Thèse n° 9463

Covariance Estimation for Random Surfaces beyond Separability

Présentée le 12 juillet 2022

Faculté des sciences de base
Chaire de statistique mathématique
Programme doctoral en mathématiques

pour l'obtention du grade de Docteur ès Sciences

par

Tomas MASÁK

Acceptée sur proposition du jury

Prof. A. C. Davison, président du jury
Prof. V. Panaretos, directeur de thèse
Prof. J. Aston, rapporteur
Prof. F. Krahmer, rapporteur
Prof. S. Olhede, rapporteuse

École
polytechnique
fédérale
de Lausanne

2022

Contents

Acknowledgements **v**

Abstract (English and French) **vii**

Declaration of Authorship **xi**

Introduction **1**
Objectives of the Thesis 3
Notation and Organization 4

1 Background Concepts **5**
1.1 Operator Theory Basics 5
1.2 Product Hilbert Spaces 8
1.3 Separability 10
1.4 Random Elements and Integral Operators 13
1.5 Inference on Function Spaces 16
1.6 Functional Data and Discrete Observations 18
1.7 Marginalization Operators 22
1.7.1 Partial Tracing 22
1.7.2 Partial Inner Product 24
1.7.3 Examples with Matrix-variate Normal Distribution 25

2 Separable-plus-Banded Model **31**
2.1 Shifted Partial Tracing 34
2.1.1 General Definition 36
2.2 Estimation 42
2.2.1 Stationarity 45
2.2.2 Choice of Bandwidth 47
2.2.3 Goodness-of-fit Testing 49
2.3 Computational Considerations 51
2.3.1 Estimation Complexity 52
2.3.2 Fast Application and Norm Calculation 54
2.3.3 Inverse Problem 55
2.3.4 Ensuring Symmetry and Positive Semi-definiteness 59

2.4 Asymptotic Properties 60
2.4.1 Fully Observed Data 60
2.4.2 Discretely Observed Data 63
2.4.3 Adaptive Bandwidth 67
2.5 Empirical Demonstration 69
2.5.1 Simulation Study 69
2.5.2 Data Analysis: Mortality Rates 75

3 Separable Component Decomposition 79
3.1 Power Iteration Method 81
3.2 Partial Inner Product 83
3.2.1 Generalized Power Iteration 86
3.3 Estimation 90
3.3.1 Degree-of-Separability Selection 92
3.4 Inversion 94
3.4.1 Prediction 97
3.5 Asymptotic Properties 98
3.5.1 Complete Observations 98
3.5.2 Discrete Observations on a Grid 101
3.5.3 Unbalanced Design 103
3.6 Simulation Study 104
3.6.1 Parametric Covariance 105
3.6.2 Weakly Separable Covariance 108
3.6.3 Superposition of Independent Separable Processes 110
3.6.4 Random Covariance 111
3.7 Data Analysis: Classification of EEG Signals 113

4 Separability under Sparse Measurements 117
4.1 Model and Observation Scheme 118
4.2 Motivation 119
4.3 Estimation of the Model Components 121
4.3.1 Alternative Number of Steps 123
4.4 Weighting Scheme 125
4.5 Implementation Details 128
4.6 Prediction 132
4.7 Asymptotic Properties 134
4.8 Simulation Study 137
4.9 Data Analysis: Implied Volatility Surfaces 143
4.9.1 Quantitative Comparison 146

Discussion and Future Directions 151
Matrix Completion 153
Weak Separability and Weak Bandedness 154

Sparsely Observed Surface-valued Time Series 155
Statistical Tests 156
Domains of Higher Dimensions 157
Links between the Proposed Methods 159
Partial Tracing vs. Partial Inner Product 160
Pseudo-MLE Approach 160
Mixed Sampling of Random Surfaces 161

Appendix A: Proofs of the Asymptotic Results 165
Separable-plus-Banded Model 165
Proof of Theorem 4 165
Proof of Proposition 8 175
Proof of Theorem 5 177
Proof of Proposition 9 181
Proof of Theorem 6 181
Separable Component Decomposition 184
Proof of Theorem 7 184
Proof of Theorem 8 187
Proof of Theorem 9 193
Separability under Sparse Measurements 194
Explicit Formula for Local Polynomial Regression 194
Kernel Averages of m-dependent Data 195
Proof of Proposition 12 197
Proof of Theorem 10 and Corollary 4 199
Proof of Proposition 13 204
Proof of Theorem 11 204

Appendix B: `surfcov` package 207

Bibliography 211

List of Abbreviations 221

List of Figures 223

List of Statements 225

List of Tables 229

Curriculum Vitae 231

Acknowledgements

First and foremost, I would like to express my immense gratitude to Victor, who has been a model advisor, and his mentorship has influenced me far beyond the scope of academic training.

Secondly, I am grateful to the members of the thesis committee, namely to Professor Anthony Davison for acting as the president of the jury, Professors John Aston and Felix Krahmer for accepting the roles of the external experts, and Professor Sofia Olhede for accepting the role of the internal expert.

I would also like to thank Professors Anthony Davison, Daniel Kressner and Fritz Eisenbrandt for their kind advice and interesting discussions throughout my studies.

Many thanks go to the members of the group, both present and past, for countless professional and social interactions: Alessia, Kartik, Laya, Leonardo, Marie-Helene, Maroussia, Matthieu, Mikael, Neda, Shahin, Soham, Tomas R., Valentina, and Yoav. I am particularly grateful to Kartik for taking care of me (or my responsibilities) whenever I was sick or quarantined (or abroad), and to Soham and Tomas R. for their contributions to this thesis through the work we have done together. Also, I was lucky to interact with members of some of the other chairs at EPFL: Christoph, Emeric, Erwan, Guillaume, Hao, Igor, Jana, Jacques, Jon, Leo, Mario, Max, Moritz, Raphael, Sonia, Soumaya, Timmy, Yousra, Zhiwen, and many others. Furthermore, I am thankful to my good friends and colleagues outside of EPFL: Christian, Claudio, Martin, and Myhra.

Finally, I am grateful to my family, in particular to my parents and Renata, as well as my closest friends for their limitless support.

Lausanne, March 10, 2022 Tomas Masak

Abstract

This thesis focuses on non-parametric covariance estimation for random surfaces, i.e. functional data on a two-dimensional domain. Non-parametric covariance estimation lies at the heart of functional data analysis, and considerations of statistical and computational efficiency often compel the use of separability of the covariance, when working with random surfaces. We seek to provide efficient alternatives to this ambivalent assumption.

In Chapter 2, we study a setting where the covariance structure may fail to be separable locally – either due to noise contamination or due to the presence of a non-separable short-range dependent signal component. That is, the covariance is an additive perturbation of a separable component by a non-separable but banded component. We introduce non-parametric estimators hinging on shifted partial tracing – a novel concept enjoying strong denoising properties. We illustrate the usefulness of the proposed methodology on a data set of mortality surfaces.

In Chapter 3, we propose a distinctive decomposition of the covariance, which allows us to understand separability as an unconventional form of low-rankness. From this perspective, a separable covariance has rank one. Allowing for a higher rank suggests a structured class in which any covariance can be approximated up to an arbitrary precision. The key notion of the partial inner product allows us to generalize the power iteration method to general Hilbert spaces and estimate the aforementioned decomposition from data. Truncation and retention of the leading terms automatically induces a non-parametric estimator of the covariance, whose parsimony is dictated by the truncation level. Advantages of this approach, allowing for estimation beyond separability, are demonstrated on the task of classification of EEG signals.

While Chapters 2 and 3 propose several generalizations of separability in the densely sampled regime, Chapter 4 deals with the sparse regime, where the latent surfaces are observed only at few irregular locations. Here, a separable covariance estimator based on local linear smoothers is proposed, which is the first non-parametric utilization of separability in the sparse regime. The assumption of separability reduces the intrinsically four-dimensional smoothing problem into several two-dimensional smoothers and allows the proposed estimator to retain the classical minimax-optimal convergence rate for two-dimensional smoothers. The proposed methodology is used for a qualitative analysis of implied volatility surfaces corresponding to call options, and for prediction of the latent surfaces based on information from the entire data set, allowing for uncertainty quantification. Our quantitative results show that the proposed methodology outperforms the common approach of pre-smoothing every implied volatility surface separately.

Throughout the thesis, we put emphasis on computational aspects, since those are the main reason behind the immense popularity of separability. We show that the covariance structures of Chapters 2 and 3 come with no (asymptotic) computational overhead relative to assuming separability. In fact, the proposed covariance structures can be estimated and manipulated with the same asymptotic costs as the separable model. In particular, we develop numerical algorithms that can be used for efficient inversion, as required e.g. for prediction. All the methods are implemented in R and available on GitHub.

Keywords: functional data analysis, covariance operator, multi-dimensional domains, non-parametric modelling, dense and sparse sampling, shifted partial tracing, partial inner product.

Résumé

Le sujet de cette thèse est l'estimation non-paramétrique de la covariance sur des surfaces aléatoires, c'est-à-dire sur des domaines bidimentionnels. L'estimation non-paramétrique de la covariance est au coeur de l'analyse des données fonctionnelles. Des considérations d'efficacité statistique et computationelle obligent souvent à assumer que la covariance est séparable sur ces surfaces aléatoires. Notre but est de présenter des alternatives efficaces à cette hypothèse ambivalente.

Dans le chapitre 2, nous étudions un contexte où la structure de la covariance ne peut pas être localement séparée – cela peut être due à une contamination avec du bruit blanc aléatoire ou bien à la présence d'un signal dépendant de courte-portée qui est non-séparable. En d'autres termes, la covariance est la somme de perturbations d'éléments séparables par des éléments non-séparables mais banded. Nous présentons des estimateurs non-paramétriques basés sur le traçage partiel décalé, un concept nouveau qui a des propriétés importantes de débruitage. Nous illustrons l'utilité de la méthode proposée sur une base de données portant sur des surfaces de mortalité.

Dans le chapitre 3, nous proposons une décomposition distinctive de la covariance, qui nous permet d'exprimer la séparabilité comme étant une forme non-conventionnelle de rang faible. Avec ce changement de perspective, le rang d'une covariance séparable est égal à 1. Ainsi, en augmentant le rang, on définit une classe structurée dans laquelle toute covariance peut-être approximée avec précision. Le produit scalaire partiel nous permet d'étendre la méthode de la puissance itérée à des espaces de Hilbert généraux et d'estimer la décomposition à partir des données. La troncature et la rétention des termes principaux impliquent automatiquement un estimateur non-paramétrique de la covariance, dont la parsimonie est donc dictée par la niveau de troncature. Nous démontrons les avantages de cette approche en l'appliquant à la classification de signaux d'éléctroencéphalographie (EEG).

Tandis que les chapitre 2 et 3 sont centrés sur des généralisations de la séparabilité pour des régimes densément échantillonés, dans le chapitre 4, nous considérons des régimes épars, où les surfaces latentes ne sont observées qu'à quelques localisations irrégulières. Dans ce cas, nous proposons un estimateur séparable de la covariance basé sur un lisseur linéaire local. Il s'agit de la première utilisation non-paramétrique de la séparabilité dans un régime épars. L'hypothèse de séparabilité permet de réduire un problème de lissage qui est intrinsèquement quadrimensionnel à plusieurs lisseurs bidimensionnel. Ainsi, l'estimateur proposé garde la vitesse de convergence optimale-minimax d'un lisseur bidimentionnel. La méthode proposée est utilisée pour l'analyse qualitative de surfaces de volatilité implicites correspondant à des options d'achat, et pour la prédiction de surfaces latentes sur la base d'informations provenant de l'ensemble des données, permettant ainsi la quantification de l'incertitude. Nos résultats quantitatifs montrent que la méthodologie proposée est plus performante que l'approche commune de pré-lissage de chaque surface de volatilité implicite séparément.

Tout au long de la thèse, nous mettons l'accent sur les aspects computationnels, puisque ceux-ci sont la raison principale de l'immense popularité de la séparabilité. Nous montrons que les structures de covariance des chapitres 2 et 3 ne présentent aucun surcoût (asymptotique) de calcul par rapport à l'hypothèse de séparabilité. En fait, les structures de covariance proposées peuvent être estimées et manipulées avec les mêmes coûts asymptotiques qu'un modèle séparable. En particulier, nous développons des algorithmes numériques qui peuvent être utilisés pour une inversion efficace, comme cela est nécessaire par exemple pour la prédiction. Toutes les méthodes sont implémentées en R et disponibles sur GitHub.

Mots clefs : analyse des données fonctionnelles, opérateur de covariance, domaines multidimentionnels, modélisation non-paramétrique, échantillonnage dense et épars, traçage partiel décalé, produit scalaire partiel.

Declaration of Authorship

I declare that this thesis titled "Covariance Estimation for Random Surfaces beyond Separability" and the work presented in it are original. This work was done wholly while in candidature for a PhD degree at EPFL.
The thesis itself consists of four main chapters. The first chapter introduces the background. The remaining three chapters are based on three research papers, which are currently under review. The first of those is based on a work I did myself under the supervision of my advisor, Victor Panaretos. The second of those is based on a collaboration with Soham Sarkar, who developed the asymptotic theory. The final one is based on a collaboration with Tomas Rubin, who developed the asymptotic theory and took part in the data analysis. All other contributions in this thesis, including

- mathematical development of shifted partial tracing,
- methodology to estimate the separable-plus-banded model, including parameter selection procedures,
- computationally efficient inversion algorithm for the separable-plus-banded estimator,
- asymptotic theory for the separable-plus-banded model,
- analysis of the mortality rates data set,
- mathematical development of the partial inner product,
- methodology to estimate the separable component decomposition, including parameter selection procedures,
- computationally efficient inversion algorithm for the R-separable estimator,
- classification of EEG signals,
- methodology to estimate a separable model from sparse measurements,
- implementation of the aforementioned methodologies in R,
- conducting all the simulation studies, and
- producing all the graphical outputs,

are my own.

Lausanne, March 10, 2022 Tomas Masak

Introduction

This thesis studies the interlinked problems of parsimonious representation, efficient estimation, and tractable manipulation of a random surface's covariance, i.e. the covariance of a random process on a two-dimensional domain. We operate in the framework of *Functional Data Analysis* (FDA, Ramsay and Silverman, 2005; Hsing and Eubank, 2015), which focusses on the problem of statistical inference on the law of a random process $X(u) : [0,1]^D \to \mathbb{R}$ given multiple realisations thereof. The process realisations are treated as elements of a separable Hilbert space $\mathcal{H}$ of functions on $[0,1]^D$, e.g. $\mathcal{L}^2([0,1]^D)$. 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 $c(u_1, u_2) = \operatorname{cov}\Big(X(u_1), X(u_2)\Big)$ 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, efficient estimation of the covariance operator C (or equivalently its kernel c) is a fundamental task in FDA, on which further methodology can be based. This is to be done on the basis of i.i.d. realisations of the random process X, say $\{X_1, \dots, X_N\}$. One wishes to do so *nonparametrically*, since the availability of replicated realisations 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 et al., 2016, for a comprehensive overview).

Though conceptually similar, things are much less straightforward in the case of random surfaces, i.e. when $D = 2$, which is the focus of this thesis. We assume that we have access to (discretized) i.i.d. realizations $X_1, \dots, X_N$ of X and wish to estimate the covariance non-parametrically and computationally feasibly, ideally via a parsimonious representation allowing for computationally tractable further manipulations (e.g. inversion) required in key tasks commonly involving the covariance (e.g. regression, prediction, or classification).

In the case of a two-dimensional domain, 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 et al., 2017, 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 most common non-parametric estimator, the empirical covariance, much less to invert it. To appreciate this, assume that each of the N i.i.d. 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 $\mathbf{X}_n \in \mathbb{R}^{K_1 \times K_2}$ and the empirical covariance is represented by the tensor $\widehat{\mathbf{C}}_N \in \mathbb{R}^{K_1 \times K_2 \times K_1 \times K_2}$, which is a discretisation of the empirical covariance kernel. The covariance tensor $\widehat{\mathbf{C}}_N$ requires $\mathcal{O}(NK_1^2K_2^2)$ operations to be estimated and $\mathcal{O}(K_1^2K_2^2)$ memory to be stored. This becomes barely feasible on a regular computer with K_1 and K_2 as small as 100. Moreover, as Aston et al. (2017) note, the statistical constraints stemming from the need to accurately estimate $\mathcal{O}(K_1^2K_2^2)$ parameters contained in $\mathbf{C}$ (i.e. the discrete version of the covariance C) from only NK_1K_2 measurements are often even tighter than the computational constraints.

This dimensionality challenge is often dealt with by imposing additional structure. The most prevalent assumption is that of separability (Gneiting, 2002; Gneiting et al., 2006), which postulates a factorization of the covariance kernel c into two kernels, corresponding to the respective domains:

$$c(t,s,t',s') = c_1(t,t')c_2(s,s'), \qquad t,s,t',s' \in [0,1].$$

As a result, the four-dimensional non-parametric problem of estimating c simplifies into the problem of estimating a pair of two-dimensional objects. In the case of data observed on a grid, this reduces the number of parameters to be estimated from $\mathcal{O}(K_1^2K_2^2)$ down to $\mathcal{O}(K_1^2 + K_2^2)$. Moreover, both estimation and subsequent manipulation (for example inversion as required in prediction) of the covariance become computationally much simpler.

However, assuming separability often encompasses oversimplification and has undesirable practical implications for real data (see Rougier, 2017). In summary, separable covariances fail to capture any space-time interactions whatsoever. A number of tests for separability of covariance operators for functional data on a two-dimensional domain have been recently developed (Aston et al., 2017; Bagchi and Dette, 2020; Constantinou et al., 2017), and their applications have demonstrated that separability is distinctly violated for several data sets previously modelled as separable. Still, separability is often assumed in practice, not because it is believed to hold, but merely due to the computational gains it offers (see Gneiting et al., 2006; Genton, 2007; Pigoli et al., 2018). In fact, when the grid sizes K_1 and K_2 are large, it may not be possible to do away with separability due to the aforementioned computational limits.

Examples of random surfaces observed on a grid densely enough to impede the usage of an unstructured covariance arise abundantly for example in biomedical imaging, see Wang et al. (2016) for a review. If separability of the covariance is rejected for a data set at hand by one of the tests cited above, an alternative simplifying assumption enjoying similar computational advantages to separability is needed, but it is lacking in present literature.

Objectives of the Thesis

While many authors focus on testing separability (a list of up-to-date references was collected by Chen et al., 2021) or assessing the departures from it (Huang and Sun, 2019; Dette et al., 2020), little work has been done to offer non-parametric alternatives to the separable model. The main aspiration of this thesis is to provide such alternatives. We adopt separability as a building block and propose several ways of generalizing it. This results in the separable-plus-banded model for the covariance studied in Chapter 2, and the separable component decomposition introduced in Chapter 3. Both the separable-plus-banded model and the separable component decomposition (after a suitable truncation) lead to covariances, which can be estimated and manipulated with ease comparable to that of a separable model. Hence they can serve as viable alternatives to separability for a whole range of applications. The methodology and related numerical routines (in particular those allowing inversion of the proposed covariance structures) are implemented in an R (R Core Team, 2020) package `surfcov`, which is available on GitHub.

We stress out that all the standard operations with the proposed covariance structures, i.e. their estimation, application, and (numerical) inversion, can be performed with little computational overhead compared to the separable model. In fact, when data are sampled on a $K \times K$ grid, all of these operations can be performed at the same (asymptotic) cost in K as matrix-matrix multiplication between pairs of the sampled observations. We set this cubic time complexity in K and quadratic memory complexity in K as a firm computational limit, preventing our methods from, for example, ever explicitly computing the empirical covariance.

The perks and flaws of separability are well-known in the case of densely observed data. Chapters 2 and 3 focus on estimation beyond separability, retaining its advantages and mitigating its drawbacks. The separable component decomposition of Chapter 3 exemplifies that separability should be seen not as much as a crucial modelling assumption, but rather as a form of regularization, trading off between bias and variance as well as between simplistic statistical interpretation and enormous computational advantages. On the other hand, in the case of sparsely observed data, even a procedure for non-parametric estimation of a separable covariance has not been previously established. The objective of Chapter 4 is to devise such a procedure. We show that separability can be leveraged to reduce complexity of covariance estimation down to that of mean estimation in the sparse regime. Also, we argue that – due to extra costs associated with smoothing as well as higher statistical complexity stemming from sparse and noisy measurements – separability is an even more powerful assumption in the sparse regime, achieving a favorable bias-variance trade-off.

Notation and Organization

We typically use upper case letters such as A as the notation for operators on a general Hilbert space or Hilbert space with a continuous domain such as $\mathcal{L}^2([0,1])$, while lower case letters such as a are used to denote their kernels. On the other hand, when working specifically with a Hilbert space with a discrete domain, such as $\mathbb{R}^{K_1 \times K_2}$, we use boldface upper case letters, such as $\mathbf{A}$, to denote both the operator and its kernel, while boldface lower case letters are used to denote vectors, such as $\mathbf{a} \in \mathbb{R}^{K_1 K_2}$. In the discrete case, we utilize the Matlab notation when integrating over some dimensions, e.g. $\sum_{i=1,\ldots,K_1} \mathbf{A}[i,:]$ results in a vector in $\mathbb{R}^{K_2}$.

We use the triple bar to denote norm of an operator, such as $|||A|||_2$, while double bar is used to denote norm of an element, for example of the kernel $\|a\|_2$. In the discrete case $\|\mathbf{A}\|_F$ is used to denote the Frobenius norm. While in the previous example the norms are strongly related, this is not the case of the operator norm $|||A|||_\infty$ and the uniform (or supremum) norm $\|a\|_\infty$, which should not be confused.

We use the "o-times" symbol $\otimes$ to denote the abstract outer product, while $\otimes_K$ denotes the Kronecker product, see Remark 1. The integers N and K are reserved to denote the sample size and the grid size, respectively. The random variables are by default assumed to be centered (without the loss of generality), whenever no care is taken about their mean.

The four main chapters of this thesis are largely self-contained. Chapter 1 introducing the background can be skipped by an experienced reader. On the other hand, reading the thesis chronologically has indisputable advantages, most importantly Chapter 4 softly builds upon the development in Chapter 3. The concluding chapter contains many potential directions for future research as its subsections.

Shorter proofs usually follow the respective statements. Contrarily, longer proofs of the asymptotic results are deferred to the appendix. List of Statements, located towards the end of the thesis, can be used for navigation, in particular to track down proofs in the appendix easily.

1 Background Concepts

This thesis operates in the framework of functional data analysis (FDA). As the name suggests, FDA seeks to utilize ideas from functional analysis – such as treating *functions as points* in certain spaces, while considering operators as functions on these points – in order to generalize statistical methods for random vectors to more complex, continuous structures, for instance random curves or random surfaces. These structures are in practice observed only discretely, leading to vector or matrix observations. However, the assumption of existence of a latent continuous object and the emphasis put on the object as a whole is an emblematic feature of FDA and separates the FDA approach from that of multivariate statistics.

In this chapter, we review some basics of operator theory, construct product Hilbert spaces, discuss different observation and asymptotic regimes for functional data, and introduce tools that can be used to construct proxies of operators on product Hilbert spaces as products of mxirginals. The purpose of this chapter is to present – from a certain point of view – the background concepts, which are important for the work presented in the remainder of this thesis. For a more complete exposition of FDA, we refer the reader to Hsing and Eubank (2015), Ramsay and Silverman (2005), or Ferraty and Vieu (2006). Furthermore, Haase (2014) and Young (1988) give overview of the elementary ideas of functional analysis that are particularly useful in FDA.

