) iid surfaces \( \{ X_n(s,t) \} \) are measured on a common grid of size \( K_1 \times K_2 \) over \([0, 1]^2\). That is, the data corresponding to a single realization \( X_n \) form a matrix \( X_n \in \mathbb{R}^{K_1 \times K_2} \) and the raw empirical covariance is represented by the tensor \( C \in \mathbb{R}^{K_1 \times K_2 \times K_1 \times K_2} \), which is a discretization of the empirical covariance kernel. If we assume \( K_1 = K_2 =: K \), the covariance tensor \( C \) requires \( O(NK^4) \) operations to be estimated and \( O(K^4) \) memory to be stored. This becomes barely feasible on a regular computer with \( K \) as small as 100. Moreover, as Aston, Pigoli, and Tavakoli (2017) note, the statistical constraints stemming from the need to accurately estimate \( O(K^4) \) parameters contained in \( C \) from only \( NK^2 \) measurements are usually even tighter than the computational constraints.

This dimensionality challenge is often handled with by imposing additional structure, for example, stationarity or separability Genton (2007), Gneiting (2002), and Gneiting, Genton, and Guttorp (2006). Either assumption reduces the four-dimensional nonparametric estimation problem into a two-dimensional one. In the case of a \( K \times K \) grid, this reduces the number of parameters from \( O(K^4) \), to \( O(K^2) \). Moreover, both estimation and subsequent manipulation (for example
inversion as required in prediction) of the covariance become computationally much simpler, owing to some explicit formulas in case of separability and to the fast Fourier transform in case of stationarity.

Though such assumptions substantially reduce the dimensionality of the problem, the imposed simplicity and structural restrictions are often quite questionable. Stationarity appears overly restrictive when replicated data are available, and indeed is seldom used for functional data. Separability is imposed much more often, despite having shortcomings of its own. A thorough discussion of the implications that separability entails is provided in Rougier (2017). In summary, separable covariances fail to model any space-time interactions whatsoever. Indeed, in recent years, several tests for separability of space-time functional data have been developed and used to demonstrate that in many datasets previously modeled as separable, the separability assumption is distinctly violated. Aston, Pigoli, and Tavakoli (2017), Bagchi and Dette (2020), and Constantinou, Kokoszka, and Reimherr (2017).

1.1. Our Contributions

We propose a more flexible framework than that offered by separability, which allows for mild (nonparametric) deviations from separability while retaining all the computational and statistical advantages that separability offers. In particular, we consider a framework where the target covariance is an additive perturbation of a separable covariance,

\[ c(t, s, t', s') = a(t, s, t', s') + b(t, s, t', s') \]  

(1)

where \( a(t, s, t', s') = a_1(t, t')a_2(s, s') \) is separable, and \( b(t, s, t', s') \) is banded, that is, supported on \( \{ \max(|t - t'|,|s - s'|) \leq \delta \} \) for some \( \delta > 0 \). Combining the two components results in a nonparametric family of models, which is much richer than the separable class. In particular, the model represents a strict generalization of separability, reducing to a separable model when \( \delta = 0 \). Intuitively, it postulates that while the global (long-range) characteristics of the process are expected to be separable, there may also be local (short-range) characteristics of the process that may be nonseparable. For some practical problems, separability might possibly fail due to some interactions between time and space, which however, do not propagate globally. These may be due to (weakly dependent) noise contamination, which can lead to local violations of separability, perturbing the covariance near its diagonal. It could also, however, be due to the presence of signal components that are nonseparable and yet weakly dependent.

Heuristically, if we were able to deconvolve the terms \( a \) and \( b \), then the term \( a \) would be easily estimable on the basis of dense observations, exploiting separability. We demonstrate that it actually is possible to access a nonparametric estimator of \( a \)—without needing to manipulate or even store the empirical covariance—by means of a novel device, which we call shifted partial tracing. This linear operation mimics the partial trace Astron, Pigoli, and Tavakoli (2017), but it is suitably modified to allow us to separate the terms \( a \) and \( b \) in (1). Exploiting this device, we produce a linear estimator of \( a \) (linear up to scaling, to be precise) that can be computed efficiently, with no computational overhead relative to assuming separability. It is shown to be consistent, with explicit convergence rates, when the processes are observed discretely on a grid, possibly corrupted with measurement error.

The bandwidth \( \delta > 0 \) is assumed constant and non-decreasing in the sample size \( N \) or the grid size \( K \). Consequently, even though \( b \) is banded, it has the same order of entries as \( c \) itself, when observed on a grid. Hence, if \( b \) is also an estimand of interest, and statistical and computational efficiency is sought, an additional structural assumption on \( b \) is needed to prevent \( b \) from being much more complicated to handle than \( a \). We focus on stationarity as a specific assumption, which seems broadly applicable, is interesting from the computational perspective, and yields a form of parsimony complementary to separability. Under this additional assumption, we show in detail that both \( a \) and \( b \) of model (1) can be estimated efficiently, and the estimator can be both applied and inverted (numerically), while the computational costs of these operations do not exceed their respective costs in the separable regime. Specifically, we show that all of these operations, that is, estimation, application, and inversion of the covariance, can be performed at the same cost as matrix-matrix multiplication between pairs of the sampled observations.

Our methodology is also capable of estimating a separable model under the presence of heteroscedastic noise. When observed on a grid, this leads to a separable covariance superposed with a diagonal structure, which has again the same order of degrees of freedom as the separable part. A heteroscedastic noise may very well arise from a discretization of a random process, which is weakly dependent and potentially even smooth at a finer resolution. In their seminal book Ramsay and Silverman (2005), Ramsay and Silverman state: “the functional variation that we choose to ignore is itself probably smooth at a finer scale of resolution.” In other words, with increasing grid size \( K \), a diagonal structure may become a banded structure. One can thus view our methodology as being able to estimate a separable model observed under heteroscedastic and/or weakly dependent noise. If the degrees of freedom belonging to the noise do not exceed the degrees of freedom of the separable part, we can use the noise structure for example, for the purposes of prediction with no computational overhead compared to the separable model.

Regardless of whether one views \( b \) as an estimand of interest or as a nuisance, the key point of this article is that the methodology we advocate, and label shifted partial tracing, can be used to estimate the separable part of model (1), provided data are densely observed.

1.2. Notation

A real separable Hilbert space \( H \) is equipped with an inner product \( \langle \cdot, \cdot \rangle \) and the induced norm \( \| \cdot \| \). The Banach space of operators on \( H \) is denoted by \( S_\infty(H) \). Using the tensor product notation, we write the SVD of \( F \in S_\infty(H) \) as \( F = \sum_{j=1}^\infty \sigma_j f_j \otimes f_j \). For \( p \geq 1 \), \( F \in S_p(H) \) belongs to \( S_p(H) \) if \( \| F \|_p := \left( \sum_{j=1}^\infty \sigma^p_j \right)^{1/p} < \infty \). When equipped with the norm \( \| \cdot \|_p \), \( S_p(H) \) is a Banach space. For \( F \in S_1(H) \), we define its trace as \( \text{Tr}(F) = \sum_{j=1}^\infty \langle f_j, f_j \rangle \), where the choice of \( \{ f_j \}_{j=1}^\infty \) is immaterial.
The tensor product of two Banach spaces $B_1$ and $B_2$, denoted by $B_1 \otimes B_2$, is the completion of the set \[ \left\{ \sum_{j=1}^{N} x_j \otimes y_j \mid x_j \in \mathcal{H}_1, y_j \in \mathcal{H}_2, N \in \mathbb{N} \right\} , \] see Weidmann (2012). We have the isometric isomorphism $S_p(H_1) \otimes S_p(H_2) \simeq S_p(H_1 \otimes H_2)$.

For $A_1 \in S_p(H_1)$ and $A_2 \in S_p(H_2)$, we define $A := A_1 \widetilde{\otimes} A_2$ as the unique operator on $S_p(H_1) \otimes S_p(H_2)$ satisfying $(A_1 \otimes A_2)(x \otimes y) = A_1 x \otimes A_2 y, \forall x \in \mathcal{H}_1, y \in \mathcal{H}_2$. By the construction above, we have $\|A_1 \otimes A_2\| = \|A_1\| \|A_2\|$.

For a random element $X$ on $\mathcal{H}$ with $\mathbb{E}|X|^2 < \infty$, we denote the mean $m = \mathbb{E}X$ and the covariance $\text{Cov}(X) = \mathbb{E}(X - m)(X - m)^\ast$ (see Hsing and Eubank 2015). Covariances are trace-class, and positive semidefinite, that is, $C \in S^+_1(L^2[0,1])$. When $H = L^2[0,1]$, the covariance operator is related to the covariance kernel $c = c(t,s)$ via $(Cf)(t) = \int_0^1 c(t,s)f(s)ds$.

We use capital letters to denote operators such as, for example, $C \in S_1(L^2[0,1])$, lower-case letters to denote their kernels such as $c \in L^2[0,1]^2$, and bold-face letters to denote discrete objects such as vectors or matrices, for example, $C(t) \in \mathbb{R}^{K_1 \times K_2}$ when $c$ is measured discretely on a $K_1 \times K_2$ grid in $[0,1]^2$. When $H = H_1 \otimes H_2$, we often think of the first dimension as time, denoted by the variable $t$, and the second dimension as space, denoted by the variable $s$. See Section A in the supplementary materials for a more detailed exposition of the notation and background concepts.

### 2. Methodology

#### 2.1. Separable-plus-Banded Covariance

We begin with a formal definition of separability and bandedness.

**Definition 1.** For $\mathcal{H} := H_1 \otimes H_2, A \in S_p(\mathcal{H})$ is called separable if $A = A_1 \otimes A_2$ for some $A_1 \in S_p(H_1)$ and $A_2 \in S_p(H_2)$. For $H = L^2[0,1]^2$, $B \in S_2(\mathcal{H})$ with kernel $b = b(t,s,t',s')$ is banded by $\delta \in [0,1)$ if $b(t,s,t',s') = 0$ almost everywhere on the set \{ $(t,s,t',s') \in [0,1]^4 | |t - t'|, |s - s'| \leq \delta$ \}. If $H = L^2[0,1]^2$, an operator $A \in S_2(\mathcal{H})$ is separable if and only if its kernel factorizes as $a(t,s,t',s') = a_1(t,t')a_2(s,s')$ almost everywhere for some marginal kernels $a_1$ and $a_2$.

We postulate the following model for the covariance of a random element $X \in L^2[0,1]^2$:

\begin{equation}
C = A_1 \otimes A_2 + B, \tag{2}
\end{equation}

where $A_1, A_2 \in S^+_1(L^2[0,1])$ and $B \in S^+_1(L^2[0,1]^2)$ is banded by $\delta \in [0,1)$. On the level of kernels, this implies for almost all $t,s,t',s' \in [0,1]$ the decomposition

\begin{equation}
c(t,s,t',s') = a_1(t,t')a_2(s,s') + b(t,s,t',s'). \tag{3}
\end{equation}

The covariance structure (2) can arise for example when $X$ is a superposition of two uncorrelated processes $Y$ and $W$, that is, $X(t,s) = Y(t,s) + W(t,s), t,s \in [0,1]$, such that the covariance of $Y$ is separable and the covariance of $W$ is banded (e.g., $W$ is a moving average process with a compactly supported window-width). Note that by choosing $\delta = 0$ (leading to $B \equiv 0$), model (2) contains separability as a submodel.

#### 2.2. Shifted Partial Tracing

Estimation of model (2) will be possible using the following device. For the sake of presentation, we assume continuity and positive semidefiniteness here. A more general (but technical and less intuitive) development, is given in the supplementary materials. The discrete case, where the integrals are replaced by sums in the usual way, will be treated in Section 3.

**Definition 2.** Let $C \in S^+_1(L^2[0,1])$ have a continuous kernel $c = c(t,s,t',s')$. Let $\delta \in [0,1)$. We define the $\delta$-shifted trace of $C$ as

\[
\text{Tr}^\delta(C) := \int_0^{1-\delta} \int_0^{1-\delta} c(t,s,t + \delta,s + \delta)ds \, dt.
\]

We also define the $\delta$-shifted partial traces of $C$, denoted $\text{Tr}^\delta_1(C)$ and $\text{Tr}^\delta_2(C)$, as the integral operators with kernels given, respectively, by

\[
c_1(t,t') := \int_0^{1-\delta} c(t,s,t',s + \delta)ds \quad \text{and} \quad c_2(s,s') := \int_0^{1-\delta} c(t,s,t + \delta,s')dt.
\]

In the special case of $\delta = 0$ the definition of shifted trace corresponds to the standard (nonshifted) trace of a trace-class operator with a continuous kernel. Also, for $\delta = 0$, $\delta$-shifted partial tracing corresponds to partial tracing as defined in Aston, Pigoli, and Tavakoli (2017). Finally, we show in Section B in the supplementary materials that shifted (partial) tracing is by definition a bounded linear functional (resp. operator).

#### 2.3. Estimation

We assume throughout the article the availability of $N$ independent (and w.l.o.g. zero-mean) surfaces, say $X_1, \ldots, X_N$, with covariance given by (2), where $\delta$ is such that $\text{Tr}^\delta(A_1)$ and $\text{Tr}^\delta(A_2)$ are nonzero. The following properties of shifted partial tracing are paramount for estimation of the separable-plus-banded model (2). The proofs are postponed to the supplementary materials.

**Proposition 1.** Let $A_1, A_2 \in S^+_1(L^2[0,1])$ have continuous kernels and $F = A_1 \otimes A_2$. Then $\text{Tr}^\delta_1(F) = \text{Tr}^\delta_1(A_1) \text{Tr}^\delta_2(A_2)$, and $\text{Tr}^\delta_2(F) = \text{Tr}^\delta_1(F) \otimes \text{Tr}^\delta_2(F)$.

**Lemma 1.** Let $B \in S^+_1(L^2[0,1]^2)$ have a continuous kernel banded by $\delta^*$. Then for any $\delta > \delta^*$ we have $\text{Tr}^\delta_1(B) = \text{Tr}^\delta_2(B) = 0$.

Therefore, shifted partial tracing works around the banded part of the process to enable a direct estimation of the separable part of the covariance.

**Example 1.** Assume we have a single continuous observation $X \in L^2[0,1]^2$ with a covariance separable $C = C_1 \otimes C_2$ and
a continuous kernel $c(t, s, t', s') = c_1(t, t')c_2(s, s')$. Assume for simplicity that $\text{Tr}(C_1) = \text{Tr}(C_2) = 1$. Partial tracing (without shifting, that is, $\delta = 0$) can be used to estimate $C_1$ and $C_2$ in the following way.

The observation $X$ is cut along the temporal axis to form a spatial sample $\{X(s)\}_{s \in [0, 1]}$, that is, any given time point $t$ provides a single curve $X(s), s \in [0, 1]$. This spatial sample is used to estimate the spatial covariance $C_2$ in a standard way, that is, outer products $X^t \otimes X^t$ are formed and averaged together as

$$\widehat{C}_2 = \int_0^1 X^t \otimes X^t \, dt \text{ or } \widehat{c}_2(s, s') = \int_0^1 X^t(s)X^t(s') \, dt.$$  

This is a moment estimator in a sense, since $E(X^t \otimes X^t) = C_2$ for any $t \in [0, 1]$. Similarly for the temporal domain.

When the covariance is instead separable-plus-banded, that is, $C = A_1 \otimes A_2 + B$ with $B$ banded by $\delta$, it is no longer true that $E(X^t \otimes X^t) = A_2$, but it is still true that $E(X^{t+\delta} \otimes X^{t+\delta}) \propto A_2$ for all $t \in [0, 1 - \delta]$. Hence, instead of taking outer products of $X^t$ with itself, we can form outer products $X^t \otimes X^{t+\delta}$ and average over these for $t \in [0, 1 - \delta]$ to obtain a scaled estimator of $A_2$. See Figure 1 for a visualization.

Using the previous lemma together with the last equality in Proposition 1, we obtain the estimating equation for model (2):

$$\text{Tr}^\delta(C)A_1 \otimes A_2 = \text{Tr}^\delta_1(C) \otimes \text{Tr}^\delta_2(C).$$  

Equation (5) suggests the following estimators for the separable part of the model:

$$\widehat{A}_1 = \text{Tr}^\delta_1(\widehat{C}_N) \quad \text{and} \quad \widehat{A}_2 = \frac{\text{Tr}^\delta_2(\widehat{C}_N)}{\text{Tr}^\delta(\widehat{C}_N)},$$  

where $\widehat{C}_N = \frac{1}{N} \sum_{n=1}^N (X_n - \bar{X}_N) \otimes (X_n - \bar{X}_N)$ is the empirical estimator of $C$. Of course, we need to assume $\text{Tr}^\delta(\widehat{C}_N) \neq 0$. Once the separable part of the model has been estimated, we can define

$$\widehat{B} = \widehat{C}_N - \widehat{A}_1 \otimes \widehat{A}_2.$$  

Optionally, we can set the kernel of $\widehat{B}$ to zero out size of the band of size $\delta$. Note that none of the estimators defined above is guaranteed to be symmetric or positive semidefinite. However, this is just a technicality, which can be handled with easily, see Section E in the supplementary materials.