1.1 Operator Theory Basics

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces equipped with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_2}$, and corresponding norms $\| \cdot \|_{\mathcal{H}_1}$ and $\| \cdot \|_{\mathcal{H}_2}$, respectively. Throughout the thesis, we work with real Hilbert spaces with countable orthonormal bases.

A linear transformation $F : \mathcal{H}_1 \to \mathcal{H}_2$ is *bounded* if

$$|||F|||_\infty := \sup_{\|x\|_{\mathcal{H}_1}=1} \|Fx\|_{\mathcal{H}_2} < \infty \,.$$

Bounded linear transformations are called *operators.*

The set of operators from H_1 to $\mathcal{H}_2$ equipped with the *operator norm* $|||\cdot|||_\infty$ is a Banach space, denoted by $\mathcal{S}_\infty(\mathcal{H}_1, \mathcal{H}_2)$. When $\mathcal{H}_1 = \mathcal{H}_2$, we abbreviate $\mathcal{S}_\infty(\mathcal{H}_1, \mathcal{H}_2)$ to $\mathcal{S}_\infty(\mathcal{H}_1)$, which is the set of operators on $\mathcal{H}_1$.

For $F \in \mathcal{S}_\infty(\mathcal{H}_1, \mathcal{H}_2)$, the unique operator $F^* \in \mathcal{S}_\infty(\mathcal{H}_2, \mathcal{H}_1)$ determined by the relation

$$\langle Fx, y\rangle = \langle x, F^* y\rangle, \qquad \forall x \in \mathcal{H}_1, y \in \mathcal{H}_2,$$

is called the *adjoint* of F. When $F \in \mathcal{S}_\infty(\mathcal{H}_1)$ and $F^* = F$, then F is called *self-adjoint.*

An, operator $F \in \mathcal{S}_\infty(\mathcal{H}_1, \mathcal{H}_2)$ is said to be *compact* if for any bounded sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{H}_1$, the sequence $\{Fx_n\}_{n\in\mathbb{N}} \subset \mathcal{H}_2$ contains a convergent subsequence. Bijective operators cannot be compact, unless $\mathcal{H}_1$ and $\mathcal{H}_2$ are finite-dimensional. However, compact operators admit eigendecomposition or singular value decomposition (SVD). As a result, they can be well-approximated by finite-dimensional operators, which makes them not too dissimilar from matrices.

Let $F \in \mathcal{S}_\infty(\mathcal{H}_1)$ be compact and self-adjoint. Then its action on $x \in \mathcal{H}_1$ can be written as

$$Fx = \sum_{j=1}^{\infty} \lambda_j \langle x, g_j\rangle g_j, \tag{1.1}$$

where $\{\lambda_j\}_{j\in\mathbb{N}}$ is the sequence of eigenvalues and $\{g_j\}_{j\in\mathbb{N}}$ is the sequence of eigenvectors of F. The sequence of eigenvalues is non-increasing in absolute value, and the sequence of eigenvectors form an orthonormal basis (ONB) of $\mathcal{H}_1$. When $F \in \mathcal{S}_\infty(\mathcal{H}_1, \mathcal{H}_2)$ is *compact*, but not necessarily self-adjoint, its action on an arbitrary $x \in \mathcal{H}_1$ can be written as

$$Fx = \sum_{j=1}^{\infty} \sigma_j \langle f_j, x\rangle e_j \,, \tag{1.2}$$

where $\{\sigma_j\}_{j=1}^{\infty}$ is a non-negative and non-increasing sequence of singular values, and $\{e_j\}_{j=1}^{\infty}$ and $\{f_j\}_{j=1}^{\infty}$ are orthonormal bases of $\mathcal{H}_1$ and $\mathcal{H}_2$. In this case, $\{e_j\}_{j\in\mathbb{N}}$ is the sequence of eigenvectors of F^*F, $\{f_j\}_{j\in\mathbb{N}}$ is the sequence of eigenvectors of FF^*, and $\{\sigma_j^2\}_{j\in\mathbb{N}}$ is the sequence of eigenvalues of both F^*F and FF^*.

When $F \in \mathcal{S}_\infty(\mathcal{H}_1)$ is compact and self-adjoint the singular value decomposition and eigendecomposition above coincide up to potential change in signs: if for $j \in \mathbb{N}$ we have $\lambda_j < 0$, we can take for example $g_j =: e_j =: -f_j$. When $\sigma_j = 0$ for all $j > R$ in (1.2), we say that the rank of F is r (or that F is rank-R, in short) and write $\text{rank}(F) = R$.

Still, a distinctive feature of matrices is that they can be thought of both as linear transformations between two vector spaces and elements of a (larger, product) vector space. To have something similar for operators on Hilbert spaces, we need a further condition on the set of singular values and the notion of product Hilbert space.

Let $p \in [1, \infty)$. Let $F : \mathcal{H}_2 \to \mathcal{H}_2$ be a compact operator such that

$$\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert F \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_p := \left(\sum_{j=1}^{\infty} \sigma_j^p \right)^{\frac{1}{p}} < \infty \,.$$

The set of all such operators is denoted by $\mathcal{S}_p(\mathcal{H}_1, \mathcal{H}_2)$, and it is a Banach space for given $p \in [1, \infty)$, when equipped with the *Schatten-p norm* $|||\cdot|||_p$. It holds that $\mathcal{S}_p(\mathcal{H}_1, \mathcal{H}_2) \subset \mathcal{S}_q(\mathcal{H}_1, \mathcal{H}_2)$ for $p < q$. We further abbreviate $\mathcal{S}_p(\mathcal{H}_1, \mathcal{H}_1) =: \mathcal{S}_p(\mathcal{H}_1)$ and denote $\mathcal{S}_p^+(\mathcal{H}_1)$ the set of all *positive semi-definite* (PSD) operators (i.e. self-adjoint operators with non-negative eigenvalues) that belong to $\mathcal{S}_p(\mathcal{H}_1)$.

We are particularly interested in the cases $p = 1$ and $p = 2$. Firstly, $\mathcal{S}_1(\mathcal{H}_1, \mathcal{H}_2)$ is the space of *trace-class* operators, and $|||\cdot|||_1$ is called *trace norm* or *nuclear norm*. For $F \in \mathcal{S}_1(\mathcal{H}_1, \mathcal{H}_2)$ we define its *trace* as

$$\mathrm{Tr}(F) := \textstyle\sum_{j=1}^{\infty} \langle (T^*T)^{1/2} e_j, e_j \rangle_{\mathcal{H}_1},$$

where $\{e_j\}$ is an orthonormal basis of $\mathcal{H}_1$. When F is positive semi-definite, then $\mathrm{Tr}(F) = |||F|||_1$. Secondly, $\mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_2)$ is the space of *Hilbert-Schmidt* operators, and it is a complete separable Hilbert space when equipped with the *Hilbert-Schmidt inner product*

$$\langle F_1, F_2 \rangle_{HS} := \sum_{j=1}^{\infty} \langle F_1 e_j, F_2 e_j \rangle_{\mathcal{H}_2} \,, \quad \forall F_1, F_2 \in \mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_2) \,.$$

Let $a \in \mathcal{H}_1$ and $b \in \mathcal{H}_2$. We define the tensor product operators $(a \otimes_1 b) : \mathcal{H}_1 \to \mathcal{H}_2$ and $(a \otimes_2 b) : \mathcal{H}_2 \to \mathcal{H}_1$ by

$$\begin{aligned} (a \otimes_1 b)x &= \langle a, x \rangle_{\mathcal{H}_1} b \,, \quad \forall x \in \mathcal{H}_1 \,, \\ (a \otimes_2 b)y &= \langle b, y \rangle_{\mathcal{H}_2} a \,, \quad \forall y \in \mathcal{H}_1 \,. \end{aligned} \tag{1.3}$$

If $\{e_j\}$ and $\{f_j\}$ are orthonormal bases in $\mathcal{H}_1$ and $\mathcal{H}_2$, then $\{e_i \otimes_1 f_j\}_{i,j=1}^{\infty}$ is an orthonormal basis of $\mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_2)$ and $\{e_i \otimes_2 f_j\}_{i,j=1}^{\infty}$ is an orthonormal basis of $\mathcal{S}_2(\mathcal{H}_2, \mathcal{H}_1)$ (Hsing and Eubank, 2015, Thm. 4.4.5).

Using the tensor product operators, we can now write the eigendecomposition (1.1) as

$$Fx = \Big(\sum_{j=1}^{\infty} \sigma_j (g_j \otimes_2 g_j) \Big) x$$

and the singular value decomposition (1.2) as

$$Fx = \Big(\sum_{j=1}^{\infty} \sigma_j (e_j \otimes_2 f_j) \Big) x,$$

so we can directly write e.g.

$$F = \sum_{j=1}^{\infty} \sigma_j (e_j \otimes_2 f_j) \tag{1.4}$$

with the series converging in the operator norm.

The eigendecomposition and the singular value decompositions enjoy certain optimality properties, see Hsing and Eubank (2015, Section 4.2) for details. Most importantly, they provide the optimal low-rank approximation of an operator. For $F \in \mathcal{S}_p(\mathcal{H}_1, \mathcal{H}_2)$ with the singular value decomposition (1.4), the optimum of the following minimization problem

$$\min_{G:\text{rank}(G) \leq R} |||F - G|||_p \tag{1.5}$$

is attained at $G = \sum_{j=1}^{R} \sigma_j (e_j \otimes_2 f_j)$. The same holds more generally for a compact F, when the Schatten-p norm in (1.5) is replaced by the operator norm $|||\cdot|||_\infty$.

1.2 Product Hilbert Spaces

We define the tensor product space of $\mathcal{H}_1$ and $\mathcal{H}_2$, denoted by $\mathcal{H} := \mathcal{H}_1 \otimes \mathcal{H}_2$, as the completion of the set of finite linear combinations of abstract tensor products

$$\Big\{ \sum_{j=1}^{N} x_j \otimes y_j \, ; \, x_j \in \mathcal{H}_1, y_j \in \mathcal{H}_2, N \in \mathbb{N} \Big\} \tag{1.6}$$

under the inner product $\langle x_1 \otimes y_1, x_2 \otimes y_2 \rangle_{\mathcal{H}} := \langle x_1, x_2 \rangle_{\mathcal{H}_1} \langle y_1, y_2 \rangle_{\mathcal{H}_2}$, for all $x_1, x_2 \in \mathcal{H}_1$ and $y_1, y_2 \in \mathcal{H}_2$ (cf. Weidmann, 2012). More precisely, set (1.6) is a vector space equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Thus its completion H is a Hilbert space. Again, if $\{e_j\}$ and $\{f_j\}$ are orthonormal bases in $\mathcal{H}_1$ and $\mathcal{H}_2$, then $\{e_i \otimes f_j\}_{i,j=1}^{\infty}$ is an orthonormal basis of $\mathcal{H}$.

The Hilbert space $\mathcal{H}$ is isometrically isomorphic to $\mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_2)$ and also to $\mathcal{S}_2(\mathcal{H}_2, \mathcal{H}_1)$. The respective isomorphisms are the linear mappings $\Phi_1 : \mathcal{H}_1 \otimes \mathcal{H}_2 \to \mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_2)$ and $\Phi_2 : \mathcal{H}_1 \otimes \mathcal{H}_2 \to \mathcal{S}_2(\mathcal{H}_2, \mathcal{H}_1)$, which are defined on the abstract tensor products as

$$\begin{aligned} &\Phi_1(x \otimes y) = x \otimes_1 y \, , \\ &\Phi_2(x \otimes y) = x \otimes_2 y \, , \quad \forall x \in \mathcal{H}_1, y \in \mathcal{H}_2 \, . \end{aligned}$$

Now we are ready to draw the connection between matrices and Hilbert-Schmidt operators. For $F \in \mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_2)$, the operator decomposition (1.4) is isometrically isomorphic to the following “element” decomposition:

$$F = \sum_{j=1}^{\infty} \sigma_j (e_j \otimes f_j),$$

where the series converges in the norm of $\mathcal{H}_1 \otimes \mathcal{H}_2$.

Hence Hilbert-Schmidt operators can be thought of both as linear transformations between two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, and elements of the product Hilbert space $\mathcal{H}_1 \otimes \mathcal{H}_2$. In the latter point of view, the singular value decomposition is in fact a decomposition of an element on the product Hilbert space w.r.t. a product basis given by the singular vectors.

The previous construction of product Hilbert spaces can be generalized to Banach spaces $\mathcal{B}_1$ and $\mathcal{B}_2$. That is, one can define $\mathcal{B} := \mathcal{B}_1 \otimes \mathcal{B}_2$ in a similar way. The only difference is that the completion is done under the norm $\|x \otimes y\|_{\mathcal{B}} := \|x\|_{\mathcal{B}_1} \|y\|_{\mathcal{B}_2}$, for $x \in \mathcal{B}_1$ and $y \in \mathcal{B}_2$.

Consider now $\mathcal{B}_1 := \mathcal{S}_p(\mathcal{H}_1)$ and $\mathcal{B}_2 := \mathcal{S}_p(\mathcal{H}_2)$, and construct the abstract tensor product space $\mathcal{S}_p(\mathcal{H}_1) \otimes \mathcal{S}_p(\mathcal{H}_2)$ as described above. Now, consider the linear map $\Phi : \mathcal{S}_p(\mathcal{H}_1) \otimes \mathcal{S}_p(\mathcal{H}_2) \to \mathcal{S}_p(\mathcal{H})$ defined on the abstract tensor products as

$$\Phi(A \otimes B) = A \tilde{\otimes} B\,, \quad A \in \mathcal{S}_p(\mathcal{H}_1), B \in \mathcal{S}_p(\mathcal{H}_2)\,,$$

where $A \tilde{\otimes} B : \mathcal{H} \to \mathcal{H}$ is the linear operator defined on the abstract tensor products in $\mathcal{H}$ as

$$(A \tilde{\otimes} B)(x \otimes y) = Ax \otimes By\,, \quad x \in \mathcal{H}_2, y \in \mathcal{H}_2. \tag{1.7}$$

The mapping Φ is an isomorphism between $\mathcal{S}_p(\mathcal{H}_1) \otimes \mathcal{S}_p(\mathcal{H}_2)$ and $\mathcal{S}_p(\mathcal{H})$. Thus we showed that the space of Schatten-p operators on an abstract tensor product space is isometrically isomorphic to the abstract tensor product of two Schatten-p operator spaces. In the following, we will prefer the former point of view, and $A \tilde{\otimes} B \in \mathcal{S}_p(\mathcal{H})$ will denote the unique operator satisfying (1.7) for $A \in \mathcal{S}_p(\mathcal{H}_1)$ and $B \in \mathcal{S}_p(\mathcal{H}_2)$. By the abstract construction we also have $|||A \tilde{\otimes} B|||_p = |||A|||_p |||B|||_p$.

Notice the difference between $A \tilde{\otimes} B$ and $A \otimes B$, the former being an operator while the latter being an element. We have introduced the “otimes-tilde” symbol $\tilde{\otimes}$ following Aston et al. (2017) to make the distinction, which is similar to the difference between $x \otimes_1 y$ and $x \otimes y$ discussed above. Only this time we have the isomorphism for all Schatten-p spaces, not just Hilbert-Schmidt operators. This is only possible because both $\mathcal{S}_p(\mathcal{H}_1) \otimes \mathcal{S}_p(\mathcal{H}_2)$ and $\mathcal{S}_p(\mathcal{H}_1 \otimes \mathcal{H}_2)$ include the abstract tensor product construction.

Remark 1. *Many authors overuse the “otimes” symbol* $\otimes$*, either using it for the Kronecker product in finite dimensions or dropping the subscripts from* $\otimes_1$ *or* $\otimes_2$*. As* *described above, not making the distinction between* $\otimes$ *and* $\otimes_1$ *is quite natural, even though formally justifiable via isometry only when working with Hilbert-Schmidt operators or product Hilbert spaces.*

For example, it is easy to verify from the definitions that

$$(x_1 \otimes_2 x_2) \tilde{\otimes} (y_1 \otimes_2 y_2) = (x_1 \otimes y_1) \otimes_2 (x_2 \otimes y_2). \tag{1.8}$$

Dropping the subscript from $\otimes_2$ *in the previous equation now shows that using* $\tilde{\otimes}$ *entails in practice a permutation of the dimensions. For example, for* $\mathbf{A} \in \mathbb{R}^{K_1 \times K_1}$ *and* $\mathbf{B} \in \mathbb{R}^{K_2 \times K_2}$, $\mathbf{A} \tilde{\otimes} \mathbf{B} \in \mathbb{R}^{K_1 \times K_2 \times K_1 \times K_2}$ *while* $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{K_1 \times K_1 \times K_2 \times K_2}$. *One needs to be careful about such permutations when implementing the methodology developed in this thesis.*

The symbol $\otimes$ *is also used in linear algebra for the Kronecker product, which we denote by* $\otimes_K$ *in this thesis. The following relation between the Kronecker product and the abstract outer product holds in the case of finite dimensional spaces:*

$$\mathrm{vec}((\mathbf{A} \tilde{\otimes} \mathbf{B})\mathbf{X}) = (\mathbf{B}^\top \otimes_K \mathbf{A})\mathbf{x}, \tag{1.9}$$

where $\mathbf{x} = \mathrm{vec}(\mathbf{X})$ *is the vectorization of matrix* $\mathbf{X}$, *and* $\mathrm{vec}(\cdot)$ *is the vectorization operator, stacking all columns of a matrix into a long vector (cf. Van Loan and Golub, 1983). In our opinion, at least some – if not all – sub-fields of linear algebra such as the rapidly growing literature on tensor decompositions (see Kolda and Bader, 2009, for an overview) would greatly benefit from refraining from the Kronecker product, replacing it with the abstract outer product instead. However, the Kronecker product is now so ubiquitous in linear algebra (as pertinently captured in the title of Van Loan, 2000) that we also use it in this thesis, whenever we borrow ideas from that field.*

1.3 Separability

Now we are ready to define *separability* of an operator on a product Hilbert space.

Definition 1. *Let* $\mathcal{H}_1$ *and* $\mathcal{H}_2$ *be Hilbert spaces and* $\mathcal{H} := \mathcal{H}_1 \otimes \mathcal{H}_2$. *An operator* $F \in \mathcal{S}_p(\mathcal{H})$ *is called separable if* $F = A \tilde{\otimes} B$ *for some* $A \in \mathcal{S}_p(\mathcal{H}_1)$ *and* $B \in \mathcal{S}_p(\mathcal{H}_2)$. *If* F *is not separable, we call it entangled.*

The relationship between the spectra of A and B and the spectrum of $A \tilde{\otimes} B$ is particularly simple.

Lemma 1. *Let* $A \in \mathcal{S}_p(\mathcal{H}_1)$ *and* $B \in \mathcal{S}_p(\mathcal{H}_2)$ *be self-adjoint with eigenvalue-eigenvector pairs* $\{(\lambda_j, e_j)\}$ *and* $\{(\rho_j, f_j)\}$. *Then* $A \tilde{\otimes} B$ *is self-adjoint with eigenvalue-eigenvector pairs* $\{(\lambda_i \rho_j, e_i \otimes f_j)\}_{i,j=1}^{\infty}$. *Furthermore, for* $p = 1$, *it holds* $Tr(A \tilde{\otimes} B) = Tr(A)\, Tr(B)$.

Proof. Let $x \in \mathcal{H}_1$ and $y \in \mathcal{H}_2$, then

$$\begin{aligned}(A \,\tilde{\otimes}\, B)(x \otimes y) &= \left[\Big(\sum_{j=1}^{\infty} \lambda_j e_j \otimes e_j\Big) \,\tilde{\otimes}\, \Big(\sum_{j=1}^{\infty} \rho_j f_j \otimes f_j\Big)\right](x \otimes y) \\ &= \Big(\sum_{j=1}^{\infty} \lambda_j e_j \otimes e_j\Big) x \otimes \Big(\sum_{j=1}^{\infty} \rho_j f_j \otimes f_j\Big) y \\ &= \left[\Big(\sum_{j=1}^{\infty} \lambda_j \langle e_j, x\rangle_{\mathcal{H}_1} e_j\Big) \otimes \Big(\sum_{j=1}^{\infty} \rho_j \langle f_j, y\rangle_{\mathcal{H}_1} f_j\Big)\right].\end{aligned}$$

For the choice of $x = e_k$ and $y = f_l$ for $k, l \in \mathbb{N}$ we have

$$(A \,\tilde{\otimes}\, B)(e_k \otimes f_l) = \lambda_k e_k \otimes \rho_l f_l = \lambda_k \rho_l (e_k \otimes f_l)\,,$$

which shows that $e_k \otimes f_l$ is an eigenvector of $A \,\tilde{\otimes}\, B$ associated with the eigenvalue $\lambda_k \rho_l$.

The additional part follows from the previous one, since by Fubini's theorem

$$\sum_{i,j=1}^{\infty} \lambda_i \rho_j = \Big(\sum_{i=1}^{\infty} \lambda_i\Big)\Big(\sum_{j=1}^{\infty} \rho_j\Big). \qquad \square$$

The previous lemma can be naturally extended to singular values and singular vectors of operators, which are not self-adjoint, only the notation gets little more complicated. Moreover, we have the following characterization.

Corollary 1. *A separable operator $A \,\tilde{\otimes}\, B$ is self-adjoint if and only if both A and B are self-adjoint. Provided the largest eigenvalues of A and B are positive, the equivalence holds also when self-adjointness is replaced by positive semi-definiteness or positive definiteness.*

Proof. It is trivial that $A \,\tilde{\otimes}\, B$ must be self-adjoint for A and B both self-adjoint.

In the other direction, assume that both A and B are non-zero, otherwise $A \,\tilde{\otimes}\, B$ is zero (because $\langle (A \,\tilde{\otimes}\, B)(u \otimes v), u \otimes v\rangle = 0$ for all u, v) and the conclusion is trivial. Note that self-adjointness of $A \,\tilde{\otimes}\, B$ gives us

$$\begin{aligned}&\langle (A \,\tilde{\otimes}\, B)(x_1 \otimes y_1), x_2 \otimes y_2\rangle = \langle x_1 \otimes y_1, (A \,\tilde{\otimes}\, B)(x_2 \otimes y_2)\rangle \\ \Rightarrow\quad &\langle A x_1, y_1\rangle \langle B x_2, y_2\rangle = \langle x_1, A y_1\rangle \langle x_2, B y_2\rangle\end{aligned}$$

for $x_1, x_2 \in \mathcal{H}_1$ and $y_1, y_2 \in \mathcal{H}_2$ arbitrary. Now choose (x_2, y_2) to be the left-right eigenvector pair of B associated with a non-negative eigenvalue. We immediately see that A must be self-adjoint, and similarly for B.