The banded part $B$ is stationary if its kernel $b$ is translation invariant: $b(t, s, t', s') = \xi(|t - t'|, |s - s'|), t, t', s, s' \in [0, 1]$, where $\xi \in L^2[0, 1]^2$ is the symbol of $B$. If we add the stationarity of $B$ into our assumptions, we take the following estimator of $B$ instead:

$$\widehat{B} = \text{Ta}(\widehat{C}_N - \widehat{A}_1 \otimes \widehat{A}_2),$$  

where $\text{Ta}(\cdot)$ is the “Toeplitz averaging” operator, that is, the projection onto the stationary operators, defined in Section C in the supplementary materials.

The next theorem shows asymptotic behavior of the proposed estimators under the following moment assumption.
(A1) Let \( \sum_{j=1}^{\infty} (E(X, \varepsilon_j))^4 < \infty \) for some orthonormal basis \( \{\varepsilon_j\}_{j=1}^{\infty} \) in \( L^2[0,1]^2 \).

This assumption ensures that \( \sqrt{N}(\hat{C}_N - C) \xrightarrow{d} Z \), where \( Z \) is a mean-zero Gaussian random element in \( S_1(L^2[0,1]^2) \), that is, the convergence is in the trace-norm topology Mas (2006).

**Theorem 1.** Let \( X_1, \ldots, X_N \sim X \) be a (w.l.o.g. centered) random sample with covariance given by (2), where \( B \) is stationary and \( \delta^* \)-banded. Let \( \delta \geq \delta^* \) such that \( Tr^2(C) \neq 0 \). Then \( \|A_1(\delta) \otimes A_2(\delta) - A_1 \otimes A_2\|_2^2 = O_p(N^{-1}) \) and \( \|\hat{B}(\delta) - B\|_2^2 = O_p(N^{-1}) \).

In fact, we show in the supplementary materials that even if model (2) is not valid, the chosen bandwidth leads to a separable-plus-banded proxy of \( C \), denoted \( \hat{C}(\delta) \), which is asymptotically optimal in the sense of (9). Recall that in this case \( \hat{C}(\delta) \) is not equal to \( C \) even asymptotically, since it is a biased proxy obtained via the proposed estimation methodology based on shifted partial tracing. We also give in the supplementary materials a version of Theorem 1 with an adaptively chosen bandwidth, that is, providing the limiting law further to the rates of convergence above, which requires a slight (and from the practical point of view unnecessary) modification of the bandwidth selection scheme.

### 2.4. Choice of Bandwidth

In order to apply our methodology in practice, it remains to provide means to choose the bandwidth \( \delta \). In this Section, we write \( \hat{C}(\delta) = A_1(\delta) \otimes A_2(\delta) + \hat{B}(\delta) \) to make the dependency on \( \delta \) explicit. Even though the separable part of the model \( A_1 \otimes A_2 \) does not depend on \( \delta \), its estimator from formula (6) does, since shifted partial tracing with a fixed bandwidth is used. If we actually knew the true covariance \( C \), we would use it in formulas (6) and (8) instead of the empirical covariance \( \hat{C}_N \) to obtain a separable-plus-banded proxy for \( C \), denoted here as \( \hat{C}(\delta) = A_1(\delta) \otimes A_2(\delta) + B(\delta) \). Under the separable-plus-banded model, it is \( \hat{C}(\delta) \sim C \) for any \( \delta \) large enough to eliminate \( B \) by \( \delta \)-shifted partial tracing. However, large bandwidths typically lead to vanishing shifted traces. Hence, a good bandwidth needs to be chosen with this tradeoff in mind.

Let \( \Delta := \{\delta_1, \ldots, \delta_n\} \) be the search grid of candidate values. If we knew \( C \), the bandwidth value leading to the best performance of our estimation methodology would be given by

\[
\delta^* := \arg \min_{\delta \in \Delta} \|C(\delta) - C\|_2^2.
\]

Here, \( \delta^* \) is a set. In particular, under model (2), \( \delta^* \) contains all such bandwidths \( \delta \) that \( B \) is banded by \( \delta \) and \( Tr^2(C) \neq 0 \). As suggested by Theorem 1, there is a range of valid bandwidths, which are asymptotically indistinguishable, and \( \delta^* \) contains all candidate values in this range. This reflects the fact that \( \delta \) is a nuisance parameter, not an estimand of interest.

Since we do not know \( C \) we cannot evaluate the objective in (9). Instead, we propose to approximate the objective by one that is fully calculable:

\[
\hat{\delta} := \arg \min_{\delta \in \Delta} \|\hat{C}(\delta)\|_2^2 = \frac{2}{N} \sum_{n=1}^{N} (X_n, \hat{C}(\delta)X_n),
\]

where \( \hat{C}(\delta) \) is our estimator constructed without the \( n \)th observation \( X_n \). In the supplementary materials, we show that (10) is \( \sqrt{N} \)-consistent for (9) up to a constant, and the following theorem provides rates of convergence with the adaptive choice of the bandwidth.

**Theorem 2.** Let \( X_1, \ldots, X_N \sim X \) be a (w.l.o.g. centered) random sample with covariance given by (2), where \( B \) is stationary and \( \delta^* \)-banded. Let (A1) hold, and let \( \hat{\delta} \) be chosen as in (10) from \( \Delta \) in which there exists \( \delta \geq \delta^* \) such that \( Tr^2(C) \neq 0 \). Then \( \|A_1(\hat{\delta}) \otimes A_2(\hat{\delta}) - A_1 \otimes A_2\|_2^2 = O_p(N^{-1}) \) and \( \|\hat{B}(\hat{\delta}) - B\|_2^2 = O_p(N^{-1}) \).

3. **Computational Considerations**

Consider now the practical scenario in which we only observe \( X \) discretely, that is, the random sample consists of matrices \( X_1, \ldots, X_N \in \mathbb{R}^{K_1 \times K_2} \). The discrete version of the covariance \( C \) is the covariance tensor \( C \in \mathbb{R}^{K_1 \times K_2 \times K_1 \times K_2} \), where \( C[i,j,k,l] = \text{cov}(X[i,j], X[k,l]) \). Assuming separability of \( C \) translates to \( C = C_1 \otimes C_2 \) for some \( C_1 \in \mathbb{R}^{K_1 \times K_1} \) and \( C_2 \in \mathbb{R}^{K_2 \times K_2} \), or entrywise \( C[i,j,k,l] = C_1[i]C_2[j]C_1[k]C_2[l] \) for \( i, k = 1, \ldots, K_1 \) and \( j, l = 1, \ldots, K_2 \).

First, assuming \( K_1 = K_2 =: K \) again for simplicity, a general covariance tensor \( C \) has \( O(K^4) \) degrees of freedom, while it only has \( O(K^3) \) degrees of freedom under the separability assumption. In comparison, the observed degrees of freedom are \( NK^2 \). Second, it takes \( O(NK^4) \) operations to calculate the empirical estimate of the covariance tensor, that is, \( \hat{C}_N = \frac{1}{N} \sum_{n=1}^{N} X \otimes X \), while this will be shown to reduce to \( O(NK^3) \) under separability. We assume throughout the article that multiplication of two \( K \times K \) matrices requires \( O(K^3) \) operations, and we set the cubic order in \( K \) as the limit of computational tractability for ourselves, which for example prevents us from ever explicitly calculating the empirical covariance \( \hat{C}_N \). Also, the degrees of freedom correspond to storage requirements, thus, although a general covariance tensor becomes difficult to manipulate on a standard computer for \( K \) as low as 100 (at that point the empirical covariance takes roughly 6 GB of memory), the situation under separability is much more favorable.

The following two properties hold for matrices \( A, B, \) and \( X \) of appropriate sizes:

\[
\begin{align*}
(A \otimes B)X &= AXB, \\
(A \otimes B)^{-1} &= A^{-1} \otimes B^{-1}.
\end{align*}
\]

These two properties are among the core reasons for popularity of the separability assumption in the space-time processes literature Gneiting, Genton, and Guttorp (2006), because they allow to apply a separable covariance fast \( O(K^3) \) instead of \( O(K^4) \) operations and solve an inverse problem involving the covariance fast \( O(K^3) \) instead of \( O(K^6) \) operations.