The parts about positive semi-definiteness and positive definiteness follow easily from the previous lemma. The additional assumption on a positive eigenvalue is to prevent a change in signs. For example, for A and B negative definite, $A \tilde{\otimes} B$ must be positive definite by the previous lemma. □

In practice, we work with finite-dimensional Hilbert spaces, e.g. $\mathcal{H}_1 = \mathcal{H}_2 = \mathbb{R}^K$, leading to $\mathcal{H} = \mathbb{R}^{K \times K}$ and $\mathcal{S}_p(\mathcal{H}) = \mathbb{R}^{K \times K \times K \times K}$ in Definition 1. In finite dimensions, all the Schatten-p classes naturally coincide, since the number of singular values of an operator is finite in this case. Here, the operator $\mathbf{F} \in \mathbb{R}^{K \times K \times K \times K}$ is a tensor of order four, acting on $K \times K$ matrices, and it is separable if and only if it can be written as $\mathbf{F} = \mathbf{A} \tilde{\otimes} \mathbf{B}$ for some matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{K \times K}$. Entry-wise, this means $\mathbf{F}[i,j,k,l] = \mathbf{A}[i,k]\mathbf{B}[j,l]$ for $i,k = 1, \ldots, K_1$ and $j,l = 1, \ldots, K_2$.

Our task will be estimation of an operator $\mathbf{F}$. If it is separable, the estimation problem becomes computationally much simpler, both from the statistical perspective (a separable $\mathbf{F}$ only has $2K^2$ degrees of freedom while a general $\mathbf{F}$ has K^4 degrees of freedom to be estimated) and the computational perspective (the smaller degrees of freedom naturally correspond to lower storage requirements). Moreover, the following two properties hold for matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{X}$ of appropriate sizes:

$$\begin{aligned} (\mathbf{A} \tilde{\otimes} \mathbf{B})\mathbf{X} &= \mathbf{A}\mathbf{X}\mathbf{B}, \\ (\mathbf{A} \tilde{\otimes} \mathbf{B})^{-1} &= \mathbf{A}^{-1} \tilde{\otimes} \mathbf{B}^{-1}. \end{aligned} \tag{1.10}$$

These two properties are among the core reasons for the popularity of the separability assumption in the space-time processes literature (Gneiting et al., 2006; Genton, 2007), because they allow to apply a separable covariance fast ($\mathcal{O}(K^3)$ instead of $\mathcal{O}(K^4)$ operations) and solve an inverse problem involving the covariance fast ($\mathcal{O}(K^3)$ instead of $\mathcal{O}(K^6)$ operations).

On the other hand, separability has been widely criticized for decades as an oversimplification, mostly because it has implications hard to justify in many, if not most, applied problems. For example, separability postulates that

$$\mathrm{Cov}\Big(X(t,s), X(t',s')\Big| X(t',s)\Big) = \mathrm{Cov}\Big(X(t,s), X(t',s')\Big| X(t,s')\Big) = 0.$$

In other words, separability fails to model any space-time interactions. We refer the reader to Rougier (2017) for a thorough discussion on the questionable ramifications of separability.

Also, consider a process X with a separable covariance observed on a grid not directly, but under additional white noise, i.e.

$$\mathbf{Y}[i,j] = \mathbf{X}[i,j] + \epsilon_{i,j}, \qquad i,j = 1, \ldots, K,$$

where $\epsilon_{i,j}$'s are zero-mean, unit-variance, and independent random variables, also independent of $\mathbf{X}$. Then we naturally have

$$\operatorname{Cov}(\mathbf{Y}) = \operatorname{Cov}(\mathbf{X}) + \mathbf{I} = \mathbf{A} \mathbin{\tilde{\otimes}} \mathbf{B} + \mathbf{I},$$

where $\mathbf{I} \in \mathbb{R}^{K\times K\times K\times K}$ is the identity tensor. Hence the covariance of $\mathbf{Y}$ cannot be separable, unless $\mathbf{X}$ itself is white noise. In other words, separability can be disrupted just by white noise errors in the data.

1.4 Random Elements and Integral Operators

The most prevalent theoretical model for functional data is a combination of the functional-analytic perspective in which functional data are realizations of an abstract random variable taking values in a Hilbert space, and the stochastic process perspective in which the functional data are sample paths of a mean-square continuous stochastic process with continuous sample paths. Occasionally, we would like to consider the observations as pointwise evaluations of a random element X, which takes values in a Hilbert space, say $\mathcal{L}^2([0,1]^D)$. But in order to make this pointwise evaluation meaningful, we will consider X to be a stochastic process, i.e.

$$X(\mathbf{u},\omega) : [0,1]^D \times (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}\,,$$

such that $X(\mathbf{u},\omega)$ is a random variable for any $\mathbf{u} \in [0,1]^D$. Then X may not be a random element of $\mathcal{L}^2([0,1]^D)$, because that requires joint measurability. However, by Theorem 7.4.2 of Hsing and Eubank (2015), we have joint measurability provided X is a mean-square continuous process with continuous sample paths. We will impose these two assumptions where needed, i.e. when we want to think of discretely observed data as pointwise evaluations of an $\mathcal{L}^2([0,1]^D)$-valued random element and a continuous stochastic process at the same time.

We will be interested in studying the second-order properties of X, that is the mean and the covariance, hence we assume that $\mathbb{E}\|X\|^2 < \infty$. From the random element viewpoint, these are defined as

$$m = \mathbb{E}X \quad \text{and} \quad C = \mathbb{E}[(X-m)\otimes(X-m)]$$

where the integrals (expectations) are defined in the Bochner sense. From the random process viewpoint, mean and covariance are defined as

$$m(\mathbf{u}) = \mathbb{E}X(\mathbf{u}) \quad \text{and} \quad c(\mathbf{u}_1,\mathbf{u}_2) = \mathbb{E}[(X(\mathbf{u}_1) - m(\mathbf{u}_1)(X(\mathbf{u}_2) - m(\mathbf{u}_2))].$$

Under the continuity assumptions above, the definitions of m coincide, and the covariance operator C is related to the covariance kernel c by

$$(Cf)(\mathbf{u}_1) = \int_{[0,1]^D} c(\mathbf{u}_1, \mathbf{u}_2) f(\mathbf{u}_2) \,\mathrm{d}\mathbf{u}_2 \quad \forall f \in \mathcal{L}^2([0,1]^D)\,.$$

The covariance C is a positive semi-definite, trace-class operator on $\mathcal{L}^2([0,1]^D)$, and its trace captures the total variance of the random element:

$$\mathrm{Tr}(C) = \mathbb{E}\Big[\mathrm{Tr}\Big((X-m) \otimes_2 (X-m)\Big)\Big] = \mathbb{E}\|X-m\|^2.$$

The covariance plays a central role in FDA (Aneiros et al., 2019). For example, uncertainty quantification about the mean requires correct recovery of the dependency structure, which is encoded in the covariance. Also, eigenfunctions of the covariance offer low-rank representation of data via the Karhunen-Loève expansion. A low-rank representation is particularly important for functional data, which are intrinsically infinite-dimensional. A large body of work in FDA is focused on non-parametric estimation of the mean and the covariance (see Li and Hsing, 2010, and references therein), while functional principal component analysis (PCA) based on non-parametric covariance estimators has become one of the most common approaches in FDA (Mueller, 2016), used for curve interpolation (Yao et al., 2005a), functional generalized regression models (Yao et al., 2005b; Müller and Stadtmüller, 2005), functional clustering and classification (Delaigle et al., 2012), or vector autoregression approach to functional times series (Aue et al., 2015).

As described below, $C \in \mathcal{S}_1(\mathcal{L}^2([0,1]^D))$ can be identified with an element of $\mathcal{L}^2([0,1]^{2D})$. In this thesis, we will be interested solely in the two-dimensional case, i.e. $D = 2$. We adopt the view that the first dimension corresponds to *time*, always denoted by variable t, and the second dimension corresponds to *space*, always denoted by variable s. This is only for the ease of presentation: both dimensions may very well be temporal or spatial, or there may be no space/time interpretation for one or both dimensions.

Note that all trace-class operators are Hilbert-Schmidt, and all Hilbert-Schmidt operators (on $\mathcal{L}^2(E)$ for $E \subset \mathbb{R}^D$ compact) are in turn integral operators. This means that for $A \in \mathcal{S}_2(\mathcal{L}^2([0,1]^2))$ there exists a kernel $a \in \mathcal{L}^2([0,1]^4)$ such that

$$(Af)(t,t') = \int_0^1 \int_0^1 c(t,s,t',s') f(s,s') \,\mathrm{d}s \,\mathrm{d}s' \quad \forall f \in \mathcal{L}^2([0,1]^2)\,.$$

On the other hand, A is fully characterized by its kernel a. In fact, there is an isometry between the space of Hilbert-Schmidt operators $\mathcal{S}_2(\mathcal{L}^2([0,1]^D))$ and the space of kernels $\mathcal{L}^2([0,1]^{2D})$. This has many implications. For example, $A \in \mathcal{S}_2(\mathcal{L}^2([0,1]^2))$ is positive semi-definite if and only if a is positive semi-definite. Secondly, we have

$$|||A|||_2^2 = \int_0^1 \int_0^1 \int_0^1 \int_0^1 \Big(a(t,s,t',s')\Big)^2 \,\mathrm{d}t \,\mathrm{d}s \,\mathrm{d}t' \,\mathrm{d}s',$$

and for $A, B \in \mathcal{S}_2(\mathcal{L}^2([0,1]^2))$, the inner product can be calculated as

$$\langle A, B\rangle_{HS} = \int_0^1 \int_0^1 \int_0^1 \int_0^1 a(t,s,t',s')b(t,s,t',s')\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}t'\,\mathrm{d}s'. \tag{1.11}$$

For $A \in \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$ with a kernel a, the trace can be calculated as (Gohberg and Krein, 1978)

$$\mathrm{Tr}(A) = \lim_{h\to 0_+} \int_0^1 \int_0^1 \left[\int_{t-h}^{t+h} \int_{s-h}^{s+h} \int_{t-h}^{t+h} \int_{s-h}^{s+h} a(u,v,x,y)\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}x\,\mathrm{d}y \right] \mathrm{d}t\,\mathrm{d}s. \tag{1.12}$$

If the kernel is continuous, the previous formula simplifies notably:

$$\mathrm{Tr}(A) = \int_0^1 \int_0^1 a(t,s,t,s)\,\mathrm{d}t\,\mathrm{d}s. \tag{1.13}$$

The singular value decomposition and eigendecomposition naturally translate to kernels. When $A \in \mathcal{S}_2(\mathcal{L}^2([0,1]^2))$ has the singular value decomposition

$$A = \sum_{r=1}^{\infty} \sigma_r (e_r \otimes f_r),$$

the corresponding kernel $a \in \mathcal{L}^2([0,1]^4)$ can be decomposed as

$$a(t,s,t',s') = \sum_{r=1}^{\infty} \sigma_r e_r(t,s) f_r(t',s'),$$

where the equality is understood in the $\mathcal{L}^2$-sense, i.e.

$$\lim_{R\to\infty} \int_{[0,1]^4} \left(a(t,s,t',s') - \sum_{r=1}^{R} \sigma_r e_r(t,s) f_r(t',s') \right)^2 \mathrm{d}t\,\mathrm{d}s\,\mathrm{d}t'\,\mathrm{d}s' = 0.$$

The eigendecomposition of the covariance operator C is particularly important due to its connection with the principal component decomposition of the corresponding random element X. Specifically, if

$$C = \sum_{r=1}^{\infty} \lambda_r (g_r \otimes g_r)$$

is the eigendecomposition of C, then X can be expanded as

$$X = m + \sum_{r=1}^{\infty} \xi_r g_r, \tag{1.14}$$

where $\xi_r = \langle X - m, g_r\rangle$, $r = 1, \ldots, R$, are uncorrelated random variables with mean zero and variance λ_r, and the equality holds in the mean-square sense, i.e.

$$\lim_{R\to\infty} \mathbb{E} \left\| X - m - \sum_{r=1}^{R} \xi_r g_r \right\|^2 = 0.$$

Equation (1.14) is the weak version ($\mathcal{L}^2$ version) of the celebrated the Karhunen-Loeve expansion (Grenander, 1981). If continuity of the covariance kernel is assumed, a strong (uniform) version of Karhunen-Loeve expansion holds, see Hsing and Eubank (2015) for details.

Finally, separability can also be characterized in terms of kernels: $A \in \mathcal{S}_2(\mathcal{L}^2([0,1]^2))$ is separable if and only if

$$a(t,s,t',s') = a_1(t,t')a_2(s,s')$$

for some kernels $a_1, a_2 \in \mathcal{L}^2([0,1]^2)$ and for almost all pairs $(t,s),(t',s') \in [0,1]^2$ (almost all in the Lebesgue sense).

1.5 Inference on Function Spaces

The strong law of large numbers and the central limit theorem (CLT) for random elements with values in a separable Hilbert space $\mathcal{H}_1$ resemble their real-valued counterparts (Hsing and Eubank, 2015). For $X_1, X_2, \ldots \in \mathcal{H}_1$ independent and identically distributed with $m = \mathbb{E}\|X_1\| < \infty$, we have

$$\frac{1}{N}\sum_{n=1}^{N} X_n \to m$$

almost surely, as $N \to \infty$. When in addition $\mathbb{E}\|X_1\|^2 < \infty$, then

$$\sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N} X_n - m\right)$$

converges (weakly) to a mean zero Gaussian random element with the covariance $C = \mathbb{E}(X_1 \otimes_2 X_1)$.

For a random element X in $\mathcal{H}_1$, the most natural estimators of its mean m and covariance C (based on a random sample $X_1, \ldots, X_N$) are the empirical mean

$$\bar{X}_N := \frac{1}{N}\sum_{n=1}^{N} X_n$$

and the empirical covariance

$$\widehat{C}_N := \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X}_N) \otimes_2 (X_n - \bar{X}_N).$$

From the above, we immediately see that the empirical mean is a consistent estimator of the mean (provided $\mathbb{E}\|X\| < \infty$), and asymptotically Gaussian random element of $\mathcal{H}_1$ (provided $\mathbb{E}\|X\|^2 < \infty$). The same can be said about the empirical covariance, which is a consistent estimator of the covariance (provided $\mathbb{E}\|X\|^2 < \infty$), and asymptotically Gaussian random element of $\mathcal{S}_2(\mathcal{H}_1)$ (provided $\mathbb{E}\|X\|^4 < \infty$).

Recall that the covariance is trace-class, and the space of trace-class operators is not a Hilbert space but a Banach space. While the strong law o large numbers can be extended to a (separable) Banach space as it stands (Bosq, 2012), the same is not true for the central limit theorem. We will only need the following CLT specifically for the empirical covariance, which is also a trace-class operator.

Theorem 1 (Mas2006)**.** *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random elements on a separable Hilbert space $\mathcal{H}_1$. Let*

$$\sum_{j=1}^{\infty}\left(\mathbb{E}\langle X, e_j\rangle^4\right)^{1/4} < \infty, \tag{1.15}$$

where $\{e_j\}_{j\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}_1$. Then $\sqrt{N}(\widehat{C}_N - C)$ converges weakly to a mean zero Gaussian random element of $\mathcal{S}_1(\mathcal{H}_1)$, where C is the covariance of X_1.

Note that condition (1.15) implies that $\mathbb{E}\|X\|^4 < \infty$. While the weaker condition implies convergence in the weaker Hilbert-Schmidt topology, the stronger condition (1.15) ensures convergence in the stronger trace-norm topology.

Finally, the continuous mapping theorem (CMT) holds both for the Hilbert-Schmidt and trace-norm topologies, as it holds even more generally for a continuous mapping between two metric spaces (Billingsley, 1999). In particular, if $Z_1, Z_2, \dots$ converges weakly to Z in $\mathcal{S}_p(\mathcal{H}_1 \otimes \mathcal{H}_2)$ and $F : \mathcal{S}_p(\mathcal{H}_1 \otimes \mathcal{H}_2) \to \mathcal{S}_p(\mathcal{H}_1)$ is a continuous mapping, then $F(Z_1), F(Z_2), \dots$ converges weakly to $F(Z)$ in $\mathcal{S}_p(\mathcal{H}_1)$.

The mapping F in the previous paragraph can be thought of as a marginalization operator, because it suppresses one of the two dimensions. In Section 1.7, we will introduce some specific marginalization operators for trace-class and Hilbert-Schmidt operator spaces.

1.6 Functional Data and Discrete Observations

Functional data analysis (FDA) covers a full gamut of statistical methods applicable to situations when the available data $X_1, \ldots, X_N$ can be thought of as independent realizations of a (continuous) random function or a random surface X, instead of a random vector or a random matrix $\mathbf{X}$. Despite the sample space (typically the Hilbert space $\mathcal{L}^2([0,1])$ or $\mathcal{L}^2([0,1]^2)$) having a continuous domain, the realizations are, quite naturally, never observed in the continuum. Under the commonly adopted model to deal with functional data (see e.g. Cai and Yuan, 2010; Li and Hsing, 2010; Zhang and Wang, 2016, and many references therein), the n-th surface is observed at M_n points $\{(t_{n1}, s_{n1}), \ldots, (t_{nM_n}, s_{nM_n})\} \subset [0,1]^2$ only approximately:

$$Y_{nm} = X_n(t_{nm}, s_{nm}) + \epsilon_{nm}, \qquad m = 1, \ldots, M_n \tag{1.16}$$

where ϵ_{nm} are mutually independent noise variables (also independent of the latent surfaces $X_1, \ldots, X_N$) with mean zero and variance σ^2.

Hence the observations themselves are discrete, and they have to be stored as such in computer memory. Still, the distinctive feature of FDA as opposed to multivariate methods is the assumption of existence of a latent continuous object, with the interest focused on that object rather than the measurements. The object does not have to be smooth per se, but smoothness is often leveraged in the analysis (Ramsay and Silverman, 2005). Another overarching theme of FDA is non-parametric inference, which is achievable due to the availability of replicated observations, and preferable due to the sought-after flexibility to model complicated data structures in infinite-dimensional spaces (Ferraty and Vieu, 2006).

The FDA literature can be categorized according to several theoretical or methodological features. By the typical number of measurements per single realization, functional data are traditionally classified as sparse (Yao et al., 2005a) or dense (Hall and Hosseini-Nasab, 2006). Although there is no formal definition of these two regimes, the convention is to consider functional data as densely sampled when the number of observations per curve converges to infinity as some power of the sample size (Zhang and Wang, 2016). Contrarily, functional data are considered sparsely sampled when the number of observations per curve is bounded. A rich source of sparsely sampled functional data are longitudinal studies. Zhang and Wang (2016) provide a comprehensive overview of the sampling regimes, categorizing them according to the achievable asymptotic properties. The sampling regimes are indeed only asymptotic concepts, and it is hard to say in practice which asymptotic regime to adhere to with a specific data set at hand. The exception arises, when data are observed on a common grid, in which case the regime has to be treated as dense from the theoretical perspective, because consistent estimation of the underlying continuous phenomenon is possible only when the grid size increases with increasing sample size.

In the dense regime, data are in fact very often observed on a grid, i.e. compared to (1.16), there is a temporal grid $\{t_1, \ldots, t_{K_1}\} \subset [0,1]$ and a spatial grid $\{s_1, \ldots, s_{K_2}\} \subset [0,1]$, and the resulting measurements of surfaces $X_1, \ldots, X_N$ take form of matrices $\mathbf{Y}_1, \ldots, \mathbf{Y}_N \in \mathbb{R}^{K_1 \times K_2}$ with entries given by

$$\mathbf{Y}[i,j] = X(t_i, s_j) + \epsilon_{ij}, \qquad i = 1, \ldots, K_1,\ j = 1, \ldots, K_2,$$

where ϵ_{ij} are again zero-mean, variance σ^2 noise variables independent of everything. Recall that, we always think of the first dimension as time, denoted by the variable t, and the second dimension as space, denoted by the variable s, but this distinction is made purely for the purposes of presentation. Mathematically, there is nothing special about time or space as variables.

While data observed on a grid can be easily addressed as multivariate, such approach suffers from disregarding smoothness as well as high dependencies between the grid points. The difference between multivariate and functional analyses is exemplified in how they approach the covariance. While invertibility issues in multivariate (and especially high-dimensional) statistics are typically caused by a low number of samples (Ravikumar et al., 2011), and the inverse of the covariance is assumed to exist on the population level, functional covariances are compact (even trace-class) and hence non-invertible.

Functional data observed on a grid (not necessarily two-dimensional) commonly arise in many fields such as

- genetics, genomics and analytical chemistry (Sørensen et al., 2013),
- growth studies (Ramsay and Silverman, 2005),
- plant science (Tessmer et al., 2013),
- biomechanics (Miller et al., 2008; Crane et al., 2011),
- chemometrics (Delaigle et al., 2012),
- electricity consumption studies (Ferraty and Vieu, 2006),
- weather and climate studies (Gneiting et al., 2006),
- linguistics (Pigoli et al., 2018),
- finance (Chen et al., 2020a),
- demography (Chen and Müller, 2012; Chen et al., 2017a),
- monitoring and tracking (Chen et al., 2017b), or
- traffic flow analysis (Chiou et al., 2014).

Another rich source of functional data observed on a grid (possibly after pre-processing steps) is biomedical imaging. Examples include

- electrocardiography (Cuevas et al., 2004),
- electroencephalograpy (Hasenstab et al., 2017),
- diffusion tensor imaging (Pomann et al., 2016),
- magnetic resonance imaging (Stoehr et al., 2020),
- positron emission tomography (Jiang et al., 2009), or
- magnetoencephalography (Bijma et al., 2005; Lynch and Chen, 2018).

Wang et al. (2016) use some of these large and rich data sets as examples of the "next-generation" functional data and suggest that novel methodologies emphasizing their computational aspects have to be developed to perform statistical analyses on these data sets, which is exactly the position taken by this thesis.

Another categorization of the FDA literature is based on the stage at which smoothing of the discretely observed data is deployed. The estimation paradigm can be classified either as the *smooth-then-estimate* approach (curves are smoothed individually before estimating the common mean and covariance) or the *estimate-then-smooth* approach (the common mean and covariance are estimated from the discrete measurements directly, and they may be later used for interpolation of the observations, see Descary, 2017, and references therein). The smooth-then-estimate approach, popularized by Ramsay and Silverman (2005, 2007) is suitable for densely observed data, particularly when observations are not gridded. The estimate-then-smooth approach is, on the other hand, necessary when working with sparsely observed data. When working on a grid, smoothing is sometimes deployed prior to estimation to overcome noise (Chen and Müller, 2012). On the other hand, gridded data can often be considered noiseless (especially when produced by complex pre-processing steps, e.g. Pigoli et al., 2018), or the noise can be kept and directly modelled instead of being suppressed (Descary and Panaretos, 2019). In these cases, we view the estimate-then-smooth approach as the more natural one, and we focus on it.

Apart from *when*, the question arises about *how* to smooth. While a whole range of smoothing techniques based on splines was utilized in FDA (see Zhang and Wang, 2016, and references therein), it is safe to say that local polynomial regression smoothers (in particular local linear smoothers) are conceptually simpler and better understood (Li and Hsing, 2010; Wang et al., 2016; Rubín and Panaretos, 2020). Also, when data are observed densely (e.g. on a grid) and the estimate-then-smooth approach is taken, the smoothing step may reduce to a simple form of interpolation.