**Remark 1.** The symbol \( \otimes \) is commonly overused in the literature. In this article, we use it as the symbol for the abstract outer
product Weidmann (2012). The symbol $\otimes$ also denotes a type of abstract outer product, but we emphasize by the tilde that we do not see for example, $A \otimes B$ as an element of a product Hilbert space $S_p(H_1) \otimes S_p(H_2)$, but rather as an operator acting on a product Hilbert space $H = H_1 \otimes H_2$. The symbol $\otimes$ is used in linear algebra for the Kronecker product, which we denote $\otimes_K$ here. The following relation between the Kronecker product and the abstract outer product holds in the case of finite dimensional spaces:

$$\text{vec}((A \otimes B)X) = (B^T \otimes_K A)x,$$  

where $x = \text{vec}(X)$ is the vectorization of matrix $X$, and $\text{vec}(\cdot)$ is the vectorization operator (see, Van Loan and Golub 1983). Properties (11) are well known in computational linear algebra, where the Kronecker product is used instead of the abstract outer product. Due to (12), the first formula in (11) can be translated to $A \otimes_K B \text{vec}(X) = \text{vec}(B^T X A)$.

In summary, separability leads to an increased estimation accuracy, lower storage requirements, and faster computations. We view our separable-plus-banded model as a generalization of separability, and the aim of this section is to show that this generalization does not come at the cost of loosing the favorable properties of the separable model described above. In fact, we show in the remainder of this section that model (2) can be estimated and manipulated under the same computational costs as the separable model.

### 3.1. Estimation Complexity

When working with discrete samples, the shifted partial tracing is defined as before, only with the Lebesque measure replaced by the counting measure (see also section B, supplementary materials). This means, for example, that for $M \in \mathbb{R}^{K_1 \times K_2 \times K_1 \times K_2}$ and $d \leq \min(K_1, K_2)$, we have

$$\text{Tr}^d(M)[i,k] = \sum_{j=1}^{K_2-d} M[i,j,k,j+d], \quad i,k = 1, \ldots, K_1.$$  

Second, one can define $\text{Ta}(M) \in \mathbb{R}^{K_1 \times K_2 \times K_1 \times K_2}$ directly as the tensor having $S[h,l] = \frac{1}{\sqrt{d}} \sum_{j=1}^{K_2} \sum_{i=1}^{K_2} M[i,j,i+h-1,j+l-1]$ as its symbol. This justifies the name “Toeplitz averaging.” The relation to the discrete Fourier basis is discussed in Section C in the supplementary materials.

Now, consider the separable-plus-banded model $C = A_1 \otimes A_2 + B$ with $B$ banded by $d$, that is, $B[i,j,k,l] = 0$ whenever $\min(|i-k|,|j-l|) \geq d$. We use $d$ to denote the discrete version of the bandwidth $\delta$; the relation for an equidistant grid of size $K \times K$ is $d = \lfloor \delta K \rfloor$. It is straightforward to translate Proposition 1 and Lemma 1 to the discrete case to obtain the estimating equation $\text{Tr}^d(C)A_1 \otimes A_2 = \text{Tr}^d(C) \otimes \text{Tr}^d(C)$, suggesting again the plug-in estimators

$$\hat{A}_1 = \text{Tr}^d(\hat{C}_N) \quad \text{and} \quad \hat{A}_2 = \frac{\text{Tr}^d(\hat{C}_N)}{\text{Tr}^d(\hat{C}_N)}.$$  

It may be useful to revisit Example 1 and Figure 1 (which is plotted discretely anyway) for an intuitive depiction of discrete shifted partial tracing.

Now we are ready to establish the estimation complexity. First, we focus on shifted partial tracing. Due to linearity, $\text{Tr}^d(\hat{C}_N) = \frac{1}{N} \sum_{n} \text{Tr}^d(X_n \otimes X_n)$ and, as can be seen from formula (13), only $K^3$ entries of the total of $K^4$ entries of $X_n \otimes X_n$ are needed to evaluate the shifted partial trace. Moreover, evaluating the shifted partial trace amounts to averaging over one dimension of the relevant 3D object, which does not have to ever be stored, hence, the time and memory complexities to estimate the separable part of the model, that is, to evaluate (14), are $O(NK^3)$ and $O(K^2)$, respectively.

To evaluate $\bar{B} = \text{Ta}(C_N - \hat{A}_1 \otimes \hat{A}_2) = \frac{1}{N} \sum_{n} \text{Ta}(X_n \otimes X_n) - \text{Ta}(A_1 \otimes A_2)$, one can use the fast Fourier transform (FFT). Every term $\text{Ta}(X_n \otimes X_n)$ can be evaluated directly on the level of data, without the necessity to form the empirical estimator, in $O(K^2 \log(K))$. The term $\text{Ta}(A_1 \otimes A_2)$ can be evaluated in $O(K^3)$ operations, again without explicitly forming the outer product. See Section C in the supplementary materials for details. Hence, estimation of the banded part is equally demanding as the estimation of the separable part.

It remains to show that $\hat{C} := \hat{A}_1 \otimes \hat{A}_2 + \bar{B}$ can be applied efficiently, that the bandwidth selection strategy is feasible, and that an inverse problem $CX = Y$ can be solved efficiently. The application of $\hat{C}$ is simple due to the additive structure: one applies the separable part using the first formula in (11), the banded part using the FFT, and sums the two, leading to the desired complexities. For bandwidth selection, one needs to evaluate the objective of (10) for all $\delta \in \Delta$. The norm in (10) can be calculated fast using the separable-plus-banded form of the estimator, see Section D in the supplementary materials. The inner products take $O(K^2)$ operations each after a fast application of $\hat{C}_{-n}(\delta)$. However, it is wasteful to re-estimate $\hat{C}_{-n}(\delta)$, holding out one observation at a time. Instead, one can split the data into for example, 10-fold, and hold out each fold as a whole. Evaluation for a single fold then takes $O(NK^3)$, since estimation is the dominating operation. Hence, overall, bandwidth selection is computationally tractable. Finally, the inverse problem is nontrivial, since it is not possible to express the inverse of a sum of two operators in terms of inverses of the two summands. This problem is handled with in the following section.

### 3.2. Inverse Problem

We need a fast solver for the linear system coming from a discretization of model (2), that is,

$$(A_1 \otimes A_2 + B)X = Y,$$  

where $B$ is stationary. Equation (15) can be rewritten in the matrix-vector form as

$$(A + B)x = y,$$  

where $A = A_2 \otimes_K A_1$ (see Remark 1), $x = \text{vec}(X)$, $y = \text{vec}(Y)$, and $B \in \mathbb{R}^{K^2 \times K^2}$ is a two-level Toeplitz matrix (i.e., a Toeplitz block matrix with Toeplitz blocks).

The naive solution to system (16) would require $O(K^6)$ operations. Since the estimation of model (2) takes $O(NK^3)$, we are looking for a solver for (16) with a complexity close to $O(K^3)$. We will develop an alternating direction implicit (ADI,
see Young (2014) solver with the per-iteration cost of $O(K^3)$ and rapid convergence.

The system (16) can be transformed into either of the following two systems:

\[
\begin{align*}
(A + \rho I)x &= y - Bx + \rho x, \\
(B + \rho I)x &= y - Ax + \rho x,
\end{align*}
\]  
(17)

where $I \in \mathbb{R}^{K \times K}$ is the identity matrix and $\rho \geq 0$ is arbitrary. The idea of the ADI method is to start from an initial solution $x^{(0)}$, and form a sequence $\{x^{(k)}\}_{k=2}^{\infty}$ by alternately solving the linearized systems stemming from (17) until convergence, specifically:

\[
\begin{align*}
(A + \rho I)x^{(k+1/2)} &= y - Bx^{(k)} + \rho x^{(k)}, \\
(B + \rho I)x^{(k+1)} &= y - Ax^{(k+1/2)} + \rho x^{(k+1/2)}.
\end{align*}
\]  
(18)

The acceleration parameter $\rho$ is allowed to vary between iterations. The optimal choice of $\rho$ based on the spectral properties of $A$ and $B$, guaranteeing a fixed number of iterations, can be made in some model examples (e.g., when $A$ and $B$ commute).

Interestingly, numerical studies suggest that the ADI method exhibits excellent performance on a large class of linear systems of the type (16) with the model choice of $\rho$, as long as matrices $A$ and $B$ are real with real spectra. Hence, we also choose $\rho$ as suggested by the model examples and, in order to boost the convergence speed, we gradually decrease its value as $\rho^{(k)} = \min(\rho^{(k-1)}, \frac{|x^{(k+1/2)} - x^{(k)}|_2}{|x^{(k)}|_2})$, $k \in \mathbb{N}$, with $\rho^{(0)} = \sqrt{\max(\rho_{\text{max}} / \rho_{\text{min}})} + \epsilon$, where $\rho_{\text{max}}$ and $\rho_{\text{min}}$ (resp. $\beta_{\text{max}}$ and $\beta_{\text{min}}$) are maximum and minimum eigenvalues of $A$ (resp. $B$), and $\epsilon$ is a small positive constant (by default the desired precision). Recall that $A$ and $B$ are positive semidefinite.