Now, let us consider how to store functional data in computer memory. One approach is to express the observed data with respect to a basis, and then store only vectors or matrices of the corresponding basis coefficients (Ramsay and Silverman, 2005, Section 6.6). This is the most prevalent approach in practice (cf. the extensive literature review by Sørensen et al., 2013), and it corresponds to the smooth-then-estimate paradigm. When adapting the estimate-then-smooth paradigm, gridded observations can be naturally stored as vectors, or matrices, while non-gridded data, such as the sparse data stemming from (1.16), can be stored as lists of triplets, of the form (t, s, Y). However, these lists are not suitable for computing purposes. For example, the most efficient implementations of local linear smoothers utilize the fast Fourier transform (FFT) algorithm (Silverman, 1982), which in turn requires an equispaced grid. Thus even when data are not observed on a grid, they are often gridded for computational reasons.

For random surfaces, sampling schemes may differ in the two dimensions. Data sampled either densely on a two-dimensional grid (Pigoli et al., 2018; Lynch and Chen, 2018; Chen et al., 2017a; Bijma et al., 2005) or sparsely in both dimensions (e.g. implied volatilities in finance, see Chapter 4) are commonly encountered. However, there are also longitudinal studies in which a functional measurement is taken at each visit, leading to data that are sparse in the temporal dimension and dense in the spatial dimension (Park and Staicu, 2015). But again, these are commonly gridded (in both dimensions) for computational purposes (Greven et al., 2011; Kidziński and Hastie, 2018).

This thesis focuses on functional observations on a two-dimensional domain and adopts the estimate-then-smooth paradigm. In the dense regime, we develop our methodology mathematically on fully observed random elements of a Hilbert space $\mathcal{H}$, which covers both the continuous case (i.e. $\mathcal{H} = \mathcal{L}^2([0,1]^2)$) and the discrete case of matrices (i.e. $H = \mathbb{R}^{K_1 \times K_2}$). While we focus on the continuous case, sometimes assuming smoothness for simplicity, the assumption of smoothness can be dropped in the methodology, and the measure can be changed from the Lebesgue measure on $[0,1]^2$ to the counting measure on $\{1,\dots,K_1\} \times \{1,\dots,K_2\}$, replacing integration by summation in the formulas. In theory, we first treat both the continuous and discrete samples as fully observed, before relating measurements on a grid and the resulting discrete estimators with their continuous latent counterparts under some smoothness assumptions using a simple form of interpolation. Whenever discussing computational issues, we assume the data are observed on a grid, coming in as matrices, i.e. $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathbb{R}^{K_1 \times K_2}$. On the other hand, smoothing is imperative in Chapter 4, where we work with the sparse sampling regime. The methodology is developed with sparse measurements, and local linear smoothing is performed in continuum, in theory. In practice (and implementation), however, we assume again that the data are observed on a grid, coming in as matrices of finite size with many entries not available.

1.7 Marginalization Operators

The goal of this thesis is to handle the covariance estimation task for random surfaces in a computationally efficient way. Two-dimensional data generally lead to four-dimensional covariances. However, four-dimensional objects (such as, for example, the empirical covariance of surface-valued random elements) are beyond the computational limits we set for ourselves. While we can work with such objects implicitly in theory, we have to avoid even constructing them in calculations. For this purpose, we need some *marginalization* tools, which would allow us to explicitly work with smaller, *marginalized* objects.

In this section, we introduce two such marginalization tools: the *partial trace* and *the partial inner product*. Both of these are well-known in the field of quantum information theory (Schumacher and Westmoreland, 2010; Wilde, 2013), where univariate distributions are characterized by positive semi-definite matrices, multi-variate distributions are characterized by positive semi-definite tensors, and the partial trace or the partial inner product are commonly used to find the marginal distributions of these *random vectors*. However, distributions in quantum information theory are discrete, and generalizing the partial trace and the partial inner product to the case of continuous domains requires some work. In the field of functional data analysis, this has been done first by Aston et al. (2017) and Bagchi and Dette (2020), respectively.

1.7.1 Partial Tracing

In accordance with Aston et al. (2017), we define the partial trace as follows.

Definition 2. *The partial tracing operators w.r.t. the first and the second argument are the unique operators* $Tr_1 : \mathcal{S}_1(\mathcal{H}_1 \otimes \mathcal{H}_2) \to \mathcal{S}_1(\mathcal{H}_1)$ *and* $Tr_2 : \mathcal{S}_1(\mathcal{H}_1 \otimes \mathcal{H}_2) \to \mathcal{S}_1(\mathcal{H}_2)$ *satisfying for all* $A \in \mathcal{S}_1(\mathcal{H}_1)$ *and* $B \in \mathcal{S}_1(\mathcal{H}_2)$

$$\begin{aligned} Tr_1(A \tilde{\otimes} B) &= Tr(B)A\,, \\ Tr_2(A \tilde{\otimes} B) &= Tr(A)B\,. \end{aligned}$$

Basic properties of partial tracing are summarized in the following proposition.

Proposition 1. *(a)* Tr_1 *and* Tr_2 *are well-defined bounded linear operators.*

(b) Let $F \in \mathcal{S}_1^+(\mathcal{H}_1 \otimes \mathcal{H}_2)$*, then* $Tr_1 F \in \mathcal{S}_1^+(\mathcal{H}_1)$ *and* $Tr_1 F \in \mathcal{S}_1^+(\mathcal{H}_2)$*.*

(c) Let $\mathcal{H} = \mathcal{H}_1 \tilde{\otimes} \mathcal{H}_2$ *be a product Hilbert space and* $F \in \mathcal{S}_1(\mathcal{H}_1 \otimes \mathcal{H}_2)$*. If* F *is separable, then*

$$Tr(F)\, F = Tr_1(F) \tilde{\otimes} Tr_2(F)\,. \tag{1.17}$$

Moreover, if F *is positive semi-definite, then also the reverse implication holds.*

Proof. We refer to Aston et al. (2017) or our more general development in Chapter 2 for the proof of part (a).

For part (b), note that $\mathcal{S}_1^+(\mathcal{H})$ is a closed subset of $\mathcal{S}_1(\mathcal{H})$ for a Hilbert space $\mathcal{H}$. Thus the previous definition could be restricted only to positive semi-definite operators, and we can repeat the proof of part (a).

The first implication in part (c) follows easily from the definition and the additional part of Lemma 1. In the other direction, positive semi-definiteness is assumed only to avoid degenerate cases of vanishing traces. If $\mathrm{Tr}(F) \neq 0$, we can set

$$A := \frac{\mathrm{Tr}_2(F)}{\sqrt{\mathrm{Tr}(F)}} \quad \text{and} \quad B := \frac{\mathrm{Tr}_1(F)}{\sqrt{\mathrm{Tr}(F)}} \,,$$

which leads to $F = A \tilde{\otimes} B$, i.e. F being separable. If $\mathrm{Tr}(F) = 0$, then due to positive semi-definiteness it must be $F \equiv 0$, and the statement is trivial. □

Equation (1.17) is paramount to estimation. We will derive similar relationships also for operators other than partial traces, and we call equations such as (1.17) *estimating equations.*

Here we provide an intuition behind partial tracing. Let $F \in \mathcal{S}_1(\mathcal{H}_1 \otimes \mathcal{H}_2)$. As a compact operator, F can be expressed with respect to a basis $\{e_i \otimes f_j\}$ as

$$F = \sum_{i,j,k,l} \sigma_{ijkl} (e_i \otimes f_j) \otimes_2 (e_k \otimes f_l) \tag{1.18}$$

where $\{e_i\}$ and $\{f_j\}$ are ONBs in $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. By linearity of the partial trace we have

$$\mathrm{Tr}_1(F) = \sum_{i,j,k,l} \sigma_{ijkl} \mathrm{Tr}_1\Big((e_i \otimes f_j) \otimes_2 (e_k \otimes f_l) \Big)$$

Now, due to (1.8), it holds that

$$\mathrm{Tr}_1(F) = \sum_{i,j,k,l} \sigma_{ijkl} \mathrm{Tr}(f_j \otimes_2 f_l) e_i \otimes_2 e_k = \sum_{i,k} \Big(\sum_j \sigma_{ijil} \Big) e_i \otimes_2 e_k$$

where the last inequality follows from the fact that

$$\mathrm{Tr}(f_j \otimes_2 f_l) = \sum_{k=1}^{\infty} \langle (f_j \otimes_2 f_l) f_k, f_k \rangle = \sum_{k=1}^{\infty} \langle f_j, f_k \rangle \langle f_l, f_k \rangle = \mathbb{1}_{[j=l]}.$$

The previous expression justifies the name for partial tracing.

Another illustration of the name comes from the integral representation for continuous kernels. Compare the following proposition to the integral trace formula (1.13).

Proposition 2. *Let* $F \in \mathcal{S}_1^+(\mathcal{L}^2([0,1]^2))$ *have a continuous kernel* $k = k(t,s,t',s')$*. Then* $Tr_1(F)$ *resp.* $Tr_2(F)$ *have continuous kernels*

$$k_1(t,t') = \int_0^1 k(t,s,t',s)\,\mathrm{d}s \qquad \textit{resp.} \qquad k_2(s,s') = \int_0^1 k(t,s,t,s')\,\mathrm{d}t\,.$$

1.7.2 Partial Inner Product

The partial inner product is an alternative to partial tracing. Unlike partial tracing, which is defined for trace-class operators only, the partial inner product naturally operates on the space of Hilbert-Schmidt operators. In this section, we revisit the development of Bagchi and Dette (2020), while a more general development will be one of the objectives in Chapter 3.

Definition 3. *Let* $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$*. The partial inner product w.r.t. the first argument is the unique operator* $T_1 : \mathcal{S}_2(\mathcal{H}) \times \mathcal{S}_2(\mathcal{H}_1) \to \mathcal{S}_2(\mathcal{H}_1)$ *given by*

$$T_1(A \tilde{\otimes} B, W_2) = \langle B, W_2\rangle_{\mathcal{S}_2(\mathcal{H}_2)} A\,, \quad A \in \mathcal{S}_2(\mathcal{H}_1),\ B, W_2 \in \mathcal{S}_2(\mathcal{H}_2)\,.$$

Similarly the partial inner product w.r.t. the second argument is the unique operator $T_2 : \mathcal{S}_2(H) \times \mathcal{S}_2(\mathcal{H}_2) \to \mathcal{S}_2(\mathcal{H}_2)$ *given by*

$$T_2(A \tilde{\otimes} B, W_1) = \langle A, W_1\rangle_{\mathcal{S}_2(\mathcal{H}_2)} B\,, \quad A, W_1 \in \mathcal{S}_2(\mathcal{H}_1),\ B \in \mathcal{S}_2(\mathcal{H}_2)\,.$$

The following proposition is analogous to Proposition 1.

Proposition 3. *(a)* T_1 *and* T_2 *are well-defined bounded bi-linear operators.*

(b) Let $F \in \mathcal{S}_2^+(\mathcal{H}_1 \otimes \mathcal{H}_2)$*,* $W_1 \in \mathcal{S}_2^+(\mathcal{H}_1)$ *and* $W_2 \in \mathcal{S}_2^+(\mathcal{H}_2)$*, then* $T_1(F, W_2) \in \mathcal{S}_2^+(\mathcal{H}_2)$ *and* $T_2(F, W_1) \in \mathcal{S}_2^+(\mathcal{H}_2)$*.*

(c) Let $F \in \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2) =: \widetilde{H}$*,* $W_1 \in \mathcal{S}_2(\mathcal{H}_1)$ *and* $W_2 \in \mathcal{S}_2(\mathcal{H}_2)$*. If* F *is separable, we have*

$$\langle F, W_1 \tilde{\otimes} W_2\rangle_{\widetilde{H}} F = T_1(F, W_2) \tilde{\otimes} T_2(F, W_1)\,. \tag{1.19}$$

Moreover, if F *is positive semi-definite and* W_1 *and* W_2 *are positive definite, the reverse implication holds as well.*

Proof. For the proof of part (a), including the uniqueness claim, we refer to Bagchi and Dette (2020) or our development in Chapter 3.

Part (b) follows again from the closedness of positive semi-definite operators.

For the final part, let $F = A \tilde{\otimes} B$ for some $A \in \mathcal{S}_2(\mathcal{H}_1)$ and $B \in \mathcal{S}_2(\mathcal{H}_2)$. Note that by the construction of the product space $\mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$ we have

$$\langle A \tilde{\otimes} B, W_1 \tilde{\otimes} W_2 \rangle_{\widetilde{H}} = \langle A, W_1 \rangle_{\mathcal{S}_2(\mathcal{H}_1)} \langle B, W_2 \rangle_{\mathcal{S}_2(\mathcal{H}_2)} .$$

The first part of (c) thus follows from the definition of partial inner product and linearity of outer product.

For the other direction, note that $W_1 \tilde{\otimes} W_2$ must by positive definite by Corollary 1, and hence $\langle F, W_1 \tilde{\otimes} W_2 \rangle > 0$ as long as F is non-zero. Thus we can define

$$A := \frac{T_2(F, W_1)}{\sqrt{\langle F, W_1 \tilde{\otimes} W_2 \rangle}} \quad \text{and} \quad B := \frac{T_2(F, W_1)}{\sqrt{\langle F, W_1 \tilde{\otimes} W_2 \rangle}}$$

to have $F = A \tilde{\otimes} B$.

In the case of $F \equiv 0$, the statement is trivial. □

Similarly to the previous section, the name for partial inner product is exemplified in the case of integral operators.

Proposition 4. *Let $F \in \mathcal{S}_2^+(\mathcal{L}^2([0,1]^2))$ have a continuous kernel $k = k(t,s,t',s')$, $W_1 \in \mathcal{S}_2(\mathcal{L}^2([0,1]))$ have kernel $w_1 = w_1(t,t')$ and $W_2 \in \mathcal{S}_2(\mathcal{L}^2([0,1]))$ have kernel $w_2 = w_2(s,s')$. Then $T_1(F, W_2)$ has kernel*

$$k_1(t,t') = \int_0^1 \int_0^1 k(t,s,t',s') w_2(s,s') \, \mathrm{d}s \, \mathrm{d}s'$$

and similarly $T_2(F, W_1)$ has kernel

$$k_2(s,s') = \int_0^1 \int_0^1 k(t,s,t',s') w_1(t,t') \, \mathrm{d}t \, \mathrm{d}t' .$$

In Chapter 3, we will see that the assumptions of continuity and positive semi-definiteness are not needed when working with a slightly more general definition of the partial inner product, since the equalities are understood in the $\mathcal{L}^2$-sense anyway.

1.7.3 Examples with Matrix-variate Normal Distribution

The development of the partial traces and the partial inner products up to now is valid for matrix-variate data as well, since the only place where we assumed smoothness up to this point were the integral representations of the marginalization operators given in Propositions 2 and 4. But these integral representations can be easily rewritten for discrete objects by the change of measure, replacing integrals with sums. It will become apparent later that such discrete approximations are sound.

In this section, we consider a multivariate separable model known as the matrix-variate normal distribution (Gupta and Nagar, 2018), which we use to demonstrate the actions of the marginalization operators. In the following, $\mathrm{vec}(\cdot) : \mathbb{R}^{K_1 \times K_2} \to \mathbb{R}^{K_1 K_2}$ is the vectorization operator, transforming a matrix into a vector by stacking the columns of the matrix under each other.

Definition 4. *A random matrix* $\mathbf{X} \in \mathbb{R}^{K_1 \times K_2}$ *is said to have the matrix variate normal distribution with mean matrix* $\mathbf{M} \in \mathbb{R}^{K_1 \times K_2}$ *and covariance* $\mathbf{\Sigma} \tilde{\otimes} \mathbf{\Psi}$*, where* $\mathbf{\Sigma} \in \mathbb{R}^{K_1 \times K_1}$ *and* $\mathbf{\Psi} \in \mathbb{R}^{K_2 \times K_2}$ *are positive semi-definite, if*

$$\mathrm{vec}(\mathbf{X}) \sim \mathcal{N}_{K_1 K_2}\Big(\mathrm{vec}(\mathbf{M}), \mathbf{\Sigma} \tilde{\otimes} \mathbf{\Psi}\Big) .$$

We shall use the notation $\mathbf{X} \sim \mathcal{N}_{K_1,K_2}(\mathbf{M}, \mathbf{\Sigma} \tilde{\otimes} \mathbf{\Psi})$

In the previous definition, $\mathbf{\Sigma}$ can be thought of as the row-specific covariance matrix, while $\mathbf{\Psi}$ can be thought of as the column-specific covariance matrix. The following theorem is imperative for numerical simulations (Gupta and Nagar, 2018, Theorem 2.3.10).

Theorem 2. *Let* $\mathbf{X} \sim \mathcal{N}_{K_1,K_2}(\mathbf{M}, \mathbf{\Sigma} \tilde{\otimes} \mathbf{\Psi})$*,* $\mathbf{A} \in \mathbb{R}^{K_1 \times n_1}$ *be of rank* $n_1 \leq K_1$*, and* $\mathbf{B} \in \mathbb{R}^{K_2 \times n_2}$ *be of rank* $n_2 \leq K_2$*. Then* $\mathbf{A}^\top \mathbf{X} \mathbf{B} \sim \mathcal{N}_{n_1,n_2}\Big(\mathbf{A}^\top \mathbf{M} \mathbf{B}, \mathbf{A}^\top \mathbf{\Sigma} \mathbf{A} \tilde{\otimes} \mathbf{B}^\top \mathbf{\Psi} \mathbf{B}\Big)$*.*

From now on, let $K_1 = K_2 =: K$. Consider a single observation from the following distribution

$$\mathbf{X} \sim \mathcal{N}_{K,K}(\mathbf{0}, \mathbf{\Sigma} \tilde{\otimes} \mathbf{I}_K) . \tag{1.20}$$

This is the most important model in multivariate statistics: the columns of $\mathbf{X}$ are i.i.d. following $\mathcal{N}_K(\mathbf{0}, \mathbf{\Sigma})$. In this classical setup, the number of rows is not constrained to be the number of columns, but here we impose this assumption for the sake of presentation.

We know from the multivariate statistics that the empirical estimate – and also the maximum likelihood estimator (MLE) – of $\mathbf{\Sigma}$ in this model is

$$\widehat{\mathbf{\Sigma}} = \frac{1}{K} \mathbf{X} \mathbf{X}^\top .$$

The covariance $\mathbf{C}$ of $\mathbf{X}$ in this model is exactly $\mathbf{C} = \mathbf{\Sigma} \tilde{\otimes} \mathbf{I}_K$, and the empirical (matrix-variate) estimate of $\mathbf{C}$ (based on a single observation only, but we still use the common notation) is

$$\widehat{\mathbf{C}}_N = \mathbf{X} \otimes \mathbf{X} \in \mathbb{R}^{K \times K \times K \times K} .$$

The partial tracing estimate of $\mathbf{\Sigma}$ would thus be

$$\widetilde{\mathbf{\Sigma}} = \frac{1}{\alpha} \mathrm{Tr}_1(\widehat{\mathbf{C}}_N) ,$$

where α is a constant. Normally, $\alpha = \sqrt{\mathrm{Tr}(\widehat{\mathbf{C}}_N)}$. Since in the more general setup it is possible to estimate the two constituents of the separable covariance only up to a scaling factor, it is natural to assume that $\mathrm{Tr}(\boldsymbol{\Sigma}) = \mathrm{Tr}(\boldsymbol{\Psi})$, which leads to this choice $\alpha = \sqrt{\mathrm{Tr}(\widehat{\mathbf{C}}_N)}$. Since in this case we know that $\boldsymbol{\Psi} = \mathbf{I}_K$, estimation of the scale is possible. Thus we choose α to be a constant such that

$$\mathrm{Tr}(\widehat{\mathbf{C}}_N) \overset{!}{=} \mathrm{Tr}\Big(\widetilde{\boldsymbol{\Sigma}} \mathbin{\tilde{\otimes}} \mathbf{I}_K\Big) = \mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}})\, \mathrm{Tr}(\mathbf{I}_K) = d\, \mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}) = d \frac{1}{\alpha} \mathrm{Tr}(\widehat{\mathbf{C}}_N)\,.$$

Thus $\alpha = K$.

And since (using the Matlab notation) we have

$$\mathrm{Tr}_1(\widehat{\mathbf{C}}) = \sum_{k=1}^{K} \widehat{\mathbf{C}}_N[:,k,:,k] = \sum_{k=1}^{K} \mathbf{X}[:,k] \otimes \mathbf{X}[:,k] = \sum_{k=1}^{K} \mathbf{X}[:,k]\mathbf{X}[:,k]^\top = \mathbf{X}\mathbf{X}^\top,$$

we obtain $\widetilde{\boldsymbol{\Sigma}} \equiv \widehat{\boldsymbol{\Sigma}}$, i.e. the partially traced estimate corresponds to the natural multivariate estimate in this case.

The special structure of the covariance in model (1.20) can be visualized, if we matricize the covariance, which corresponds to the covariance matrix of the vectorized random element. Specifically, $\mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{d^2}$ is distributed as a Gaussian random vector with mean zero and the Kronecker product covariance $\mathbf{I}_K \otimes_K \boldsymbol{\Sigma}$, cf. Remark 1 (note the flip of the two covariance factors between the outer product and the Kronecker product). The block structure of $\mathbf{I}_K \otimes_K \boldsymbol{\Sigma}$ is shown in Figure 1.1 (left). The blocks around the diagonal are all the same, while the remaining entries are zero. Partial tracing takes this into account and estimates the unknown $\boldsymbol{\Sigma}$ by taking an average of the diagonal blocks of the empirical version of the covariance.

In the case of independent rows sampled from $\mathcal{N}_K(\mathbf{0}, \boldsymbol{\Psi})$, i.e.

$$\mathbf{X} \sim \mathcal{N}_{K,K}(\mathbf{0}, \mathbf{I}_K \otimes \boldsymbol{\Psi})\,, \tag{1.21}$$

the empirical (and also the maximum likelihood) multi-variate estimate of $\boldsymbol{\Psi}$ is

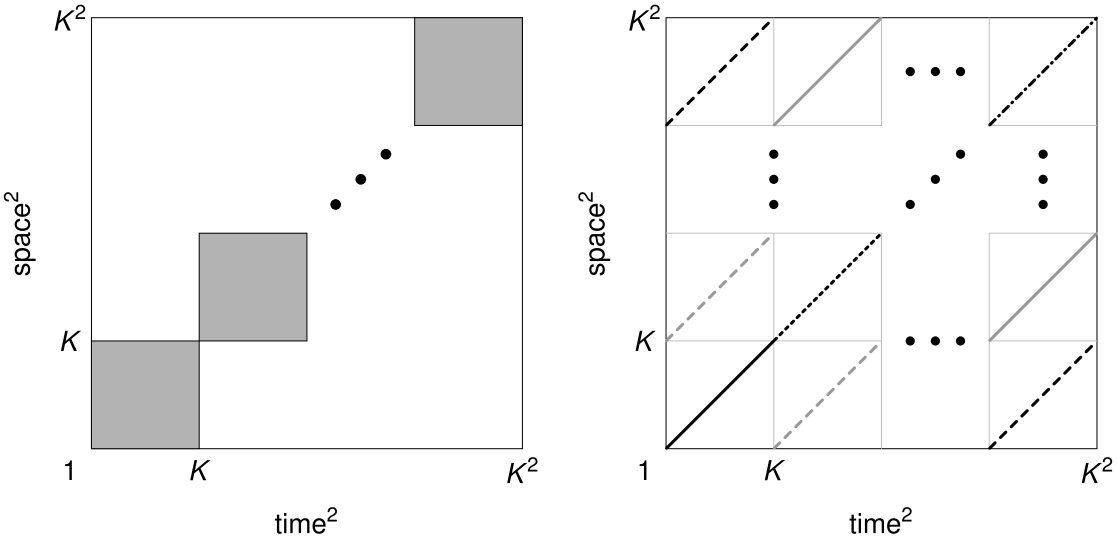
$$\widehat{\boldsymbol{\Psi}} = \frac{1}{K}\mathbf{X}^\top\mathbf{X}\,.$$

This again corresponds to the partially traced estimate of $\boldsymbol{\Psi}$ in a similar manner. Vectorization of distribution (1.21) is a multivariate Gaussian distribution with mean zero and covariance $\boldsymbol{\Psi} \otimes_K \mathbf{I}_K$, whose stripe structure is depicted in Figure 1.1 (right). In this case, all diagonal entries of a single block are the same, and partial tracing again exploits this, averaging the corresponding elements of the empirical covariance.

In the general case

$$\mathbf{X} \sim \mathcal{N}_{K,K}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$$

Figure 1.1: Left: Block structure of $\mathbf{I}_K \otimes_K \mathbf{\Psi}$. **Right**: Stripe structure of $\mathbf{\Sigma} \otimes_K \mathbf{I}_K$.



a natural question arises: do the estimates $\widehat{\mathbf{\Sigma}} = \frac{1}{K}\mathbf{X}\mathbf{X}^\top$ and $\widehat{\mathbf{\Psi}} = \frac{1}{K}\mathbf{X}^\top\mathbf{X}$ still work here? We know that the partially traced estimates are

$$\widetilde{\mathbf{\Sigma}} = \frac{1}{\sqrt{\operatorname{Tr}(\widehat{\mathbf{C}}_N)}}\operatorname{Tr}_1(\widehat{\mathbf{C}}_N) \quad \text{and} \quad \widetilde{\mathbf{\Psi}} = \frac{1}{\sqrt{\operatorname{Tr}(\widehat{\mathbf{C}}_N)}}\operatorname{Tr}_2(\widehat{\mathbf{C}}_N).$$

It will be shown later in Chapter 2 that they are consistent, and it still holds that $\operatorname{Tr}_1(\widehat{\mathbf{C}}_N) = \mathbf{X}\mathbf{X}^\top$ and $\operatorname{Tr}_2(\widehat{\mathbf{C}}_N) = \mathbf{X}^\top\mathbf{X}$. The scaling constants $1/K$ are not appropriate in general, but they work e.g. when $\mathbf{X}$ is standardized so both the rows and the columns have unit variances, which is a common practice in many applications. The partially traced estimates, however, do not correspond to the maximum likelihood estimates in the general case.

In a general matrix variate Gaussian model with full-rank covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Psi}$, the maximum likelihood estimates do not have a closed form solutions and they are found by the following alternating minimization algorithm (Dutilleul, 1999):

```
INPUT: Data $\mathbf{X}_1, \dots, \mathbf{X}_N$, an initial guess for $\mathbf{\Psi}$.

REPEAT

    $$\begin{aligned} \mathbf{\Sigma} &:= \frac{1}{K_1 N}\sum_{n=1}^{N}(\mathbf{X}_n - \bar{\mathbf{X}}_N)\mathbf{\Psi}^{-1}(\mathbf{X}_n - \bar{\mathbf{X}}_N)^\top \\ \mathbf{\Psi} &:= \frac{1}{K_2 N}\sum_{n=1}^{N}(\mathbf{X}_n - \bar{\mathbf{X}}_N)^\top\mathbf{\Sigma}^{-1}(\mathbf{X}_n - \bar{\mathbf{X}}_N) \end{aligned} \qquad (1.22)$$

UNTIL convergence
```

The relative magnitudes of $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ can change throughout the iterations, possibly even diverge, which can cause numerical instability. It is advisable to incorporate some sort of scaling, e.g. to ensure that $\mathrm{Tr}(\mathbf{\Sigma}) = \mathrm{Tr}(\mathbf{\Psi})$ throughout the iterations.

To finish this chapter, let us compare the partial trace and the partial inner product against the maximum likelihood algorithm above. Partial tracing estimates the covariance matrices as scaled averages of some of the entries of the empirical covariance matrix, as depicted in Figure 1.1. It will become clear later that equations (1.22) can be rewritten, up to the scaling constants, using the partial inner products as

$$\mathbf{\Sigma} \approx T_1(\widehat{\mathbf{C}}_N, \mathbf{\Psi}^{-1}) \quad \text{and} \quad \mathbf{\Psi} \approx T_2(\widehat{\mathbf{C}}_N, \mathbf{\Sigma}^{-1}).$$

Secondly, we will see that, without the inverses in the previous equation, the algorithm solves the following optimization problem:

$$\operatorname*{arg\,min}_{\mathbf{\Sigma},\mathbf{\Psi}} \left|\!\left|\!\left| \widehat{\mathbf{C}}_N - \mathbf{\Sigma} \otimes \mathbf{\Psi} \right|\!\right|\!\right|_F . \tag{1.23}$$

The solution to (1.23) has been studied by Van Loan and Pitsianis (1993), and was suggested as a separable estimator of the covariance by Genton (2007). And finally, it holds that

$$\mathrm{Tr}_1(\widehat{\mathbf{C}}_N) = T_1(\widehat{\mathbf{C}}_N, \mathbf{I}_K) \quad \text{and} \quad \mathrm{Tr}_2(\widehat{\mathbf{C}}_N) = T_2(\widehat{\mathbf{C}}_N, \mathbf{I}_K),$$

hence the partial inner product can be thought of as a generalization of partial tracing when working on finite-dimensional domains (with matrices or tensor). On a general Hilbert space, the identity need not be a Hilbert-Schmidt operator and hence the previous equation need not make sense. Altogether, we can say that partial tracing can be thought of as both the first step in either the maximum likelihood algorithm for the matrix-variate normal distribution, and the first step in the algorithm for computing the solution to 1.23, i.e. finding the nearest Kronecker product (NKP) approximation to the empirical covariance matrix[1]. However, the maximum likelihood algorithm cannot be applied on a general Hilbert space, due to the inverses used. And as for the algorithm to solve (1.23), which is based on the partial inner product, initialization by partial tracing is not very natural for the same reason.

Remark 2. *(Notation.) The subscripts for partial traces and partial inner products correspond to the dimension that is being kept. For example, $Tr_1(C) \in \mathcal{S}_1(\mathcal{H}_1)$ corresponds to a temporal covariance. The opposite convention was adopted by Aston et al. (2017) and Bagchi and Dette (2020), who use the subscript to designate which dimension is being integrated out (e.g. $Tr_1(C) \in \mathcal{S}_1(\mathcal{H}_2)$). Both of these different conventions in notation have their pros and cons. We use the subscript to denote the dimension that is being*

---

[1] In statistics literature, the Kronecker product, which is nearest to the empirical covariance matrix, is often called the *best separable approximation*, a term coined by Genton (2007). However, this latter term can be misleading, and we will only use it in this thesis when the true (unknown) covariance is being approximated, instead of the empirical covariance in (1.23)

kept, since we view the partial tracing and the partial inner product as marginalization operators*.*

2 Separable-plus-Banded Model

In this chapter, we postulate the following model for the covariance C of a random element $X \in \mathcal{L}^2([0,1]^2)$:

$$C = A_1 \tilde{\otimes} A_2 + B, \tag{2.1}$$

where $A_1, A_2 \in \mathcal{S}_1^+(\mathcal{L}^2[0,1])$ and $B \in \mathcal{S}_1^+(\mathcal{L}^2([0,1]^2))$ is *banded* by $\delta \in [0,1)$ in the sense of its kernel vanishing far away from the diagonal, as explained below. On the level of kernels, this implies the decomposition

$$c(t,s,t',s') = a_1(t,t')a_2(t,t') + b(t,s,t',s') \tag{2.2}$$

for almost all $t, s, t, s' \in [0,1]$, where $b(t,s,t',s') = 0$ almost everywhere on

$$\Big\{(t,s,t',s') \in [0,1]^4 \Big| \max(|t-t'|, |s-s'|) \geq \delta\Big\}.$$

In the case of measurements on the grid $\{(t_i, s_j), i,j = 1,\dots,K\}$, equations (2.1) and (2.2) become

$$\mathbf{C}[i,j,i',j'] = \mathbf{A}_1[i,i'] \tilde{\otimes} \mathbf{A}_2[j,j'] + \mathbf{B}[i,j,i',j'], \qquad i,i',j,j' = 1,\dots,K,$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are square matrices of sizes $K \times K$, and $\mathbf{B} \in \mathbb{R}^{K\times K\times K\times K}$ is a tensor satisfying $\mathbf{B}[i,j,i',j'] = 0$ if $\max(|i-i'|, |j-j'|) \geq d$, where $d = \lceil \delta K \rceil \in \{1,\dots,K\}$ is the discrete version of the bandwidth δ. Note that by choosing $d = 1$, the errors-in-measurement model (1.16) with a separable covariance is contained in the separable-plus-banded model as a sub-model. In fact, the separable-plus-banded model allows the underlying process X to be corrupted by errors that propagate in space and time. That is when we observe $Y = X + W$, where X and W are uncorrelated processes, and W has a banded covariance. Of course, W may not be just an error process, but possibly an object of its own value.

Combining the separable and the banded components into a single model results in a non-parametric family, which is much richer than the separable class. In particular, the model represents a strict generalisation 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) non-separable characteristics of the process. 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 non-separable 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 non-parametric 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* (Aston et al., 2017), but it is suitably modified to allow us to separate the terms a and b in (2.2). 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 it 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 (2.2) 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, i.e. 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) 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 corresponding to noise may become a banded structure corresponding to another signal. One can thus view our methodology as being able to estimate a separable model observed under heteroscedastic and/or weakly dependent noise. If the number of degrees of freedom belonging to the noise does not exceed the degrees of freedom of the separable part, we can utilize the noise structure e.g. 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 chapter is that the methodology we advocate, and label shifted partial tracing, can be used to estimate the separable part of model (2.2), provided data are densely observed.

Before developing the apparatus for estimation of model (2.1), let us show that the model is identifiable up to scaling in A_1 and A_2, unless A_1 and A_2 are themselves banded.

Lemma 2. *Let $A_1, A_2, \widetilde{A}_1, \widetilde{A}_2 \in \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$ be operators with continuous kernels $a_1(t,t')$, $a_2(s,s')$, $\widetilde{a}_1(t,t')$ and $\widetilde{a}_2(s,s')$, respectively. Let $B, \widehat{B} \in \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$ be operators with continuous kernel $b(t,s,t',s')$ and $\widetilde{b}(t,s,t',s')$ banded by $\delta \in [0,1)$ and $\widetilde{\delta} \in [0,1)$, respectively. Let there be $t_1, t_2 \in [0,1]$ and $s_1, s_2 \in [0,1]$ such that $|t_1 - t_2| > \max(\delta, \widetilde{\delta})$ and $|s_1 - s_2| > \max(\delta, \widetilde{\delta})$, and $a_1(t_1,t_2) \neq 0$ and $a_2(s_1,s_2) \neq 0$. Then*

$$a_1(t,t')a_2(s,s') + b(t,s,t',s') = \widetilde{a}_1(t,t')\widetilde{a}_2(s,s') + \widetilde{b}(t,s,t',s')\,, \quad \forall t,s,t',s' \in [0,1]$$

if and only if

$$a_1(t,t')a_2(s,s') = \widetilde{a}_1(t,t')\widetilde{a}_2(s,s') \quad \& \quad b(t,s,t',s') = \widetilde{b}(t,s,t',s')\,, \quad \forall t,s,t',s' \in [0,1]$$

Proof. Only the left-to-right implication is interesting. To this end, the assumptions give us

$$\begin{aligned} a_1(t,t')a_2(s_1,s_2) &= \widetilde{a}_1(t,t')\widetilde{a}_2(s_1,s_2)\,, \quad \forall t,t' \in [0,1]\,, \\ a_1(t_1,t_2)a_2(s,s') &= \widetilde{a}_1(t_1,t_2)\widetilde{a}_2(s,s')\,, \quad \forall s,s' \in [0,1]\,, \\ a_1(t_1,t_2)a_2(s_1,s_2) &= \widetilde{a}_1(t_1,t_2)\widetilde{a}_2(s_1,s_2)\,. \end{aligned}$$

Multiplying the first two equations and dividing the result by the third one gives the equality between the separable parts. The equality between the banded parts follows naturally. □

The previous proof suggests an estimation strategy, which would be, however, quite wasteful. Hence we take a slightly different route in the following section.

2.1 Shifted Partial Tracing

Here we develop methodology capable of estimating model (2.1). For the sake of presentation, we first work under additional assumptions (positive semi-definiteness and continuity of the kernel) in this section, before providing more general (but less intuitive) development in the following sub-section. We start by defining the *shifted trace.*

Definition 5. *Let $F \in \mathcal{S}_1^+(\mathcal{L}^2([0,1]))$ have a continuous kernel $k = k(t,s)$. Let $\delta \in [0,1)$. We define the δ-shifted trace of F as*

$$Tr^{\delta}(F) := \int_0^{1-\delta} k(t, t+\delta)\, \mathrm{d}t\,.$$

In the special case of $\delta = 0$ the definition corresponds to the standard (non-shifted) trace of a trace-class operator with a continuous kernel, cf. (1.13). The definition of the shifted trace is naturally extended to higher dimensions. Next, we define the shifted partial trace.

Definition 6. *Let $\delta \in [0,1)$. Let $F \in \mathcal{S}_1^+(\mathcal{L}^2([0,1]^2))$ have a continuous kernel $k = k(t,s,t',s')$. We define the $\delta-$shifted partial traces of F, denoted $Tr_1^{\delta}(F)$ and $Tr_2^{\delta}(F)$, as the integral operators with kernels given respectively by*

$$k_1(t,t') := \int_0^{1-\delta} k(t,s,t',s+\delta) ds \qquad \& \qquad k_2(s,s') := \int_0^{1-\delta} k(t,s,t+\delta,s')\, \mathrm{d}t\,. \tag{2.3}$$

Again, for $\delta = 0$, δ-shifted partial tracing corresponds to partial tracing as defined by Aston et al. (2017). Also, notice that shifted partial tracing is linear. While partial tracing of Aston et al. (2017) can be used to estimate a separable model, we will later see that introducing a shift like in the previous definitions will allow us to work around short-range dependencies in the data, estimating the separable part of the separable-plus-banded model.

At this point, it is not immediately clear that the integral defining shifted (partial) traces are finite. While this could be shown directly here using trace-classness, positive semi-definiteness and continuity, we opt to skip these proofs for the sake of brevity. Correctness of Definitions 5 and 6 will ultimately follow from the more general development in Section 2.1.1.

Remark 3. *The definition of shifted partial tracing above is not symmetric, meaning that the result of the shifted partial trace is not necessarily self-adjoint. We could define a symmetrized shifted partial trace instead, but this is (due to linearity of shifted partial* *tracing and symmetry of the kernel* k*) equivalent to symmetrizing the result. The latter is used in practice for its computational convenience, while the former is hypothetically done in theory, but we avoid it without loss of generality to ease the presentation (see Section 2.3.4).*

The shifted partial trace has the following properties, which resemble those of the standard (non-shifted) partial trace.

Proposition 5. *Let* $A_1, A_2 \in \mathcal{S}_1^+(\mathcal{L}^2([0,1]))$ *have continuous kernels and* $F = A_1 \tilde{\otimes} A_2$*. Then*

- *(a)* $\mathrm{Tr}_1^\delta(F) = \mathrm{Tr}^\delta(A_2)\, A_1$,
- *(b)* $\mathrm{Tr}^\delta(F) = \mathrm{Tr}^\delta(A_1)\, \mathrm{Tr}^\delta(A_2)$*, and*
- *(c)* $\mathrm{Tr}^\delta(F)\, F = \mathrm{Tr}_1^\delta(F) \tilde{\otimes} \mathrm{Tr}_2^\delta(F)$.

Proof. First note that all the shifted (partial) traces are well defined. The claims follow from the definitions and separability of F. The kernel of $\mathrm{Tr}_1^\delta(F)$ is given by

$$k_1(t,t') = \int_0^{1-\delta} a_1(t,t') a_2(s,s+\delta) ds = a_1(t,t') \int_0^{1-\delta} a_2(s,s+\delta)\,\mathrm{d}s = a_1(t,t')\mathrm{Tr}^\delta(A_2)\ ,$$

which shows part (a). For the second part, using Fubini's theorem, we have

$$\begin{aligned} \mathrm{Tr}^\delta(F) &= \int_0^1 \int_0^1 a_1(t,t+\delta) a_2(s,s+\delta)\,\mathrm{d}t\,\mathrm{d}s \\ &= \int_0^1 a_1(t,t+\delta) dt \int_0^1 a_2(s,s+\delta)\,\mathrm{d}s = \mathrm{Tr}^\delta(A_1)\,\mathrm{Tr}^\delta(A_2)\ . \end{aligned}$$

Part (c) follows naturally by combining part (a) with part (b). ☐

The following lemma illustrates the importance of shifted partial tracing for estimation of model (2.1).

Lemma 3. *Let* $B \in \mathcal{S}_1^+(\mathcal{L}^2([0,1])^2)$ *be banded by* $\delta^\star$ *and have a continuous kernel* b*. Then for any* $\delta > \delta^\star$ *we have* $\mathrm{Tr}_1^\delta(B) = \mathrm{Tr}_2^\delta(B) = 0$.

Proof. The kernel of $\mathrm{Tr}_1^\delta(B)$ is $b_1(t,t') = \int_0^{1-\delta} b(t,s,t',s+\delta)\,\mathrm{d}s = 0$ due to bandedness of B, because $b(t,s,t,s') = 0$ for $|s-s'| > \delta^\star$. Similarly for $\mathrm{Tr}_2^\delta(B)$. ☐

Hence, shifted partial tracing allows us to work around the banded part of the model. Or, more precisely, banded operators belong to the kernel of shifted partial tracing, while separable operators do not. Due to linearity, shifted partial tracing enables us to deconvolve the two parts of model (2.1) as will be shown in Section 2.2.

In the following section, we will see that the statements of Proposition 5 and Lemma 3 hold even without the assumption of continuity and positive semi-definiteness.

2.1.1 General Definition

To ease the exposition, we assumed continuity and positive semi-definiteness in the definition of shifted (partial) tracing given above. But in order to prove the asymptotic results of Section 2.4, it is necessary to generalize the notions of shifted (partial) tracing to general trace-class operators on $\mathcal{L}^2([0,1]^2)$, i.e. to covariances of random elements on $\mathcal{L}^2([0,1]^2)$, which are not necessarily continuous or have continuous sample paths, and to differences of such operators, which are typically not positive semi-definite. If the reader is not interested in the proofs of asymptotic properties, which are provided in the appendix, this section can be skipped upon noting that, for discretely observed data, the shifted (partial) trace corresponds to the standard discretization of the continuous definition above, c.f. Definition 10 below.

In this section, we provide alternative definitions of the shifted (partial) traces, which require neither continuity nor positive semi-definiteness. These will be denoted by "T", replacing "Tr", to make the distinction. It will be shown subsequently that, under continuity, they coincide with Definitions 5 and 6. Note that Remark 3 is still applicable here.

In Definitions 5 and 6, we introduced shifts to integrate along off-diagonals, in hope to avoid the banded part of our process. In the following definition, we rather shift the whole kernel, and then integrate along the diagonal of the shifted kernel. Intuitively, the two approaches are equivalent, but some work is required to show this formally.

Definition 7. *We define the shifting operator* $\mathrm{S}^\delta : \mathcal{S}_1(\mathcal{L}^2([0,1])) \to \mathcal{S}_1(\mathcal{L}^2([0,1]))$ *by its action on kernels. For* $F \in \mathcal{S}_1(\mathcal{L}^2([0,1]))$ *with a kernel* $k = k(t,s)$*,* $\mathrm{S}^\delta(F)$ *have kernel*

$$k^\delta(t,s) = \begin{cases} k(t, s+\delta), & s < 1-\delta, \\ 0, & \textit{otherwise}. \end{cases} \tag{2.4}$$

It is straightforward to check that S^δ is a well-defined linear operator on $\mathcal{S}_1(\mathcal{L}^2([0,1]))$. To check boundedness, let $F = \sum_j \sigma_j g_j \otimes h_j$ be the SVD of F. Then we have $k^\delta(t,s) = \sum_j \sigma_j g_j(t) h_j^\delta(s)$, where the equality is understood in the $\mathcal{L}^2$-sense, and $h_j^\delta(s) = h_j(s+\delta)$

for $s \leq 1-\delta$, and $h_j^\delta(s) = 0$ otherwise. Then

$$\begin{aligned}\left|\!\left|\!\left|\mathrm{S}^\delta(F)\right|\!\right|\!\right|_1 &= \left|\!\left|\!\left|\sum_{j=1}^\infty \sigma_j \mathrm{S}^\delta(g_j \otimes h_j)\right|\!\right|\!\right|_1 \leq \sum_{j=1}^\infty \sigma_j \left|\!\left|\!\left| g_j \otimes h_j^\delta \right|\!\right|\!\right|_1 \\ &= \sum_{j=1}^\infty \sigma_j \|g_j\| \|h_j^\delta\| \leq \sum_{j=1}^\infty \sigma_j = |\!|\!| F |\!|\!|_1\end{aligned}$$

where we used the triangle inequality (the first inequality) and the fact that $\|h_j^\delta\| \leq \|h_j\|$ (the second inequality). The calculation above shows that $\left|\!\left|\!\left|\mathrm{S}^\delta\right|\!\right|\!\right|_\infty \leq 1$.

Definition 8. *(a) For $F \in \mathcal{S}_1(\mathcal{L}^2([0,1]))$, we define $T^\delta(F) = Tr(S^\delta F)$.*

(b) For $F \in \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$ we define $T^\delta(F) = Tr\Big[(S^\delta \otimes S^\delta)F\Big]$.

T^δ is well defined bounded linear functional on $\mathcal{S}_1(\mathcal{L}^2([0,1]))$ or $\mathcal{S}_1(\mathcal{L}^2([0,1]^2))$, since

$$|\mathrm{T}^\delta(F)| = |\mathrm{Tr}(\mathrm{S}^\delta F)| \leq \left|\!\left|\!\left|\mathrm{S}^\delta F\right|\!\right|\!\right|_1 \leq \left|\!\left|\!\left|\mathrm{S}^\delta\right|\!\right|\!\right|_\infty |\!|\!| F |\!|\!|_1 \leq |\!|\!| F |\!|\!|_1,$$

where the second inequality is of Hölder-type for Schatten-p spaces.

Definition 9. *For $F \in \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$ separable, i.e. of the form $F = A \tilde{\otimes} B$, we define $T_1^\delta(F) = T^\delta(B)A$.*

For other than separable operators, T_1^δ is defined by linear extension. Such a construction is viable due to the following proposition.

The proofs of the following two propositions borrow ideas from Aston et al. (2017).