Now it remains to show how to efficiently solve the linear subproblems (18). The first subproblem has an analytic solution given in the matrix form by

\[
X^{(k+1/2)} = V[G \odot U^T (Y - BX^{(k)}) + \rho X^{(k)}]V^T,
\]

where $V$ and $U$ are formed by the eigenvectors of $A_1$ and $A_2$, respectively, $G$ depends on the eigenvalues of $A_1$ and $A_2$, and $\odot$ denotes the Hadamard (element-wise) product. This solution is computable in $O(K^3)$ operations. The second subproblem in (18) involves a two-level Toeplitz matrix as its left-hand side and can be solved iteratively via preconditioned conjugate gradient, with a single-step complexity of $O(K^2 \log(K))$. See Section F in the supplementary materials for details.

In summary, we devised a doubly iterative algorithm to solve inverse problems in the context of the separable-plus-stationary model. The outer iterative scheme requires solution of two linear systems, one solvable in $O(K^3)$ iterations, the other in $O(\eta_{ps} K^2 \log(K))$, where $\eta_{ps}$ is the number of the iterations of the inner scheme. In Section 5.1, we demonstrate empirically that $\eta_{ps}$ does not increase with increasing $K$, and hence, the overall complexity of the algorithm is $O(\eta_{adi} K^3)$, where $\eta_{adi}$ is the number of outer iterations. As demonstrated again in Section 5.1, $\eta_{adi}$ also does not depend on $K$, leading to an overall complexity $O(K^3)$. Hence, we have a tractable inversion algorithm for the separable-plus-stationary model.

Note that stationarity of $B$ is used at two instances: in the top right-hand side of (18), $B$ needs to be applied fast, and in (B + \rho I)x = y needs to be solved fast. Both of these are easy if, for example, $B$ is diagonal rather than stationary. Hence, we also have inversion algorithm when a separable covariance is observed under heteroscedastic noise.

4. Asymptotics under Discrete Noisy Measurements

While Theorem 1 establishes the asymptotic behavior of our estimators under complete observations, in practice one only observes discrete and potentially noisy samples, which are the topic of this section. Let $[0, 1]^2 = \bigcup_{i=1}^K \bigcup_{j=1}^K I_{ij}$, where $I_{ij}$ is a Cartesian product of two subintervals of $[0, 1]$. Let $I_{ij} \cap I_{i', j'} = \emptyset$ for $(i, j) \neq (i', j')$, and $|I_{ij}| = K^{-2}$ for all $i, j = 1, \ldots, K$. The observations are assumed to be of the form

\[
\tilde{X}^K_{ij} = X^K_{ij} + E^K_{n,ij}, \quad i, j = 1, \ldots, K,
\]  
(19)

where the matrices $X_1, \ldots, X_N \in \mathbb{R}^{K \times K}$ are discretely measured versions of the latent surfaces $X_1, \ldots, X_N \in \mathbb{L}^2[0, 1]^2$, and $E^K_{n}$ are measurement errors. We will consider two types of sampling, which relate the latent surfaces $X_1, \ldots, X_N \in \mathbb{L}^2[0, 1]^2$ to the discrete data $X_1, \ldots, X_N \in \mathbb{R}^{K \times K}$.

(S1) $X_n, n = 1, \ldots, N$, are observed pointwise on a grid, that is, there exist $t^K_{1}, \ldots, t^K_{K} \in [0, 1]$ and $s^K_{1}, \ldots, s^K_{K_2} \in [0, 1]$ such that $(t^K_{1}, s^K_{1}) \in I_{ij}$ and

\[
X^K_{n,ij} = X_n(t^K_{i}, s^K_{j}), \quad i, j = 1, \ldots, K,
\]

Note that to make such point evaluations of $X$ meaningful, we have to assume that realizations of $X$ are continuous (see, Hsing and Eubank 2015).

(S2) The average value of $X_n$ on the pixel $I^K_{ij}$ is observed for every pixel, that is,

\[
X^K_{n,ij} = \frac{1}{|I^K_{ij}|} \int_{I^K_{ij}} X_n(t,s)dt, \quad i, j = 1, \ldots, K.
\]

As for the measurement error arrays $(E^K_{n,ij})_{i,j=1}^K$, these are assumed to be iid (with respect to $n$) and uncorrelated with $X_n$, satisfying the following fourth order moment conditions for $i, j, k, l, i', j', k', l' = 1, \ldots, K$ and $n = 1, \ldots, N$:

\[
\begin{align*}
E(E^K_{n,ij} | k, l) &= 0, \\
\sigma^2 < \infty &\iff E(E^K_{n,ij} | k, l) \leq \sigma^2 \mathbf{1}_{[i \neq k, j \neq l]}, \\
E(E^K_{n,ij} | k, l)X^K_{n,ij} &\leq E(E^K_{n,ij} | k, l)X^K_{n,k',l'} = E(E^K_{n,ij} | k, l)\mathbb{E}(X^K_{n,i',j'} | X^K_{n,k',l'}).
\end{align*}
\]

The previous inequality allows for heteroscedasticity of the noise process: we allow the variance to change with location, but assume it is bounded over the domain by an unknown constant $\sigma^2$. Note that under the sampling scheme (S1) and homoscedasticity (constraining the inequality to equality), Equation (19) corresponds to the commonly adopted errors-in-measurements model Yao, Müller, and Wang (2005) and Zhang and Wang (2016).

Let us denote $X^K(t,s) = \sum_{i=1}^{K} \sum_{j=1}^{K} X^K_{ij}I_{|(t,s)\in I^K_{ij}}$ that is, $X^K$ is the piecewise constant function of $X^K$. One can
readily verify that pointwise sampling (scheme S1) corresponds to pointwise evaluations of the covariance, that is, \( \text{var}(X^K) = C^K \), where \( C^K \) has kernel

\[
e^K(t, s, t', s') = \sum_{i,j,k,l=1}^{K} c_{ijkl} E_{((t_i,s_i)\in E^K_{(i)})} E_{((t',s',i')\in E^K_{(i')})},
\]

while pixel-wise sampling (scheme S2) corresponds in turn to pixelization of the covariance. Namely, if we denote \( g^K_{ijkl}(t, s) = K E_{((t_i,s_i)\in E^K_{(i)})} \) then we have \( \text{var}(X^K) = C^K \) with

\[
X^K = \sum_{i,j=1}^{K} (X_i g^K_{ij} d) g^K_{ij}, \quad C^K = \sum_{i,j,k,l=1}^{K} (C_{ijkl} \otimes g^K_{ij}) (\delta \otimes g^K_{kl}).
\]

(20)

In the same spirit, \( C^K \) is the piecewise constant continuation of \( C^K = E(X^K \otimes X^K) \).

If we constrain ourselves to the noiseless multivariate setting and consider the discrete version of the covariance to be the ground truth, it is straightforward to obtain the multivariate version of Theorem 1, regardless of the sampling scheme. When both \( N \) and \( K \) diverge, Theorem 1 does not apply, but we can still obtain convergence rates. To this aim, we first ought to clarify how bandedness of \( B, B^K \) (the discrete version of \( B \) and \( B^K \) (the piece-wise continuation of \( B^K \)) are related. It can be seen that if \( B \) is banded by \( \delta \), then \( B^K \) is banded by \( d^K = \delta K \), while \( B^K \) is banded by \( d_K = d^K/K \), which decreases monotonically down to \( \delta \) for \( K \to \infty \). In the following theorem, \( \hat{A}^1 \) and \( \hat{A}^2 \) denote piecewise constant continuations of \( \hat{A}^1 = \text{Tr}^d (\hat{C}^1_K) \) and \( \hat{A}^2 = \text{Tr}^d (\hat{C}^2_N)/\text{Tr}^d (\hat{C}^2_K) \), where \( \hat{C}^K = \frac{1}{N} \sum_{i=1}^{N} \hat{X}_i^K \otimes \hat{X}_i^K \) is the empirical covariance based on the observed (noisy) data (19).

**Theorem 3.** Let \( X_1, \ldots, X_N \) be iid copies of \( X \in L^2[0,1]^2 \), which has (w.l.o.g. mean zero and) covariance given by (2), where the separable part \( A := A_1 \otimes A_2 \) has kernel \( \alpha(t_i, s_i, t'_i, s'_i) \), which is Lipschitz continuous on \([0,1]^4 \) with a Lipschitz constant \( L > 0 \). Let \( \|X_i\|^4 < \infty \) and \( \delta \in (0, 1) \) be such that \( B \) from (2) is banded by \( \delta \) and \( \|B(A)\| \neq 0 \). Let the samples come from (19) via measurement scheme (S1) or (S2) with \( \text{var}(E_0[i, j]) \leq \sigma^2 = O(\sqrt{K}) \). Then we have

\[
\|\hat{A}^1 \otimes \hat{A}^2 - A_1 \otimes A_2\|^2 = \mathcal{O}(p(N^{-1}) + 2K^{-2}L^2),
\]

(21)

where the \( \mathcal{O}(p(N^{-1})) \) term is uniform in \( K \), for all \( K \geq K_0 \) for a certain \( K_0 \in \mathbb{N} \). Furthermore, if \( A_1 = \sum_{j \in \mathbb{N}} \hat{\lambda}_j \otimes f_j, A_2 = \sum_{j \in \mathbb{N}} \lambda_j \otimes \hat{f}_j, \) and \( A_3 = \sum_{j \in \mathbb{N}} \hat{\lambda}_j \otimes \hat{f}_j \) are eigendecompositions, then \( ||\hat{\lambda}_j \otimes \lambda_j \otimes f_j - \lambda_j \otimes f_j||^2 \) follows the rate given in (21), and if the eigenspace associated with \( \lambda_j \) is one-dimensional, then also \( ||\hat{\lambda}_j \otimes \hat{\lambda}_j \otimes f_j - \lambda_j \otimes f_j||^2 \) follows the rate given in (21).

**Remark 2.** Since the roles of \( A_1 \) and \( A_2 \) are symmetric, one naturally obtains the rates for \( f_j^2 \) as well. Second, one can show that the rate in (21) is valid also in the uniform norm, see Theorem G.2 in the supplementary materials. Finally, a version of Theorem 3 with the bandwidth chosen adaptively as in Section 2.4 is also valid, see Theorem G.3 in the supplementary materials.

The proofs are postponed to the supplementary materials, but we make several comments here. First, there is a concentration in \( K \) due to shifted partial tracing (recall Figure 1), hence, the variance of the errors is allowed to grow with \( K \) as stated in Theorem 3. Second, the estimators \( \hat{A}_1^1 \) and \( \hat{A}_2^2 \) are only defined if \( \text{Tr}^d (\hat{C}^1_K) \neq 0 \). Since \( \hat{C}^1_N \to C^K \) for \( N \to \infty \) entry-wise apart from the diagonal, we have \( \text{Tr}^d (\hat{C}^1_N) \to \text{Tr}^d (C^K) \), so we require \( \text{Tr}^d (C^K) \neq 0 \). Due to continuity of the kernel \( c \) and the fact that \( d_K \to \delta \) for \( K \to \infty \), the assumption \( \text{Tr}^d (A) \neq 0 \) implies \( \text{Tr}^d (C^K) \neq 0 \) for a sufficiently large \( K \). This is the only reason why we require \( K \) larger than a certain \( K_0 \) in order for the \( \mathcal{O}(p(N^{-1}) \) term to be uniform in \( K \). Finally, the Lipschitz continuity assumption can be weakened. For example, continuity almost everywhere is sufficient for the bias to converge to zero, though without an explicit rate in \( K \).

In case the banded part of the covariance is also of interest, the same rates can be achieved in the noiseless setting (\( \sigma^2 = 0 \)) under smoothness assumptions on the banded part. Without the assumption of stationarity on \( B \), that is, without Toeplitz averaging, one has:

\[
|||\hat{B}^K - B|||^2 \leq \|\hat{B}^K - B^K|||^2 + \|B^K - B|||^2 \\
\leq \|\hat{C}^1_N - A_1^1 \otimes A_2^2 - (C^K - A_1^1 \otimes A_2^2)|||^2 \\
+ ||B^K - B|||^2 \\
\leq \|\hat{C}^1_N - C^K|||^2 + ||A_1^1 \otimes A_2^2 - A_1^1 \otimes A_2^2|||^2 \\
+ ||B^K - B|||^2 \\
\]

where the separable term can be treated as before and \( ||\hat{C}^1_N - C^K|||^2 \) can be bounded similarly. When Toeplitz averaging is used, nothing essential changes in the noiseless case.

The noisy case (\( \sigma^2 > 0 \)) is trickier however, because we cannot estimate the diagonal of \( B \). In such a case, one would need to smooth the estimated symbol of \( B \) as in Yao, Müller, and Wang (2005). We omit the details here. However, we note that full covariance smoothing is obviously not computationally tractable, hence, any smoothing should either be applied on the level of data (presmoothing) or on the level of the estimated 2D parts of the covariance (postsmeoothing). Nonetheless, as exemplified by the previous theorem, the mere presence of noise does not call for smoothing when the target of inference is the separable component.

**Remark 3.** In the noiseless case, the convergence rates in Theorem 3 are immediately applicable to the special case of a separable model and standard (nonshifted) partial tracing, as used by Aston, Pigoli, and Tavakoli (2017). In the noisy case, however, shifted partial tracing (with an arbitrarily small shift) is needed to remove the noise. Due to continuity, a small shift should have a small impact on the quality of the estimator. Hence, it might be recommended to always use shifted partial tracing with the minimal possible shift instead of the standard (nonshifted) partial tracing.
5. Empirical Demonstration

In this section, we demonstrate how our methodology can be used to estimate a covariance from surface data observed on a grid, and how it compares to the empirical covariance estimator and the separable model, estimated via partial tracing Astón, Pogli, and Tavakoli (2017) or as the nearest Kronecker product Genton (2007). We begin with simulated data in Section 5.1, where we focus on a weakly dependent contamination of separability, and then move on to real data in Section 5.2, where we find evidence for heteroscedastic white noise contamination.

5.1. Simulation Study

The data generation procedure is as follows. First, we create covariances \( A_1, A_2 \in \mathbb{R}^{K \times K} \) and draw \( Y_1, \ldots, Y_N \) independently from the matrix-variate Gaussian distribution with mean zero and covariance \( A = A_1 \otimes A_2 \). Second, we draw enough \( \mathcal{N}(0,1) \) entries (independent of everything), arrange them on a grid, and perform space-time averaging using a window of size \( d \in \{1, 3, \ldots, 19\} \) to obtain a sample \( W_n \) for every \( n = 1, \ldots, N \). This sample is drawn from a distribution with mean zero and covariance \( B \in \mathbb{R}^{K \times K \times K} \), which is by construction stationary and banded by \( d \). We set the sample size \( N = 300 \) and the grid size \( K = 100 \), so the discrete bandwidth \( d \) approximately corresponds to the continuous bandwidth \( \delta \) in percentages. Finally, we form our dataset \( X_1, \ldots, X_N \in \mathbb{R}^{K \times K} \) as \( X_n = \sqrt{\tau} Y_n + W_n \), \( n = 1, \ldots, N \), where \( \tau \geq 1 \). Thus, \( X_1, \ldots, X_N \in \mathbb{R}^{K \times K} \) are drawn from a zero-mean distribution with a separable-plus-banded covariance \( C = \tau A_1 \otimes A_2 + B \). Since \( A_1, A_2 \) and \( B \) are standardized to have norm one, \( \tau \) can be understood as a signal-to-noise ratio. The separable constituents \( A_1 \) and \( A_2 \) are chosen both as rank-7 covariances with linearly decaying eigenvalues and shifted Legendre polynomials as the eigenvectors, while \( B \) corresponds to the covariance of a spatio-temporal moving average process, as per the construction above. See Section H in the supplementary materials for a detailed description and for other simulation results in additional setups.

Note that our methodology based on shifted partial tracing first estimates the separable part of the model, and subsequently estimates \( B \) using the estimates for the separable part. Therefore, the signal-to-noise ratio \( \tau \) naturally influences difficulty of the estimation problem. The second parameter governing the difficulty of the estimation problem is the bandwidth \( d \). However, the effect of \( d \) is discontinuous: a small \( d \) does not correspond to a nearly separable model; only \( d = 0 \) formally leads to separable model with no contamination.