Proposition 6. *Let $\delta \geq 0$, then $T_1^\delta : \mathcal{S}_1(\mathcal{L}^2([0,1]^2)) \to \mathcal{S}_1(\mathcal{L}^2([0,1]))$ is well defined, linear, and bounded. Moreover, for $F \in \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$ we have*

$$Tr(G T_1^\delta(F)) = Tr([S^\delta \tilde{\otimes} G]F), \quad \forall G \in \mathcal{S}_1(\mathcal{L}^2([0,1])). \tag{2.5}$$

Proof. Let $F = \sum_{r=1}^R A_r \tilde{\otimes} B_r$. Then for any $G \in \mathcal{S}_1(\mathcal{L}^2([0,1]))$ we have

$$\begin{aligned}\mathrm{Tr}(G\ \mathrm{T}_1^\delta(F)) &= \sum_{r=1}^R \mathrm{Tr}(S^\delta B_r)\mathrm{Tr}(GA_r) = \sum_{r=1}^R \mathrm{Tr}\Big(GA_r) \tilde{\otimes} (S^\delta B_r)\Big] \\ &= \sum_{r=1}^R \mathrm{Tr}\Big[(G \tilde{\otimes} S^\delta)(A_r \tilde{\otimes} B_r)\Big] = \mathrm{Tr}\Big[(G \tilde{\otimes} S^\delta)F\Big]\end{aligned} \tag{2.6}$$

By Lemma 1.6 of the supplementary material of Aston et al. (2017), the space

$$\mathcal{X} := \left\{ \sum_{r=1}^R A_r \tilde{\otimes} B_r \;\middle|\; A_r, B_r \in \mathcal{S}_1(\mathcal{L}^2([0,1])), r \in \mathbb{N} \right\}$$

is dense in $\mathcal{S}_1(\mathcal{L}^2([0,1]^2))$. Using the following characterization of the trace norm,

$$\vert\vert\vert F\vert\vert\vert_1 = \sup_{\vert\vert\vert G\vert\vert\vert_\infty=1} |\mathrm{Tr}(GF)|,$$

we obtain from (2.6) that

$$\begin{aligned}\left\vert\left\vert\left\vert T_1^\delta(F)\right\vert\right\vert\right\vert_1 &= \sup_{\vert\vert\vert G\vert\vert\vert_\infty=1} |\mathrm{Tr}(GT_1^\delta(F))| = \sup_{\vert\vert\vert G\vert\vert\vert_\infty=1} \left|\mathrm{Tr}\left[(G\,\tilde{\otimes}\,S^\delta)F\right]\right| \\ &\leq \sup_{\vert\vert\vert U\vert\vert\vert_\infty=1} |\mathrm{Tr}(UF)| = \vert\vert\vert F\vert\vert\vert_1,\end{aligned} \tag{2.7}$$

since $\left\vert\left\vert\left\vert \mathrm{S}^\delta\right\vert\right\vert\right\vert_\infty \leq 1$.

Hence T_1 can be extended continuously to $\mathcal{S}_1(\mathcal{L}^2([0,1]^2))$. Equation (2.5) now follows from (2.6) also by continuity, and we have

$$\left\vert\left\vert\left\vert \mathrm{Tr}_1^\delta(F)\right\vert\right\vert\right\vert_1 = \sup_{\vert\vert\vert G\vert\vert\vert_\infty=1} \left|\mathrm{Tr}\left[(G\,\tilde{\otimes}\,S^\delta)F\right]\right|. \tag{2.8}$$

for any F. $\square$

The following proposition states that the functional specified in Definition 8 and the operator specified in Definition 9 correspond under the continuity assumption to the shifted trace and the shifted partial trace, respectively.

Proposition 7. *Let $A \in \mathcal{S}_1^+(\mathcal{L}^2([0,1]))$ and $F \in \mathcal{S}_1^+(\mathcal{L}^2([0,1]^2))$ have continuous kernels $a = a(t,s)$ and $k = k(t,s,t',s')$. Then $T^\delta(A) = Tr^\delta(A)$ and $T_1^\delta(F) = Tr_1^\delta(F)$.*

Proof. We begin by showing the assertion for the shifted trace. We define the continuous version of the shifting operator S^δ, denoted as S_τ^δ. It is defined by Definition 7 with S^δ replaced by S_τ^δ and k^δ replaced by

$$k_\tau^\delta(t,s) = \begin{cases} k(t,s+\delta), & s < 1-\delta-\tau, \\ (s+\delta+\tau)k(t,1-\delta-\tau) + (s+\delta-\tau)k(t,1-\delta+\tau), & |s-(1-\delta)| \leq \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Then by continuity, $\mathrm{Tr}(S_\tau^\delta F) \overset{\tau\to 0_+}{\longrightarrow} \mathrm{T}^\delta(F)$ and at the same time $\mathrm{Tr}(S_\tau^\delta F) \overset{\tau\to 0_+}{\longrightarrow} \mathrm{Tr}^\delta(F)$, implying the equality of the limits.

We now proceed to the shifted partial trace. By Lemma 1.7 of the supplementary material of Aston et al. (2017), for any $\epsilon > 0$ there exists a finite rank operator $F_R = \sum_{r=1}^R A_r \,\tilde{\otimes}\, B_r$ with kernel k_R such that $\vert\vert\vert F - F_R\vert\vert\vert_1 < \epsilon$ and $\|k - k_R\|_\infty < \epsilon$.

Fixing $\epsilon > 0$, we have from the triangle inequality that

$$\begin{aligned}\left|\!\left|\!\left|\mathrm{T}_1^\delta(F) - \mathrm{Tr}_1^\delta(F)\right|\!\right|\!\right|_2 \leq \left|\!\left|\!\left|\mathrm{T}_1^\delta(F) - \mathrm{T}_1^\delta(F_R)\right|\!\right|\!\right|_2 &+ \left|\!\left|\!\left|\mathrm{T}_1^\delta(F_R) - \mathrm{Tr}_1^\delta(F_R)\right|\!\right|\!\right|_2 \\ &+ \left|\!\left|\!\left|\mathrm{Tr}_1^\delta(F_R) - \mathrm{Tr}_1^\delta(F)\right|\!\right|\!\right|_2\end{aligned}$$

The middle term is zero, which follows from linearity of the operators and the first half of this proof.

The first term can be bounded by

$$\left|\!\left|\!\left|\mathrm{T}_1^\delta(F - F_R)\right|\!\right|\!\right|_1 \leq |||F - F_R|||_1 < \epsilon,$$

and for the final term we have

$$\begin{aligned}\left|\!\left|\!\left|\mathrm{Tr}_1^\delta(F_R - F)\right|\!\right|\!\right|_2 &= \left(\int_0^1 \int_0^1 \left(\int_0^1 \Big[k_R(t,s,t',s+\delta) - k(t,s,t',s+\delta)\Big] ds\right)^2 dt dt'\right)^{1/2} \\ &\leq \|k - k_R\|_\infty\end{aligned}$$

Altogether, we have that $\left|\!\left|\!\left|\mathrm{T}_1^\delta(F) - \mathrm{Tr}_1^\delta(F)\right|\!\right|\!\right|_2 < 2\epsilon$. Since ϵ was arbitrarily small, the proof is complete. □

The development of shifted partial tracing with respect to the second argument can be done similarly. Also, Proposition 5 holds with the general definitions of the shifted (partial) traces and without the assumptions of positive semi-definiteness and continuity placed on A_1 and A_2, and their kernels, respectively. This can be simply checked using the definitions. It thus remains to generalize the proof of Lemma 3 to the case where continuity is not assumed.

Proof of Lemma 3. We will use two simple auxiliary results, a certain decomposition of B that will be only introduced in the next chapter, and the limiting argument of Gohberg and Krein (1978), c.f. equation (1.12). Firstly, it holds for any operators A and B that

$$(A \tilde{\otimes} B) = (A \tilde{\otimes} Id)(Id \tilde{\otimes} B), \tag{2.9}$$

which can be verified on the rank one elements:

$$(A \tilde{\otimes} Id)(Id \tilde{\otimes} B)(x \otimes y) = (A \tilde{\otimes} Id)(x \otimes By) = a\ x \otimes By = (A \tilde{\otimes} B)(x \otimes y).$$

Secondly, we have formula (2.8).

Hence it suffices to show that $\mathrm{Tr}\Big[(G \tilde{\otimes} S^\delta)B\Big] = 0$ for any G with unit operator norm. Let $B = \sum_{r=1}^\infty \sigma_r \widetilde{U}_r \tilde{\otimes} V_r$ be the separable component decomposition of B, see Chapter 3. Let $U_r := \sigma^r \widetilde{U}_r$, so $B = \sum_{r=1}^\infty U_r \tilde{\otimes} V_r$. We know by Lemma 6 that for B banded, all V_r, $r = 1, 2, \ldots$ are banded as well, with the same bandwidth. Using (2.9) and cyclicity of trace, we have

$$\operatorname{Tr}\Big[(G \tilde{\otimes} S^\delta)B\Big] = \operatorname{Tr}\Big[(Id \tilde{\otimes} S^\delta)B(G \tilde{\otimes} Id)\Big].$$

Using (2.9) again, we obtain

$$B(G \tilde{\otimes} Id) = \sum_{r=1}^{\infty} (GU_r) \tilde{\otimes} V_r,$$

which is still banded in the spatial dimensions (corresponding to V's). Therefore, $(Id \tilde{\otimes} S^\delta)B(G \tilde{\otimes} Id)$ has a kernel, which is 0 along the diagonal. Since that kernel clearly gives rise to a trace-class operator, the limiting argument (1.12) can be used, showing that the trace is zero. ☐

We have just shown that the conclusions of Section 2.1 remain valid even without the assumption of continuity. In practice, however, data is observed discretely, and shifted partial tracing needs to be performed on a tensor.

Definition 10. *Let* $\mathbf{M} \in \mathbb{R}^{K_1 \times K_2 \times K_1 \times K_2}$. *For* $d \leq \min(K_1, K_2)$, *we define*

$$\begin{aligned}
Tr_1^d(\mathbf{M})[i,k] &= \sum_{j=1}^{K_2 - d} \mathbf{M}[i,j,k,j+d]\,, && i,k = 1,\ldots,K_1\,, \\
Tr_2^d(\mathbf{M})[j,l] &= \sum_{i=1}^{K_1 - d} \mathbf{M}[i,j,i+d,l]\,, && j,l = 1,\ldots,K_2\,, \\
Tr^d(\mathbf{M}) &= \sum_{i=1}^{K_1 - d} \sum_{j=1}^{K_2 - d} \mathbf{M}[i,j,i+d,j+d]\,.
\end{aligned} \tag{2.10}$$

The previous definition is in a natural agreement with Definition 6, as stated in the following lemma.

Lemma 4. *Let* $\mathbf{M} \in \mathbb{R}^{K \times K \times K \times K}$. *Let* $F \in \mathcal{S}_2(\mathcal{L}^2([0,1]^2))$ *be the pointwise continuation of* $\mathbf{M}$, *i.e. the kernel* k *of* F *is given by*

$$k(t,s,t',s') = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} \mathbf{M}[i,j,k,l] \mathbb{1}_{[(t,s) \in I_{i,j}^K]} \mathbb{1}_{[(t',s') \in I_{k,l}^K]}\,,$$

where $I_{i,j} = \Big[\frac{i-1}{K_1}, \frac{i}{K_1}\Big) \times \Big[\frac{j-1}{K_2}, \frac{j}{K_2}\Big)$. *Then*

- *(a) For* $\delta \in [0,1)$ *such that* $d := \delta K_2 \in \mathbb{N}_0$, *we have* $\left|\!\left|\!\left| Tr_1^\delta(F) \right|\!\right|\!\right|_2 = K_1^{-1} K_2^{-1} \left\| Tr_1^d(\mathbf{M}) \right\|_F$.
- *(b) For* $\delta \in [0,1)$ *such that* $d := \delta K_1 \in \mathbb{N}_0$, *we have* $\left|\!\left|\!\left| Tr_2^\delta(F) \right|\!\right|\!\right|_2 = K_1^{-1} K_2^{-1} \left\| Tr_2^d(\mathbf{M}) \right\|_F$.
- *(c) For* $K_1 = K_2 =: K$ *and* $\delta \in [0,1)$ *such that* $d := \delta K \in \mathbb{N}_0$, *we have* $Tr^\delta(F) = K_1^{-1} K_2^{-1} Tr^d(\mathbf{M})$.

Proof. We only show the first part, since the other two parts are similar.

Let $g_{i,j}(t,s) = \sqrt{K_1 K_2}\mathbb{1}_{[(t,s)\in I^K_{i,j}]}$. Since

$$k(t,s,t',s') = K_1^{-1}K_2^{-1}\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2}\mathbf{M}[i,j,k,l]g_{i,j}(t,s)g_{k,l}(t',s')\,,$$

we can express F as

$$F = K_1^{-1}K_2^{-1}\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2}\mathbf{M}[i,j,k,l]g_{i,j}\otimes g_{k,l}\,.$$

It now follows from linearity of shifted partial tracing that

$$\mathrm{Tr}_1^\delta(F) = K_1^{-1}K_2^{-1}\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2}\mathbf{M}[i,j,k,l]\mathrm{Tr}_1^\delta(g_{i,j}\otimes g_{k,l})\,. \tag{2.11}$$

Since $I^K_{i,j}$ is a cartesian product of two intervals, we can write $I^K_{i,j} = I^K_i \times I^K_j$. Then $g_{i,j} = g_i^{(1)}\otimes g_j^{(2)}$ with $g_i^{(1)}(t) = \sqrt{K_1}\mathbb{1}_{[t\in I_i^K]}$ and $g_j^{(2)}(s) = \sqrt{K_1}\mathbb{1}_{[s\in I_j^K]}$. Furthermore,

$$g_{i,j}g_{k,l} = g_i^{(1)}\otimes g_i^{(2)}\otimes g_k^{(1)}\otimes g_l^{(2)} = (g_i^{(1)}\otimes g_k^{(1)})\,\tilde{\otimes}\,(g_j^{(2)}\otimes g_l^{(2)})$$

and hence by Definition 9 we have $\mathrm{Tr}_1^\delta(g_{i,j}\otimes g_{k,l}) = \mathrm{Tr}^\delta(g_j^{(2)}\otimes g_l^{(2)})g_i^{(1)}\otimes g_k^{(1)}$. Note that due to the limiting argument (1.12), $\mathrm{Tr}^\delta(g_j^{(1)}\otimes g_l^{(1)}) = \mathbb{1}_{[j=l+\delta K_1]}$, hence from (2.11) we have

$$\mathrm{Tr}_1^\delta(F) = K_1^{-1}K_2^{-1}\sum_{i=1}^{K_1}\sum_{k=1}^{K_1}\left(\sum_{j=1}^{(1-\delta)K_2}\mathbf{M}[i,j,k,j+\delta K]\right)g_i^{(2)}\otimes g_k^{(2)}\,. \tag{2.12}$$

Thus it is $\left\|\left|\mathrm{Tr}_1^\delta(F)\right|\right\|_2 = K_1^{-1}K_2^{-1}\left[\sum_{i=1}^{K_1}\sum_{k=1}^{K_1}\left(\sum_{j=1}^{(1-\delta)K_2}\mathbf{M}[i,j,k,l]\right)^2\right]^{1/2}$, while in the discrete case we have $\left\|\mathrm{Tr}_1^d(\mathbf{M})\right\|_F = \left[\sum_{i=1}^{K_1}\sum_{k=1}^{K_1}\left(\sum_{j=1}^{(1-\delta)K_2}\mathbf{M}[i,j,k,l]\right)^2\right]^{1/2}$ from Definition 10. ☐

Note that we have actually proven something more general. We can write from (2.12) that the kernel of $\mathrm{Tr}_1^\delta(F)$ is

$$k_1(t,t') = \sum_{i=1}^{K_1}\sum_{k=1}^{K_1}\left(\frac{1}{K_2}\sum_{j=1}^{(1-\delta)K_2}\mathbf{M}[i,j,k,j+\delta K]\right)\mathbb{1}_{t\in I_i}\otimes\mathbb{1}_{t'\in I_k}\,,$$

where the term inside the parentheses is almost the (i,k)-th element of discrete partial tracing, but instead of summing in the discrete case we have to average in the continuous case (which corresponds to the difference between the Lebesque measure on piecewise constant function on $[0,1]$ with at most K jumps and the counting measure on the set $\{1,\dots,K\}$).

Shifted partial tracing could still have been defined in slightly greater generality. However, the definition requires the notion of a "shift" and hence it requires an explicit set to act on. We could instead of $\mathcal{L}^2([0,1]^2)$ take $\mathcal{L}^2(\Omega)$ with $(\Omega,\mathcal{A},\mu)$ a measure space with Ω a linearly ordered metric space and μ a finite measure. The choice of $\Omega=\{1,\dots,K_1\}\times\{1,\dots,K_2\}$ and μ being the counting measure would then lead to Definition 10. We have not gone down this path since this formalism would not be particularly useful in practice anyway. Note, however, that Definitions 6 and 10 are compatible in this way, with the difference between them stemming from the change of measure, as depicted in Lemma 4.

2.2 Estimation

We assume here availability of N independent (and w.l.o.g. zero-mean) surfaces, say $X_1,\dots,X_N$, with covariance satisfying the separable-plus-banded model (2.1), where δ is such that $\mathrm{Tr}^\delta(A_1)$ and $\mathrm{Tr}^\delta(A_2)$ are non-zero. For now, let the surfaces be fully observed; discrete observations are considered in Sections 2.3 and 2.4. Firstly, we focus on estimation of the separable part of model (2.1) by shifted partial tracing.

The following example explains how 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 $X\in\mathcal{L}^2([0,1]^2)$ with a separable covariance $C=C_1\otimes C_2$, which has a continuous kernel $c(t,s,t',s')=c_1(t,t')c_2(s,s')$. Assume for simplicity that Tr(C_1) = Tr$(C_2)=1$. Partial tracing (without shifting, i.e. $\delta=0$) 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 $\{X^t(s)\}_{t\in[0,1]}$, i.e. any given time point t is providing us with a single curve $X^t(s)$, $s\in[0,1]$. This spatial sample is used to estimate the spatial covariance C_2 in a standard way, i.e. outer products $X^t\otimes X^t$ are formed and averaged together as

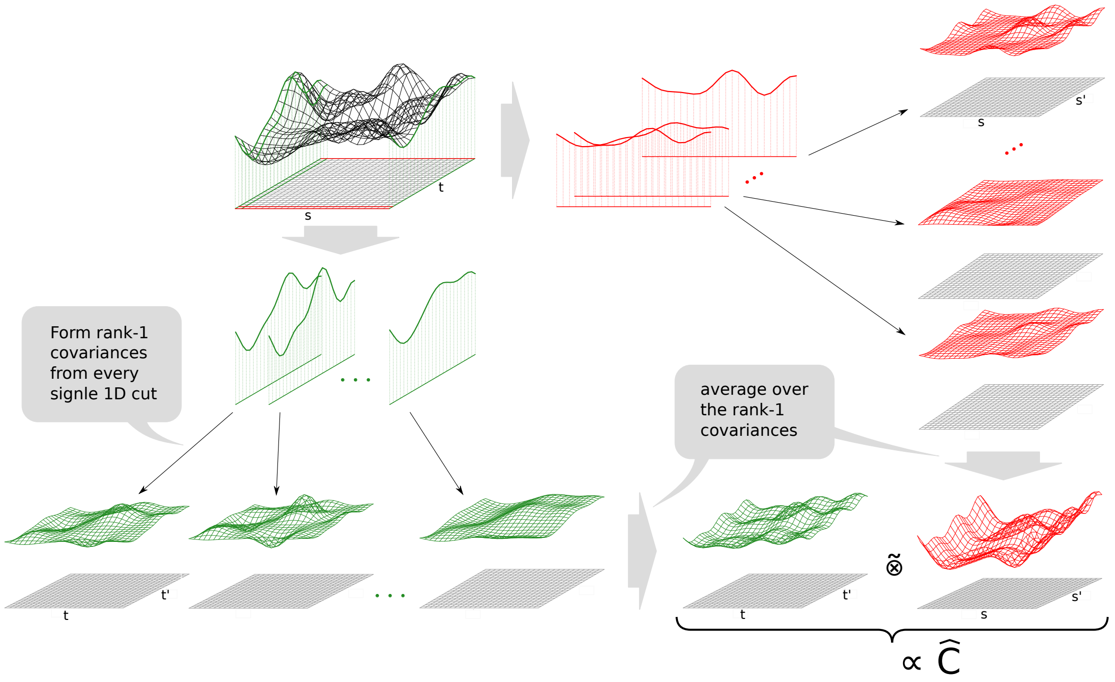
$$\widehat{C}_2=\int_0^1 X^t\otimes X^t\,\mathrm{d}t \qquad \textit{or equivalently} \qquad \widehat{c}_2(s,s')=\int_0^1 X^t(s)X^t(s')\,\mathrm{d}t\,.$$

This is a moment estimator in a sense, since $\mathbb{E}(X^t\otimes X^t)\propto C_2$ for any $t\in[0,1]$. Similarly for the temporal domain: a temporal sample $\{X^s(t)\}_{s\in[0,1]}$ is formed by cutting X along the spatial domain, and the temporal covariance is then estimated as

$$\widehat{C}_1=\int_0^1 X^s\otimes X^s ds \qquad \textit{or equivalently} \qquad \widehat{c}_1(t,t')=\int_0^1 X^s(t)X^s(t')\,\mathrm{d}t\,.$$

This process is captured in Figure 2.1. When multiple surfaces are observed, the described

Figure 2.1: Estimation of a separable model via partial tracing based on a single observation. The observation is cut along the temporal domain to obtain a temporal sample (in green), from which the temporal part of the separable covariance is empirically estimated. Similarly for the spatial part (in red).


s
t
Form rank-1 covariances from every signle 1D cut
t
t'
s
s'
average over the rank-1 covariances
$\tilde{\otimes}$
$\propto \widehat{C}$

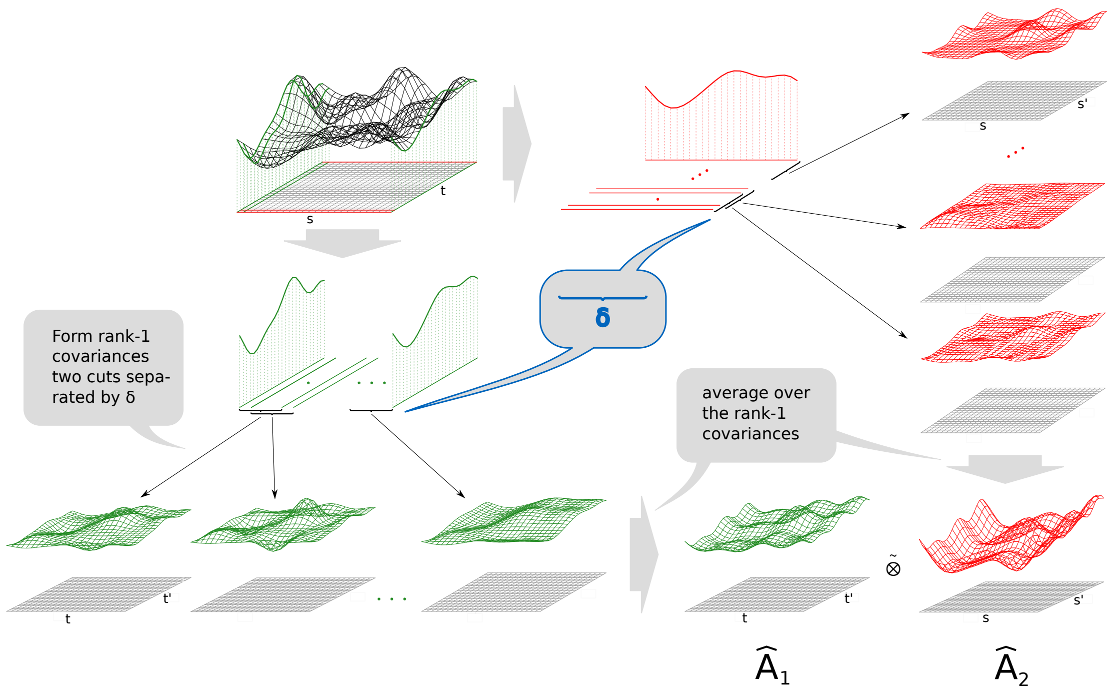
procedure is repeated for all of them, and the results are averaged together for the resulting estimator.