The following methods were used to estimate \( C \): SPT-\( d \)—shifted partial tracing, the proposed methodology of Section 2.3, provided with the true bandwidth \( d \); SPT-CV—shifted partial tracing with \( \delta \) chosen as in (10); PT—partial tracing Astón, Pogli, and Tavakoli (2017), an approach assuming separability; NKP—nearest Kronecker product Genton (2007), another approach assuming separability; ECE—the standard empirical covariance estimator. For several different settings, we calculate the relative estimation error \( ||C - \hat{C}||_F / ||C||_F \), where \( C \) is an estimator computed by one of the above-listed methods. The plots also show the bias of a separable estimator, calculated as the best separable approximation to the true covariance \( C \) Van Loan and Pitsianis (1993).

Figure 2 depicts how the estimation error evolves when one of the three difficulty-governing parameters (bandwidth \( d \), signal-to-noise ratio \( \tau \), and sample size \( N \)) varies, while the remaining two parameters are held fixed at any given plot (at \( d = 9, \tau = 3 \) or \( N = 300 \)). There are several remarks to be made about the results in Figures 2:

1. Shifted partial tracing outperforms both the separable model (estimated either by partial tracing or as the nearest Kronecker product) and the empirical covariance.
2. Bandwidth selection works well, leading to the same or even better performance than with known \( \delta \) (see the right tail of the left and middle plots in Figure 2). This is because the banded part \( B \) decays away from the diagonal, and sometimes choosing a smaller bandwidth than the true one can lead to a better bias-variance tradeoff.
3. When the truth is separable (i.e., \( d = 0 \)) or nearly separable (i.e., \( \tau \) large), partial tracing leads to the best results. In these cases, the bandwidth selection strategy correctly chooses a very small bandwidth, and hence, the performance of SPT-CV matches the one of PT.
4. Note the extreme rise at the beginning of the error curves belonging to the empirical or the separable estimators in Figure 2 (left). While \( d = 0 \) corresponds to a separable model, \( d = 1 \) is already quite nonseparable. Even though the amount of nonseparability (see, the bias curve) is rather low, it is enough to substantially deteriorate performance of the separable estimators or the empirical covariance, while
performance of the proposed methodology does not suffer too much.

5. The previous point is manifested again for large sample sizes
   \( N \) (see Figure 2, right). While the amount of nonseparability
   of \( \mathbf{C} \) is still low and one would expect the performance of
   the separable estimators to be quite good, this is not the case.
   Actually, the separable estimators can be outperformed even
   by the empirical covariance here. Altogether, we can say that
   presence of a banded structure strikingly obstructs separable
   estimation, even when the amount of nonseparable bias is
   relatively low.

In the remainder of this section, we examine the functional
nature of our problem, behavior of the ADI algorithm of Section 3,
and the number of iterations needed by the algorithm to converge.
We simulate data as described before in the Legendre case, but now we vary the grid size \( K \in \{10(2j + 1); j = 1, \ldots, 10\} \), fix \( \delta \) at 10\% (i.e., \( d = K/10 \)), and we keep \( \tau = 3 \) and \( N = 300 \) for all the grid sizes.

Let \( \hat{\mathbf{C}} = \hat{\mathbf{A}}_1 \otimes \hat{\mathbf{A}}_2 + \hat{\mathbf{B}} \) denote the estimator obtained by
shifted partial tracing. \( \hat{\mathbf{A}}_1, \hat{\mathbf{A}}_2 \) and \( \hat{\mathbf{B}} \) are subsequently projected
onto positive semidefinite matrices, as described in Section E in the
supplementary materials. Also, a ridge regularization of order \( 10^{-5} \) is added to \( \hat{\mathbf{C}} \). This is not necessary, because \( \hat{\mathbf{B}} \)
is positive definite, and thus the problem is well defined even
without any ridge regularization. However, the performance of
the ADI method heavily depends on the condition number of
the system matrix, as is the case for any numerical method.
Adding the ridge regularization ensures that the condition
number stays roughly the same, regardless of \( K \). Then, a random
\( \mathbf{X} \in \mathbb{R}^{K \times K} \) is generated, and we set \( \mathbf{Y} = \mathbf{CX} \). Subsequently, the
ADI algorithm is called on the inverse problem \( \hat{\mathbf{C}} \mathbf{X} = \mathbf{Y} \) with \( \hat{\mathbf{C}} \)
and \( \mathbf{Y} \) given. The desired relative accuracy for the ADI scheme
is set to \( 10^{-6} \). We do not report the relative reconstruction
errors of \( \mathbf{X} \), because these varied between \( 10^{-7} \) and \( 10^{-11} \)
for every single run, leaving no doubt that the ADI scheme always
converged to the truth with the desired precision. Instead, we
report estimation errors and number of iterations needed by the
ADI scheme in Figure 3.

As suggested by our theoretical results, the relative estimation
error does not depend on the grid size (see Figure 3, left).
Additionally, both the number of outer iterations (ADI) and the
number of inner iterations (PCG) does not seem to increase
with the grid size (see Figure 3, right). This suggests super-linear
convergence of the algorithm.

5.2. Real Data

We analyze a dataset \( \mathbf{X} \in \mathbb{R}^{N \times K_1 \times K_2} \), where \( \mathbf{X}[n, k_1, k_2] \) denotes
the mortality rate for the \( n \)th country, on the \( k_1 \)th calendar
year and for subjects of age \( k_2 \). We consider the same set of 32
countries as Chen and Müller (2012) and Chen, Delicado, and
Müller (2017), with \( k_1 \) ranging in the 50 year span 1964–2014,
and we too focus on the mortality rates of older individuals aged
between \( 60 \leq k_2 < 100 \). Hence, \( \mathbf{X} \in \mathbb{R}^{12 \times 50 \times 40} \). For a single
country, we thus have a mortality rate surface of two arguments:
the calendar year and the age of subjects in the population.
This surface is observed discretely since both the calendar year
and age are integers. Figure 4 shows the raw mortality surfaces
for two sample countries. The underlying continuous surfaces
for different countries are assumed to be iid functional observa-
tions. The data were obtained from the Human Mortality
Database (Wilmoth et al. 2007, www.mortality.org, downloaded
on 12/4/2019).

An in-depth analysis of mortality surfaces was provided in
Chen and Müller (2012). Mortality surfaces were also considered
by the authors of Chen, Delicado, and Müller (2017), who—presumably motivated by Aston, Pigoli, and
Tavakoli (2017) and aiming for computational efficiency—
calculated the so-called marginal kernels \( \text{Tr}_1(\hat{\mathbf{C}}_N) \) and \( \text{Tr}_2(\hat{\mathbf{C}}_N) \),
found the leading eigenfunctions of these marginal kernels, say
(\( \hat{\phi}_i \)) and (\( \hat{\psi}_j \)), and used the tensor product approximation
\( \hat{\mathbf{C}}_N \approx \sum_{i=1}^I \sum_{j=1}^J \hat{\gamma}_{ij} (\hat{\phi}_i \otimes \hat{\psi}_j) \). Indeed, we highlight that
using the marginal eigenfunctions as building blocks for
a low-rank approximation of the empirical covariance can
be meaningful even if the covariance \( \mathbf{C} \) is not separable
Lynch and Chen (2018).

Compared to Chen and Müller (2012) and Chen, Delicado,
and Müller (2017), we consider the mortality data with a slightly
larger span of calendar years (the maximal span in which no data
are missing). Our aim here is not to provide a novel analysis of

Figure 3. Left: Estimation errors for several competing methods depending on the grid size \( K \) with the bandwidth fixed at \( d = K/10 \). Right: Number of iterations needed by the outer iteration scheme (ADI) and the inner iteration scheme (PCG) of the inversion algorithm of Section 3.2.
the mortality dataset, but merely to illustrate the usefulness of shifted partial tracing.

First, when investigating the sample curves in Figure 4, it seems that the discrete observations of the mortality rate surfaces are observed with additional noise, which is likely heteroscedastic with variance increasing with the age of the subjects. This is presumably due to the fact that the size of the population of subjects of a given age shrinks with increasing age. To probe whether the (most likely heteroscedastic) noise disrupts separability, we can use the bandwidth selection procedure. We do not assume stationarity, and we set the estimator of the banded part (7) to zero outside of the current bandwidth in every step. The objective of (10) is maximized at $\hat{d} = 1$. We plot the objective curve in Figure 5, providing a strong evidence for presence of noise. Since $\hat{d} = 1$, we are in the separable-plus-noise regime, which is computationally feasible even under heteroscedasticity. Figure 5 also shows a heatmap of the estimated variance (or rather its logarithm, for visualization purposes) of the noise depending on the location. The heatmap is in alignment with the conjecture that the noise variance is increasing with age.