When the covariance is instead separable-plus-banded, i.e. $C = A_1 \otimes A_2 + B$ *with* B *banded by* δ*, it is no longer true that* $\mathbb{E}(X^t \otimes X^t) \propto A_2$*, but it is still true that* $\mathbb{E}(X^t \otimes X^{t+\delta}) \propto A_2$ *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 products for* $t \in [0, 1-\delta]$ *to obtain a scaled estimator of* A_2*, see Figure 2.2. In other words, one estimates the temporal factor in the separable part of the model by cutting the observations along the temporal axis and introducing a spatial shift when taking outer products.* A_1 *can be estimated in a similar way (cutting the observations along the spatial axis and introducing a temporal shift), and the only remaining question is how to determine the scaling constants.*

Using Lemma 3 together with Proposition 5 (c), we obtain the following estimating equation for model (2.1):

$$\mathrm{Tr}^{\delta}(C) A_1 \tilde{\otimes} A_2 = \mathrm{Tr}^{\delta}_1(C) \tilde{\otimes} \mathrm{Tr}^{\delta}_2(C). \tag{2.13}$$

Figure 2.2: Estimation of the separable-plus-banded model via shifted partial tracing based on a single observation. The observation is cut along the temporal domain to obtain a temporal sample (in green), from which the temporal part of the separable covariance is empirically estimated by introducing a shift in space. Similarly for the spatial part (in red).


s
t
δ
Form rank-1 covariances two cuts sepa-rated by δ
average over the rank-1 covariances
s
s'
t
t'
t
t'
s
s'
$\tilde{\otimes}$
$\widehat{A}_1$
$\widehat{A}_2$

Equation (2.13) suggests the following estimators for the separable part of the model:

$$\widehat{A}_1 = \mathrm{Tr}_1^\delta(\widehat{C}_N) \qquad \& \qquad \widehat{A}_2 = \frac{\mathrm{Tr}_2^\delta(\widehat{C}_N)}{\mathrm{Tr}^\delta(\widehat{C}_N)}, \tag{2.14}$$

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 $\mathrm{Tr}^\delta(\widehat{C}_N) \neq 0$. Once the separable part of the model has been estimated, we can define

$$\widehat{B} = \widehat{C}_N - \widehat{A}_1 \tilde{\otimes} \widehat{A}_2 \, . \tag{2.15}$$

Optionally, we can set the kernel of $\widehat{B}$ to zero outsize of the band of size δ. Note that none of the estimators defined above is guaranteed to be symmetric or positive semi-definite. However, this is just a technicality, which can be dealt with easily, see Section 2.3.4.

Consider the separable-plus-banded model $\mathbf{C} = \mathbf{A}_1 \tilde{\otimes} \mathbf{A}_2 + \mathbf{B}$ with $\mathbf{B}$ banded by d, i.e. $\mathbf{B}[i,j,k,l] = 0$ whenever $\min(|i-k|,|j-l|) \geq d$. We denote by d the discrete version of the bandwidth δ; the relation for an equidistant grid of size $K \times K$ is $d = \lceil \delta K \rceil + 1$. It is straightforward to translate Proposition 5 and Lemma 3 to the discrete case to obtain the following estimating equation

$$\mathrm{Tr}^d(\mathbf{C}) \mathbf{A}_1 \tilde{\otimes} \mathbf{A}_2 = \mathrm{Tr}_1^d(\mathbf{C}) \tilde{\otimes} \mathrm{Tr}_2^d(\mathbf{C}) \, ,$$

suggesting again the plugin estimators

$$\widehat{\mathbf{A}}_1 = \mathrm{Tr}_1^d(\widehat{\mathbf{C}}_N) \quad \text{and} \quad \widehat{\mathbf{A}}_2 = \mathrm{Tr}_2^d(\widehat{\mathbf{C}}_N)/\mathrm{Tr}^d(\widehat{\mathbf{C}}_N)\,. \tag{2.16}$$

It may be useful to revisit Example 1 and Figure 2.1 (which is plotted discretely anyway) for intuitive depiction of these definitions.

Remark 4. *While the proposed estimators are based on the empirical covariance, the empirical covariance is only considered implicitly. It is never even calculated, let alone stored. Computation of the empirical covariance is outside of the computational limits we set for ourselves. We seek a methodology which would be as efficient as matrix-matrix multiplication between pairs of the sampled observations.*

2.2.1 Stationarity

As previously noted, model (2.1) is interesting from two perspectives, depending on whether the banded part b is only seen as a nuisance or whether it is also of interest (e.g. when the inverse of $a + b$ needs to be applied for the purpose of prediction):

- If b is indeed an estimand of interest, then one needs to make further structural assumptions on b, for reasons of computational and statistical efficiency. This is because the bandwidth $\delta > 0$ is assumed constant and non-decreasing in N or K – consequently, even though b is banded, it has the same order of entries as c itself, when observed on a grid. In our development, we show how one can also estimate b under the additional assumption that it is stationary (Hall et al., 1994; Cai et al., 2013). We focus on stationarity as a specific assumption which seems broadly applicable and yields a form of parsimony complementary to separability. Under this additional assumption, we show in detail that both a and b of model (2.2) 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, i.e. estimation, application, and inversion of the covariance, can be performed at the same cost as matrix-matrix multiplication of two sampled observations.
- Contrarily, if the banded part b is only viewed as a nuisance, no additional assumption to bandedness needs to be made to allow for the estimation of the separable part a. This is the case when the data are believed to be separable up to a weakly dependent contamination, which carries no information of any value for an analyst, i.e. only long-term dependencies are of interest. In this case again, estimation, application and inversion of the covariance (including noise) can be performed at the asymptotic cost of matrix-multiplying two sampled observations.
- A special case in the previous considerations is when the discrete bandwidth is

equal to one, i.e. a separable covariance model is observed under additional white noise. In that case, the covariance of the observations is separable-plus-diagonal. This situation corresponds to the widely used errors-in-measurements model (1.16). If we allow for heteroscedasticity of noise, the diagonal has the same number of (discrete) degrees of freedom as the separable model.

Hence we see that shifted partial tracing can be used to estimate a separable model under noisy regimes, which can be either heteroscedastic (separable-plus-diagonal) or weakly dependent (separable-plus-stationary), and we will show that in both of these cases, no additional computational costs have to be paid compare to when separability is assumed. Contrarily, when both weakly separable and heteroscedastic noise is present, shifted partial tracing can still be used to estimate the separable part of the model at the same cost, but when an estimator of the error structure is also needed (e.g. for the purposes of prediction), additional computational costs are present due to the fact that the noise is more complicated than the signal, in this particular case.

If at this point we add the stationarity of B into our assumptions (i.e. let the kernel b be translation invariant: $b(t,s,t',s') = \varsigma(|t-t'|,|s-s'|)$, $t,t',s,s' \in [0,1]$, where $\varsigma \in \mathcal{L}^2([0,1]^2)$ is the *symbol* of B) we take the following estimator of B instead of (2.15):

$$\widehat{B} = \mathrm{Ta}(\widehat{C}_N - \widehat{A}_1 \tilde{\otimes} \widehat{A}_2), \tag{2.17}$$

where $\mathrm{Ta}(\cdot)$ is the "Toeplitz averaging" operator, i.e. the projection onto the stationary operators, defined as follows.

Definition 11. *For $F \in S_1(\mathcal{L}^2([0,1]^2))$ self-adjoint and $\{e_j\}_{j\in\mathbb{Z}}$ the complete orthonormal basis of trigonometric functions in $\mathcal{L}^2([0,1])$, let*

$$F = \sum_{i,j,k,l\in\mathbb{Z}} \gamma_{ijkl}(e_i \otimes e_j) \otimes (e_k \otimes e_l). \tag{2.18}$$

Then we define

$$Ta(F) = \sum_{i,j\in\mathbb{Z}} \gamma_{ijij}(e_i \otimes e_j) \otimes (e_i \otimes e_j). \tag{2.19}$$

Let us comment on the previous definition. If $\{e_j\}_{j\in\mathbb{Z}}$ is the trigonometric basis on $\mathcal{L}^2([0,1])$, then $\{e_i \otimes e_j\}_{i,j\in\mathbb{Z}}$ is the trigonometric basis on $\mathcal{L}^2([0,1])^2$, so every compact operator F can be expressed with respect to this basis as in (2.18). For F trace class, the Fourier coefficients $\{\gamma_{ijkl}\}$ are absolutely summable, leading to $\mathrm{Ta}(F)$ in (2.19) being also trace-class. Secondly, a stationary operator has the trigonometric basis as its eigenbasis, as shown below. Thirdly, $\mathrm{Ta}(\cdot)$ as defined in (2.19) is clearly an orthogonal projection. Altogether, $\mathrm{Ta}(\cdot)$ is the orthogonal projection onto the space of stationary operators in $S_1(\mathcal{L}^2([0,1]^2))$, which is itself a Banach space.

Now we show that a self-adjoint stationary integral operator on $\mathcal{L}^2([0,1])$ has the Fourier basis as its eigenbasis. We work with $\mathcal{L}^2([0,1])$ for simplicity, the argument translates easily to higher dimensions.

Let F be a stationary integral operator on $\mathcal{L}^2([0,1])$ with kernel $k = k(u_1, u_2)$, i.e. $k(u_1, u_2) = h(u_1 - u_2)$, $t, s \in [0,1]$, for a symmetric function $h : [-1,1] \to \mathbb{R}$. We expand h into its Fourier series as $h(x) = \sum_{j\in\mathbb{Z}} \phi_j e^{-2\pi i j x}$. Thus we have

$$k(u_1, u_2) = \sum_{j\in\mathbb{Z}} \phi_j e^{-2\pi i j u_1} e^{2\pi i j u_2}.$$

To see that the previous expansion is in fact an eigen-decomposition, note that for $l = 0, 1, \ldots$, we have

$$\int_0^1 k(t,s) e^{-2\pi i l u_2}\, \mathrm{d}s = \sum_{j\in\mathbb{Z}} \phi_j e^{-2\pi i j u_1} \int_0^1 e^{-2\pi i (l-j) u_2}\, \mathrm{d}s = \theta_l e^{-2\pi i l u_1},$$

and similarly for $-l \in \mathbb{N}$ due to self-adjointness.

The previous justifies the definition of the Toeplitz averaging operator in the continuous case. In the discrete case, Toeplitz averaging is defined as follows.

Definition 12. *For* $\mathbf{F} \in \mathbb{R}^{K_1\times K_2\times K_1\times K_2}$, *we define*

$$\mathbf{S}[h,l] = \frac{1}{K^2} \sum_{i=1}^{K-h} \sum_{j=1}^{K-l} \mathbf{F}[i, j, i+h-1, j+l-1] \tag{2.20}$$

for $h = 1, \ldots, K_1$ *and* $l = 1, \ldots, K_2$, *and* $Ta(\mathbf{F}) \in \mathbb{R}^{K_1\times K_2\times K_1\times K_2}$ *is the tensor having* $\mathbf{S}$ *as its symbol, i.e.* $Ta(\mathbf{F})[i,j,k,l] = \mathbf{S}[1+|i-k|, 1+|j-l|]$ *for* $i,k = 1,\ldots,K_1$ *and* $j,l = 1,\ldots,K_2$.

Unlike in continuum, the discrete Fourier basis is not necessarily the eigenbasis of a stationary operator, hence the need for an alternative definition, which does not bear immediate resemblance with Definition 11. Formula (2.20) directly utilizes the stationarity assumption by averaging over the elements that ought to be the same (under the stationarity assumption), hence the name “Toeplitz averaging”. The relation to the discrete Fourier basis, which is important for efficient manipulation, is discussed in Section 2.3.2.

2.2.2 Choice of Bandwidth

It remains to provide means to choose the band size δ, in order to make the methodology applicable in practice. Recall that δ has to be large enough to eliminate B from the model (2.1), but small enough so $\mathrm{Tr}^{\delta}(A_1 \tilde{\otimes} A_2)$ does not vanish. We develop a strategy, which picks δ among some candidate values. In practice, data is observed discretely, so a finite set of candidate bandwidths is easy to choose. At the same time, a whole range of values for δ is asymptotically indistinguishable (we will formally observe this in Section 2.4). Hence the number of candidate values should increase with the grid size only up to a certain reasonable value.

In this section, we make the estimator of Section 2.2 depend explicitly on δ as

$$\widehat{C}(\delta) = \widehat{A}_1(\delta) \tilde{\otimes} \widehat{A}_2(\delta) + \widehat{B}(\delta).$$

Even though the separable part of the model $A_1 \tilde{\otimes} A_2$ does not depend on δ, the estimator of the separable part $\widehat{A}_1(\delta) \tilde{\otimes} \widehat{A}_2(\delta)$ does, since shifted partial tracing with some fixed δ is used to obtain the estimator. Note that if we actually knew C, we would use it in formulas (2.14) and (2.17) instead of the empirical estimator $\widehat{C}_N$ to obtain a separable-plus-banded proxy of C, denoted here as

$$C(\delta) = A_1(\delta) \tilde{\otimes} A_2(\delta) + B(\delta).$$

Under the separable-plus-banded model, it is $C(\delta) = C$ for any δ large enough to eliminate B by δ-shifted partial tracing. However, among all such bandwidths, the smaller ones will lead to better empirical performance.

Let $\Delta := \{\delta_1, \ldots, \delta_m\}$ 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

$$\delta^\star := \operatorname*{arg\,min}_{\delta \in \Delta} |||C(\delta) - C|||_2^2. \tag{2.21}$$

Here, $\delta^\star$ is a set. In particular, under model (2.1), $\delta^\star$ contain all such bandwidths $\delta \in \Delta$ that B is banded by δ. We identify $\delta^\star$ with the minimum of this set. This arbitrary choice reflects the fact that δ is a nuisance parameter, not an estimand of interest. And as suggested by Theorem 3, there is a range of valid values, which are asymptotically indistinguishable.

Since we do not know C we cannot evaluate the objective in (2.21). Instead, we propose to approximate the objective by one that is fully calculable:

$$\widehat{\delta} := \operatorname*{arg\,min}_{\delta \in \Delta} \left|\left|\left|\widehat{C}(\delta)\right|\right|\right|_2^2 - \frac{2}{N} \sum_{n=1}^{N} \langle X_n, \widehat{C}_{-n}(\delta) X_n \rangle, \tag{2.22}$$

where $\widehat{C}_{-n}(\delta)$ is our estimator constructed without the n-th observation X_n. It can be shown that (2.22) is root-n consistent for (2.21) up to a constant (see Proposition 15), which provides a justification for the adaptive choice of bandwidth. We cannot speak of consistency here, since a whole range of values for δ is asymptotically undistinguishable, provided the separable-plus-banded model holds. However, we will see that (even if the separable-plus-banded model is not valid) the adaptively chosen bandwidth $\widehat{\delta}$ leads to a separable-plus-banded proxy of C, denoted $\widehat{C}(\widehat{\delta})$, which is asymptotically optimal in the sense of (2.21).

The procedure above corresponds to a leave-one-out cross-validation (CV), which is computationally prohibitive. In practice, we use 10-fold CV, where we split the data into 10 folds, and every fold gets held-out as a whole, instead of holding out only a single observation as in (2.22).

Finally, let us discuss whether it is reasonable to have a fixed set of candidate values Δ. Should we not allow the number of candidate bandwidth values increase with increasing grid size K? The answer is negative simply because there is a whole range of equally good candidate values (large enough to eliminate the banded part) among which to pick. This range does not depend on K, we only need K large enough such that at least one candidate discrete bandwidth falls inside this range. At the same time, Δ needs to contain this suitable candidate. However, this is always satisfied for a finite grid size K and a finite cardinality of Δ. For example, when when $\mathrm{Tr}^{\delta}(C) \neq 0$ for all $\delta \in (0,1)$ and the true bandwidth $\delta^\star$ is smaller than 0.5, then for an equidistant grid of size $K \geq 2$ it is enough to choose $\Delta = \{1/3, 2/3\}$. Of course, in practice, $\delta^\star$ is unknown, usually much smaller, and we would like to approximate it more closely, so we choose a larger set of candidate values Δ. However, it is clear that the cardinality of Δ should not depend on the grid size K.

2.2.3 Goodness-of-fit Testing

In this section, we develop a testing procedure to check validity of the separable-plus-banded model, generalizing the bootstrap separability test of Aston et al. (2017).

We begin by reviewing the seminal test of Aston et al. (2017). For a covariance $C \in \mathcal{S}_2(\mathcal{H})$ with $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$, a separable proxy is given by

$$C_1 \tilde{\otimes} C_2 = \frac{\mathrm{Tr}_1(C) \tilde{\otimes} \mathrm{Tr}_2(C)}{\mathrm{Tr}(C)},$$

where Tr_1 and Tr_2 are partial traces, i.e. shifted partial traces with the zero shifts. Plugging in the empirical covariance estimator, we obtain

$$\widehat{C}_1 \tilde{\otimes} \widehat{C}_2 = \frac{\mathrm{Tr}_1(\widehat{C}_N) \tilde{\otimes} \mathrm{Tr}_2(\widehat{C}_N)}{\mathrm{Tr}(\widehat{C}_N)}$$

a separable estimator of the covariance. Testing for separability is now based on the following operator

$$D_N = \widehat{C}_N - \widehat{C}_1 \tilde{\otimes} \widehat{C}_2. \tag{2.23}$$

Under the hypothesis of separability, the norm of D_N (a distance to separability) converges to zero as $N \to \infty$.

While a test can be based directly on the asymptotic distribution of D_N (given as a special case of Theorem 3 later), such a test would require full calculation of the empirical covariance, and even worse calculation of the asymptotic variance, which is an eight-dimensional structure. Hence Aston et al. (2017) propose to test separability only on a subspace of $\mathcal{S}_2(\mathcal{H})$ and use bootstrap to avoid calculation of the asymptotic variance. Namely, let $\mathcal{U} = \operatorname{span}\{u_1, \ldots, u_m\} \subset \mathcal{H}$, then $\mathcal{U} \otimes \mathcal{U}$ determines a subspace of $\mathcal{S}_2(\mathcal{H})$ via isometry. Let $T_{\mathcal{U}\otimes\mathcal{U}}$ denote the orthogonal projection to the subspace $\mathcal{U} \otimes \mathcal{U}$. Then we have

$$|||T_{\mathcal{U}\otimes\mathcal{U}} D_N|||_2^2 = \sum_{r=1}^{m} \langle u_r, D_N u_r \rangle^2. \tag{2.24}$$

The most natural choice of $\mathcal{U}$ is given by the eigenfunctions of the separable estimator. There are two reasons for this. Firstly, since we constrain the subspace on which separability will be tested, it makes sense to focus on the subspace on which further analysis (e.g. PCA) will likely be performed. Secondly, using the separable eigenfunctions allows for a fast calculation of the test statistic. Let $\widehat{C}_1 = \sum \widehat{\lambda}_j \widehat{e}_j \otimes \widehat{e}_j$ and $\widehat{C}_2 = \sum \widehat{\gamma}_j \widehat{f}_j \otimes \widehat{f}_j$ be the eigendecompositions and let $\mathcal{U} = \operatorname{span}\{\widehat{e}_i \otimes \widehat{f}_j; i = 1, \ldots, I, j = 1, \ldots, J\}$. Then we have

$$|||T_{\mathcal{U}\otimes\mathcal{U}} D_N|||_2^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \langle \widehat{e}_i \otimes \widehat{f}_j, D_N\, \widehat{e}_i \otimes \widehat{f}_j \rangle^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{1}{N} \sum_{n=1}^{N} \langle X_n, \widehat{e}_i \otimes \widehat{f}_j \rangle^2 - \widehat{\lambda}_j \widehat{\gamma}_j \right)^2 .$$

As for the bootstrap, Aston et al. (2017) propose to approximate the distribution of $|||T_{\mathcal{U}\otimes\mathcal{U}} D_N|||_2^2$ by

$$|||T_{\mathcal{U}\otimes\mathcal{U}}(D_N - D_N^\star)|||_2^2, \tag{2.25}$$

where $D_N^\star$ is the distance-to-separability operator calculated based on a bootstrap sample $\{X_1^\star, \ldots, X_N^\star\}$ drawn from the set $\{X_1, \ldots, X_N\}$ with replacement. The reason for using bootstrap statistic (2.25) instead of simply $|||T_{\mathcal{U}\otimes\mathcal{U}}(D_N^\star)|||_2^2$ is that the latter would approximate the distribution of $|||T_{\mathcal{U}\otimes\mathcal{U}} D_N|||_2^2$ under the true C, i.e. not necessarily under the null, which is that the C is separable.

It is natural to modify the separability test described above by changing the definition of the distance-to-separability (2.23). From now on, let

$$D_N = \widehat{C}_N - \widehat{A}_1 \,\tilde{\otimes}\, \widehat{A}_2 - \widehat{B}, \tag{2.26}$$

where $\widehat{A}_1$, $\widehat{A}_2$ and $\widehat{B}$ are the estimators proposed in (2.14) and (2.17). The test statistic is still taken as (2.24), where $\mathcal{U}$ is still given by the eigenfunctions of the separable part $\widehat{A}_1 \,\tilde{\otimes}\, \widehat{A}_2$. We still approximate the distribution of (2.26) by the bootstrap statistic

$$|||T_{\mathcal{U}\otimes\mathcal{U}}(D_N - D_N^\star)|||_2^2, \tag{2.27}$$

where $D_N^\star$ is calculated like D_N in (2.26) from a bootstrapped sample. Drawing for example 10^3 bootstrap samples, the bootstrapped p-value is given by

$$\frac{1}{10^3+1}\sum_{m=1}^{10^3}\mathbb{1}_{\left[\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \mathcal{T}_{\mathcal{U}\otimes\mathcal{U}}(D_N-D_{N,m}^\star)\right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_2^2 > \vert\kern-0.25ex\vert\kern-0.25ex\vert \mathcal{T}_{\mathcal{U}\otimes\mathcal{U}} D_N \vert\kern-0.25ex\vert\kern-0.25ex\vert_2^2\right]},$$

where $D_{N,m}$ is calculated like in (2.26) from the m-th bootstrapped sample. The statistic can be evaluated efficiently using the linear structure:

$$\vert\kern-0.25ex\vert\kern-0.25ex\vert \mathcal{T}_{\mathcal{U}\otimes\mathcal{U}} D_N \vert\kern-0.25ex\vert\kern-0.25ex\vert_2^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{1}{N}\sum_{n=1}^{N}\langle X_n, \widehat{e}_i\otimes\widehat{f}_j\rangle^2 - \widehat{\lambda}_j\widehat{\gamma}_j - \langle \widehat{e}_i\otimes\widehat{f}_j, \widehat{B}\,\widehat{e}_i\otimes\widehat{f}_j\rangle\right)^2,$$

where $\widehat{B}\widehat{e}_i\otimes\widehat{f}_j$ is calculated using the fast Fourier transform. To implement this test, we modified the codes available in the `covsep` package (Tavakoli, 2016).