Second, we compare spectra of the marginal kernels $\text{Tr}_1(\hat{C}_N)$ and $\text{Tr}_2(\hat{C}_N)$ to their shifted counterparts $\text{Tr}_1(\hat{C}_N)$ and $\text{Tr}_2(\hat{C}_N)$. When partial tracing is used to obtain the marginal kernels, one has to keep 16 and 4 eigenfunctions, respectively, to capture 90% of the marginal variance (see, Lynch and Chen 2018) in both dimensions. When shifted partial tracing is used instead, one only needs to retain 4 and 2 eigenfunctions, respectively. Hence, shifted partial tracing offers a more parsimonious representation.

Third, the empirical bootstrap test of Aston, Pigoli, and Tavakoli (2017) with 4 and 2 marginal eigenfunctions (which...
seems to be the most reasonable choice, also used by Lynch and Chen (2018)) leads to a borderline p-value of 0.06. The test of Aston, Pigoli, and Tavakoli (2017) can be generalized to testing separable-plus-banded model instead, see Section I in the supplementary materials. In comparison, the p-value for this test is over 0.4, suggesting that the separable-plus-banded model cannot be rejected for this dataset.

Finally, keeping only two eigenfunctions in both dimensions (explaining 83% and 96% of the variance, respectively) leads to a plausible interpretation when shifted partial tracing is used. The eigenfunctions are plotted in Figure 6. The first eigenfunctions in both dimensions capture the overall trend: $\hat{\phi}_1$ captures the decreasing variance in calendar years (the first dimension) and $\hat{\psi}_1$ the increasing variance in age (the second dimension). The second eigenfunction in the first dimension $\hat{\phi}_2$ distinguishes between countries having either a ”U-shape” (the Czech Republic, for example) or reversed ”U-shape” in calendar years (Switzerland, for example). This ”U-shape” is more prominent in older ages, but it barely visible by eye in the raw data plotted in Figure 4. Finally, the second eigenfunction in the second dimension $\hat{\psi}_2$ contrasts the old age (around 85) and the oldest age (post 90) mortalities. However, this is only the case if shifted partial tracing is used. The eigenfunction $\hat{\psi}_2$ obtained from $\text{Tr}_2(\hat{C}_N)$ does not have this interpretation; it is in fact not interpretable. Interestingly, the same qualitative conclusions as those drawn here by using shifted partial tracing were drawn in Chen and Müller (2012) based on a different methodology.

6. Discussion

The immense popularity of separability stems mainly from the computational advantages it entails. The separable-plus-banded model we propose is an additive generalization of separability. To retain the computational advantages, many natural operations (such as forming the empirical covariance estimator or naively inverting the estimated model) are prohibited. Efficient estimation of the separable-plus-banded model can be achieved with shifted partial tracing—a novel methodology for estimation of the covariance of surface-valued processes, working on the level of data.

From another point of view, shifted partial tracing can be used to estimate a separable model, even when data are corrupted by heteroscedastic and/or weakly dependent noise. The noise can be incorporated for tasks such as prediction with no computational overhead compared to the simple separable model, whenever it (or equivalently the banded part of the model) does not exceed the separable model in terms of the degrees of freedom it possesses. The latter is true for example when the noise is heteroscedastic and white (as was the case in the real data analysis) or when the noise is weakly dependent but stationary (as was the case in the simulation study). If the noise is both heteroscedastic and weakly dependent, our methodology can still be used. However, one has to pay extra computational costs (compared to just a noiseless separable model), if one wishes to explicitly work with a noise structure, which has higher complexity than that of the separable model.

Following Aston, Pigoli, and Tavakoli (2017), partial tracing has become the method of choice for calculation of the marginal kernels, that is, for calculating a separable proxy of the covariance. However, given the theoretical development here and the practical evidence found in the mortality dataset, it seems that shifted partial tracing should in general be preferred, due to its denoising properties.

A straightforward extension of our work would be to consider other forms of parsimony than stationarity, which can be imposed on the banded part of the covariance to maintain the computational advantages of our model. In principle, any form of parsimony, which reduces the complexity of storing $B$ to $O(K^2)$ and the number of flops required to apply $B$ to $O(K^3)$, may be considered. The precise form of parsimony capable of complementing separability well in the additive model will depend on the specific application.

Supplementary Materials

The supplementary materials generalize the development of shifted partial tracing, provide computational details and proofs of the asymptotic results, and develop a bootstrap test for validity of the separable-plus-banded model. They also contain R codes to reproduce all the results reported in the paper.

Acknowledgments

We thank the associate editor and the two reviewers for their insightful comments. Also, we thank Prof. John Aston (Cambridge), Prof. Daniel Kressner (EPFL), and Prof. Shahin Tavakoli (Warwick) for very fruitful discussions.

Funding

Research supported by a Swiss National Science Foundation grant.

References

Aston, J. A. D., Pigoli, D., and Tavakoli, S. (2017), “Tests for Separability in Nonparametric Covariance Operators of Random Surfaces,” The Annals of Statistics, 45, 1431–1461. [2562,2563,2564,2569,2570,2571,2572,2573]
Bagchi, P., and Dette, H. (2020), “A Test for Separability in Covariance Operators of Random Surfaces,” Annals of Statistics, 48, 2302–2322. [2563]
Chen, K., Delicado, P., and Müller, H.-G. (2017), “Modelling Function-Valued Stochastic Processes, with Applications to Fertility Dynamics,” Journal of the Royal Statistical Society, Series B, 79, 777–808. [2571]
Chen, K., and Müller, H.-G. (2012), “Modeling Repeated Functional Observations,” Journal of the American Statistical Association, 107, 1599–1609. [2571,2573]
Constantinou, P., Kokoszka, P., and Reimherr, M. (2017), “Testing Separability of Space-Time Functional Processes,” Biometrika, 104, 425–437. [2563]
Genton, M. G. (2007), “Separable Approximations of Space-Time Covariance Matrices,” Environmetrics, 18, 681–695. [2562,2570]
Gneiting, T. (2002), “Nonseparable, Stationary Covariance Functions for Space–Time Data,” Journal of the American Statistical Association, 97, 590–600. [2562]
Gneiting, T., Genton, M. G., and Guttorp, P. (2006), Geostatistical Space-Time Models, Stationarity, Separability, and Full Symmetry, pp. 151–175, Boca Raton, FL: Chapman and Hall/CRC. [2562,2566]
Hsing, T., and Eubank, R. (2015), Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators, Chichester: Wiley. [2562,2564,2568]
Lynch, B., and Chen, K. (2018), “A Test of Weak Separability for Multiway Functional Data, with Application to Brain Connectivity Studies,” Biometrika, 105, 815–831. [2571,2572,2573]
Mas, A. (2006), “A Sufficient Condition for the CLT in the Space of Nuclear Operators—Application to Covariance of Random Functions,” Statistics & Probability Letters, 76, 1503–1509. [2566]
Ramsay, J. O., and Silverman, B. W. (2002), Applied Functional Data Analysis: Methods and Case Studies, New York: Springer. [2562]
(2005), Functional Data Analysis, New York: Springer-Verlag. [2563]
Rougier, J. (2017), “A Representation Theorem for Stochastic Processes with Separable Covariance Functions, and its Implications for Emulation,” arXiv preprint arXiv:1702.05599. [2563]
Van Loan, C. F., and Golub, G. H. (1983), Matrix Computations, Baltimore, MD: Johns Hopkins University Press. [2567]
Van Loan, C. F., and Pitsianis, N. (1993), “Approximation with Kronecker Products,” in Linear Algebra for Large Scale and Real-Time Applications, eds. M. S. Moonen, G. H. Golub, and B. L. R. Moor, pp. 293–314, Dordrecht: Springer. [2570]
Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016), “Functional Data Analysis,” Annual Review of Statistics and Its Application, 3, 257–295. [2562]
Weidmann, J. (2012), Linear Operators in Hilbert Spaces (Vol. 68), New York: Springer. [2564,2567]
Wilmott, J. R., Andreev, K., Jdanov, D., Glei, D. A., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., and Vachon, P. (2007), “Methods Protocol for the Human Mortality Database,” University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock, 9, 10–11. http://mortality.org. [2571]
Yao, F., Müller, H.-G., and Wang, J.-L. (2005), “Functional Data Analysis for Sparse Longitudinal Data,” Journal of the American Statistical Association, 100, 577–590. [2568,2569]
Young, D. M. (2014), Iterative Solution of Large Linear Systems, New York: Elsevier. [2568]
Zhang, X., and Wang, J.-L. (2016), “From Sparse to Dense Functional Data and Beyond,” The Annals of Statistics 44, 2281–2321. [2568]