The procedure outlined above allows for goodness-of-fit testing of the separable-plus-banded model using the ideas of Aston et al. (2017). It can be vaguely though of as testing whether a separable model holds outside of a band.

2.3 Computational Considerations

Assume availability of discrete observations, i.e. that N independent realizations of $\mathbf{X}\in\mathbb{R}^{K\times K}$ were sampled (let $K_1 = K_2 =: K$ for now, for simplicity) and denoted as $\mathbf{X}_1,\ldots,\mathbf{X}_N$. Firstly, a general covariance tensor $\mathbf{C}$ has $\mathcal{O}(K^4)$ degrees of freedom, while it only has $\mathcal{O}(K^2)$ degrees of freedom under the separability assumption. In comparison, the observed degrees of freedom are only NK^2. Secondly, it takes $\mathcal{O}(NK^4)$ operations to calculate the empirical estimate of the covariance tensor, i.e. $\widehat{\mathbf{C}}_N = \frac{1}{N}\sum_{n=1}^N \mathbf{X}\otimes\mathbf{X}$, while this will be shown to reduce to $\mathcal{O}(NK^3)$ under separability. We assume throughout the thesis that multiplication of two $K\times K$ matrices requires $\mathcal{O}(K^3)$ 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 $\widehat{\mathbf{C}}_N$. 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.

Recall that 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 (2.1) can be estimated and manipulated under the same computational costs as the separable model.

2.3.1 Estimation Complexity

Recall that if we assume that $\mathbf{B}$ is stationary, we take $\widehat{\mathbf{B}} = \mathrm{Ta}(\widehat{\mathbf{C}}_N - \widehat{\mathbf{A}}_1 \tilde{\otimes} \widehat{\mathbf{A}}_2)$, where is the Toeplitz averaging operator. Now we establish the estimation complexity. Firstly, we focus on the shifted partial tracing. Due to linearity, $\mathrm{Tr}_1^d(\widehat{\mathbf{C}}_N) = \frac{1}{N}\sum_n \mathrm{Tr}_1^d(\mathbf{X}_n \otimes \mathbf{X}_n)$, and as can be seen from formula (2.10), only K^3 entries of the total of K^4 entries of $\mathbf{X}_n \otimes \mathbf{X}_n$ are needed to evaluate the shifted partial trace. Moreover, evaluating the shifted partial trace amounts to averaging over one dimension of the 3D array, which does not have to ever be stored, hence the time and memory complexities to estimate the separable part of the model, i.e. to evaluate (2.16), are $\mathcal{O}(NK^3)$ and $\mathcal{O}(K^2)$, respectively.

To evaluate $\widehat{\mathbf{B}} = \frac{1}{N}\sum_n \mathrm{Ta}(\mathbf{X}_n \otimes \mathbf{X}_n) - \mathrm{Ta}(\widehat{\mathbf{A}}_1 \tilde{\otimes} \widehat{\mathbf{A}}_2)$, one can utilize the fast Fourier transform (FFT). Every term $\mathrm{Ta}(\mathbf{X}_n \otimes \mathbf{X}_n)$ can be evaluated directly on the level of data, without the necessity to form the empirical estimator, in $\mathcal{O}(K^2\log(K))$ using the FFT. This is because even in the discrete case, there is relation between stationary operators and the Fourier transform. It is a well known fact in the time series literature that the periodogram is both the real part of the discrete Fourier transform (DFT) of the autocovariance function, i.e. of the first row of the (Toeplitz) covariance matrix, and the squared DFT of the data (Brockwell et al., 1991). This is a consequence of the Wiener-Khinchin theorem, and it allows one to compute the autocovariance function fast using the FFT. It is straightforward to show that the previous generalizes to the case of 2D data, which is done next for completeness.

Note that in the case of a 1D time series, the 2D covariance operator is captured by the 1D autocovariance. In the case of a 2D datum $\mathbf{X} \in \mathbb{R}^{K_1 \times K_2}$, the 4D covariance estimator $\widehat{\mathbf{C}}_N \in \mathbb{R}^{K_1 \times K_2 \times K_1 \times K_2}$ will be captured by the 2D *symbol* $\widehat{\mathbf{\Gamma}}_N \in \mathbb{R}^{K_1 \times K_2}$. The latter is defined as

$$\widehat{\mathbf{\Gamma}}_N[h_1, h_2] = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \mathbf{X}[k_1, k_2]\mathbf{X}^*[k_1 + h_1, k_2 + h_2].$$

The DFT of $\mathbf{X}$, denoted as $\mathbf{Z}$, is defined by

$$\mathbf{X}[k_1, k_2] = \frac{1}{\sqrt{K_1 K_2}} \sum_{a=1}^{K_1} \sum_{b=1}^{K_2} \mathbf{Z}[a, b] e^{-i\omega k_1 a} e^{-i\theta k_2 b},$$

where $\omega = 2\pi/K_1$ and $\theta = 2\pi/K_2$. Thus plugging the DFT of $\mathbf{X}$ into the formula for $\widehat{\mathbf{\Gamma}}_N$, we obtain

$$\begin{aligned}
\boldsymbol{\Gamma}[h_1,h_2] &= \frac{1}{(K_1K_2)^2}\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\sum_{a=1}^{K_1}\sum_{b=1}^{K_2}\sum_{t=1}^{K_1}\sum_{s=1}^{K_2}\mathbf{Z}[a,b]\mathbf{Z}^*[t,s]\cdot \\
&\qquad\qquad \cdot\, e^{-i\omega k_1 a}e^{-i\theta k_2 b}e^{i\omega(k_1+h_1)t}e^{i\theta(k_2+h_2)s} \\
&= \frac{1}{(K_1K_2)^2}\sum_{a=1}^{K_1}\sum_{b=1}^{K_2}\sum_{t=1}^{K_1}\sum_{s=1}^{K_2}\mathbf{Z}[a,b]\mathbf{Z}^*[t,s]\cdot \\
&\qquad\qquad \cdot\, e^{i\omega h_1 t}e^{i\theta h_2 s}\underbrace{\left[\sum_{k_1=1}^{K_1}e^{-i\omega k_1(a-t)}\right]}_{=K_1\mathbb{1}_{[a=t]}}\underbrace{\left[\sum_{k_2=1}^{K_2}e^{-i\theta k_2(b-s)}\right]}_{=K_2\mathbb{1}_{[b=s]}} \\
&= \frac{1}{K_1K_2}\sum_{a=1}^{K_1}\sum_{b=1}^{K_2}\mathbf{Z}[a,b]\mathbf{Z}^*[a,b]e^{i\omega h_1 a}e^{i\theta h_2 b} = \mathbf{W}[h_1,h_2],
\end{aligned}$$

where $\mathbf{W}$ is the inverse DFT applied to the DFT of $\mathbf{X}$ squared element-wise. Symbolically $\mathbf{W} = \text{ifft}\Big(|\text{fft}(\mathbf{X})|^2\Big)$, where $|\cdot|^2$ is applied element-wise. This shows that $\text{Ta}(\mathbf{X}\otimes\mathbf{X})$ can be calculated fast using the FFT.

Finally, the term $\text{Ta}(\widehat{\mathbf{A}}_1\,\tilde{\otimes}\,\widehat{\mathbf{A}}_2)$ can be evaluated directly in $\mathcal{O}(K^3)$ operations, again without explicitly forming the outer product. For example, $\text{Ta}(\mathbf{A}_1\,\tilde{\otimes}\,\mathbf{A}_2)[1,1,1,1]$ is the average of the diagonal elements of $\mathbf{A}_1\,\tilde{\otimes}\,\mathbf{A}_2$, which can be calculated as a product of the average diagonal element of $\mathbf{A}_2$ and average diagonal element of $\mathbf{A}_2$. Also $\text{Ta}(\mathbf{A}_1\,\tilde{\otimes}\,\mathbf{A}_2) \in \mathbb{R}^{K_1\times K_2\times K_1\times K_2}$ can be stored in the "autocovariance form" as an element of $\mathbb{R}^{K_1\times K_2}$. Altogether, the memory complexity and the number of operations needed for computing the estimator (2.17) in the case of $K_1 = K_2 = K$ is $\mathcal{O}(K^2)$ and $\mathcal{O}(NK^2\log K + K^3)$, respectively. Hence estimation of the banded part is equally demanding as the estimation of the separable part.

Altogether, we showed that a separable-plus-banded model can be estimated efficiently. It remains to show that $\widehat{\mathbf{C}} := \widehat{\mathbf{A}}_1\,\tilde{\otimes}\,\widehat{\mathbf{A}}_2 + \widehat{\mathbf{B}}$ can be applied efficiently, and that an inverse problem $\widehat{\mathbf{C}}\mathbf{X} = \mathbf{Y}$ can be solved efficiently. The application of $\widehat{\mathbf{C}}$ is easy to analyse due to the superposition structure: one simply applies the separable part using the first formula in (1.10), the banded part using the FFT, which is demonstrated in the following section, and sums the two, leading to the desired complexities. On the other hand, the inverse problem is non-trivial, 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 dealt with in Section 2.3.3.

2.3.2 Fast Application and Norm Calculation

To achieve fast application of the separable-plus-banded covariance, we can apply the separable and banded parts separately, and sum the results. The separable part can be applied fast using formula (1.10). For the banded part, we need to show that a matrix-vector product involving a Toeplitz matrix can be calculated efficiently. For this, we utilize *circulant matrices* (c.f. Davis, 2013). Recall that a matrix $\mathbf{Q} \in \mathbb{R}^{m \times n}$ is circulant if $\mathbf{Q} = (q_{ij}) = (q_{j-i+1 \mod n})$, where $\mathbf{q} \in \mathbb{R}^n$ is the *symbol* of the matrix, i.e. $\mathbf{q}^\top$ is the first row of $\mathbf{Q}$. Every circulant matrix is obviously a Toeplitz matrix. Contrarily, every Toeplitz matrix can be embedded into a larger circulant matrix (note that this embedding is not unique). For example, a symmetric Toeplitz matrix $\mathbf{T} \in \mathbb{R}^{n \times n}$ with symbol $\mathbf{t} \in \mathbb{R}^n$ can be embedded into a symmetric circulant matrix $\mathbf{Q} \in \mathbb{R}^{(2n-1)\times(2n-1)}$ with symbol $\mathbf{q} = (t_1, \ldots, t_n, t_n, \ldots, t_2)$. In the case of $n = 3$, we have

$$\mathbf{Q} = \left(\begin{array}{ccc|cc} t_1 & t_2 & t_3 & t_3 & t_2 \\ t_2 & t_1 & t_2 & t_3 & t_3 \\ t_3 & t_2 & t_1 & t_2 & t_3 \\ \hline t_3 & t_3 & t_2 & t_1 & t_2 \\ t_2 & t_3 & t_3 & t_2 & t_1 \end{array}\right) = \left(\begin{array}{c|c} \mathbf{T} & \cdot \\ \hline \cdot & \cdot \end{array}\right).$$

This embedding is useful due to the well known fact that circulant matrices are diagonalizable by the DFT, hence $\mathbf{Q} = \mathbf{E}^* \operatorname{diag}(\boldsymbol{\lambda})\mathbf{E}$, where $\mathbf{E}$ is matrix with the discrete Fourier basis in its columns, i.e. $\mathbf{E}[j,k] = \frac{1}{\sqrt{n}} e^{2\pi i jk/n}$. Hence the eigenvalues of $\mathbf{Q}$ can be calculated as the FFT of the symbol $\mathbf{q}$, namely $\boldsymbol{\lambda} = \operatorname{fft}(\mathbf{q})$. This implies that a matrix-vector product involving a circulant matrix can be calculated in $\mathcal{O}(n \log n)$ as

$$\mathbf{Q}\mathbf{v} = \mathbf{E}^* \operatorname{diag}(\boldsymbol{\lambda})\mathbf{E}\mathbf{v} = \operatorname{ifft}\Big(\boldsymbol{\lambda} \odot \mathbf{E}\mathbf{v}\Big) = \operatorname{ifft}\Big(\operatorname{fft}(\mathbf{q}) \odot \operatorname{fft}(\mathbf{v})\Big), \tag{2.28}$$

where $\operatorname{ifft}(\cdot)$ is the inverse FFT and $\odot$ denotes the Hadamard (element-wise) product. Thus using the circulant embedding, the product of a Toeplitz matrix $\mathbf{T} \in \mathbb{R}^{n \times n}$ with a vector $\mathbf{v} \in \mathbb{R}^n$ can also be calculated in $\mathcal{O}(n \log n)$:

$$\mathbf{Q}\left(\begin{array}{c} \mathbf{v} \\ \hline \mathbf{0} \end{array}\right) = \left(\begin{array}{c|c} \mathbf{T} & \cdot \\ \hline \cdot & \cdot \end{array}\right)\left(\begin{array}{c} \mathbf{v} \\ \hline \mathbf{0} \end{array}\right) = \left(\begin{array}{c} \mathbf{T}\mathbf{v} \\ \hline \cdot \end{array}\right). \tag{2.29}$$

The previous machinery can be naturally extended to higher dimensions, using two-level Toeplitz (resp. circulant) matrices, i.e. Toeplitz (resp. circulant) block matrices with Toeplitz (resp. circulant) blocks. For example, the tensor-matrix product $\widehat{\mathbf{B}}\mathbf{X}$ can be written as $\widehat{\mathbf{B}}_{\text{mat}}\operatorname{vec}(\mathbf{X})$, where $\widehat{\mathbf{B}}_{\text{mat}}$ is the matricization of $\widehat{\mathbf{B}}$, which is a two-level Toeplitz matrix. This product can be calculated by embedding $\widehat{\mathbf{B}}_{\text{mat}}$ into a two-level circulant matrix $\mathbf{Q}_{\text{mat}}$ and using analogs of (2.28) and (2.29). Notably, equation (2.28) becomes

$$\mathbf{Q}_{\text{mat}}\mathbf{X} = \text{i2Dfft}\Big(\text{2Dfft}(\boldsymbol{\Gamma}) \odot \text{2Dfft}(\mathbf{X})\Big),$$

where 2Dfft is the 2D DFT, i2Dfft is its inverse counterpart, and $\boldsymbol{\Gamma} \in \mathbb{R}^{(2K_1-1)\times(2K_2-1)}$ is the symbol of $\mathbf{Q}$, which is the tensorization of $\mathbf{Q}_{\text{mat}}$. Note that the $K_1 \times K_2$ top-left sub-matrix of $\boldsymbol{\Gamma}$ is the symbol of $\widehat{\mathbf{B}}$.

Altogether, we have shown that a stationary $\mathbf{B}$ or its estimate can be applied within our computational limits. Next, we show the same for norm calculations involving the separable-plus-banded model.

The bandwidth selection strategy discussed in Section 2.2.2 requires calculations of norms of separable-plus-stationary covariances. More generally, norms of the following form needs to be calculated:

$$|||\mathbf{A}_1 \tilde{\otimes} \mathbf{A}_2 + \mathbf{B} - \mathbf{C}_1 \tilde{\otimes} \mathbf{C}_2 - \mathbf{D}|||_2. \tag{2.30}$$

Assuming we work on a $K \times K$ grid, we have $\mathbf{A}_1, \mathbf{A}_2, \mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{K\times K}$ and $\mathbf{B}, \mathbf{D} \in \mathbb{R}^{K\times K\times K\times K}$ (being stationary) in the previous formula. A naive calculation of the norm then requires $\mathcal{O}(K^4)$ flops. In this section, we show that the special structure can be used to reduce the complexity to $\mathcal{O}(K^3)$.

One only needs to realize that both a separable covariance and a stationary covariance of size $K \times K \times K \times K$ can be re-arranged into a matrix of size $K^2 \times K^2$ with $K \times K$ blocks such that every block is a rank-one matrix. For example, the diagonal entries of $\mathbf{A}_1 \tilde{\otimes} \mathbf{A}_2 + \mathbf{B} - \mathbf{C}_1 \tilde{\otimes} \mathbf{C}_2 - \mathbf{D}$ are also entries of

$$\text{diag}(\mathbf{A}_1)\,\text{diag}(\mathbf{A}_2)^\top + \mathbf{B}[1,1,1,1]\mathbf{1}\mathbf{1}^\top - \text{diag}(\mathbf{C}_1)\,\text{diag}(\mathbf{C}_2)^\top - \mathbf{D}[1,1,1,1]\mathbf{1}\mathbf{1}^\top, \tag{2.31}$$

where $\mathbf{1} \in \mathbb{R}^K$ is the vector of ones. The matrix (2.31) is of size $K \times K$ and also rank-3. The squared Frobenius norm of this rank-3 matrix can be calculated using Gram-Schmidt orthogonalization in only $\mathcal{O}(K)$ flops. Summing together the total of K^2 of these blocks, we can calculate the square of (2.30) in $\mathcal{O}(K^3)$ flops. Therefore the norm calculation is within our computational constraints.

2.3.3 Inverse Problem

In this section, we develop a fast solver to a linear system coming from a discretization of model (2.1), i.e.

$$(\mathbf{A}_1 \tilde{\otimes} \mathbf{A}_2 + \mathbf{B})\mathbf{X} = \mathbf{Y}, \tag{2.32}$$

where $\mathbf{B} \in \mathbb{R}^{K\times K\times K\times K}$ is stationary.

Equation (2.32) can be rewritten in a matrix-vector form as

$$(\mathbf{A}+\mathbf{B})\mathbf{x}=\mathbf{y}, \tag{2.33}$$

where $\mathbf{A}=\mathbf{A}_2\otimes_K\mathbf{A}_1$ (cf. Remark 1), $\mathbf{x}=\mathrm{vec}(\mathbf{X})$, $\mathbf{y}=\mathrm{vec}(\mathbf{Y})$, and $\mathbf{B}\in\mathbb{R}^{K^2\times K^2}$ is a two-level Toeplitz matrix (i.e. a Toeplitz block matrix with Toeplitz blocks, see Chan and Jin, 2007).

The naive solution to system (2.33) would require $\mathcal{O}(K^6)$ operations, while if $\mathbf{B}\equiv 0$, i.e. if the system were separable, the solution could be found in $\mathcal{O}(K^3)$ operations. Since the estimation of model (2.1) takes $\mathcal{O}(NK^3)$, we are looking for a solver for (2.33) with a complexity close to $\mathcal{O}(K^3)$. We will develop an alternating direction implicit (ADI, cf. Young, 2014) solver with the per-iteration cost of $\mathcal{O}(K^3)$ and rapid convergence.

The system (2.33) can be transformed into either of the following two systems:

$$\begin{aligned}(\mathbf{A}+\rho\mathbf{I})\mathbf{x}&=\mathbf{y}-\mathbf{B}\mathbf{x}+\rho\mathbf{x},\\(\mathbf{B}+\rho\mathbf{I})\mathbf{x}&=\mathbf{y}-\mathbf{A}\mathbf{x}+\rho\mathbf{x},\end{aligned}\tag{2.34}$$

where $\mathbf{I}\in\mathbb{R}^{K^2\times K^2}$ is the identity matrix and $\rho\geq 0$ is arbitrary. The idea of the ADI method is to start from an initial solution $\mathbf{x}^{(0)}$, and form a sequence $\{\mathbf{x}^{(k)}\}_{k;2k\in\mathbb{N}}$ by alternately solving the linearized systems stemming from (2.34) until convergence, specifically:

$$\begin{aligned}(\mathbf{A}+\rho\mathbf{I})\mathbf{x}^{(k+1/2)}&=\mathbf{y}-\mathbf{B}\mathbf{x}^{(k)}+\rho\mathbf{x}^{(k)},\\(\mathbf{B}+\rho\mathbf{I})\mathbf{x}^{(k+1)}&=\mathbf{y}-\mathbf{A}\mathbf{x}^{(k+1/2)}+\rho\mathbf{x}^{(k+1/2)}.\end{aligned}\tag{2.35}$$

The acceleration parameter ρ (also called the shift) is allowed to vary between iterations. The optimal choice of ρ based on the spectral properties of $\mathbf{A}$ and $\mathbf{B}$, guaranteeing a fixed number of iterations, can be made in some model examples (e.g. when $\mathbf{A}$ and $\mathbf{B}$ commute). Interestingly, numerical studies suggest that the ADI method exhibits excellent performance on a large class of linear systems of the type (2.33) with the model choice of ρ, as long as matrices $\mathbf{A}$ and $\mathbf{B}$ are real with real spectra (Young, 2014). Hence we also choose ρ as suggested by the model examples and, in order to boost the convergence speed, we gradually decrease its value from the starting one

$$\rho^{(0)}=\sqrt{\max(\alpha_{\max}\alpha_{\min},\beta_{\max}\beta_{\min})}+\epsilon$$

as

$$\rho^{(k+1)}=\min\left(\rho^{(k)},\frac{\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|_2}{\|\mathbf{x}^{k}\|_2}\right),k\in\mathbb{N},$$

where α and β are the vectors of eigenvalues of $\mathbf{A}$ and $\mathbf{B}$, respectively, and ϵ is a small positive constant (by default the desired precision). Recall that $\mathbf{A}$ and $\mathbf{B}$ are positive semi-definite (we ensure this after estimation, as described in Section 2.3.4).

Now it remains to show how to efficiently solve the linear sub-problems (2.35).

Separable Equation $(\mathbf{A} + \rho\mathbf{I})\mathbf{x} = \mathbf{y}$

Even though $\mathbf{A} = \mathbf{A}_2 \otimes_K \mathbf{A}_1$, the matrix $\mathbf{A} + \rho\mathbf{I}$ does not generally posses the Kronecker structure. Nonetheless, the system can be rewritten in the matrix form as

$$\mathbf{A}_1\mathbf{X}\mathbf{A}_2 + \rho\mathbf{X} = \mathbf{Y}\,, \tag{2.36}$$

which is the well-known discrete Stein's equation. Even though there exist specialized solvers for this particular equation (see Simoncini, 2016, for an overview), they are not suitable here due to the fact that ρ is usually very small. Instead of using these specialized solvers, we show that, in our case of $\mathbf{A}_1$ and $\mathbf{A}_2$ being positive semi-definite, equation (2.36) has in fact an analytic solution computable in $\mathcal{O}(K^3)$ operations.

We compute the eigendecompositions $\mathbf{A}_1 = \mathbf{U}\,\mathrm{diag}(\boldsymbol{\phi})\mathbf{U}^\top$ and $\mathbf{A}_2 = \mathbf{V}\,\mathrm{diag}(\boldsymbol{\psi})\mathbf{V}^\top$. Then, using the knowledge of the spectra of Kronecker products (c.f. Lemma lem:ct and Remark 1), system (2.36) can be vectorized as

$$(\mathbf{U} \otimes_K \mathbf{V})\,\mathrm{diag}\left[\mathrm{vec}(\boldsymbol{\phi}\boldsymbol{\psi}^\top)\right](\mathbf{U} \otimes_K \mathbf{V})^\top\mathbf{x} + \rho\mathbf{x} = \mathbf{y}\,,$$

where $\boldsymbol{\phi}\boldsymbol{\psi}^\top$ is a matrix corresponding to the vector of eigenvalues of $\mathbf{A}$, which is subsequently rearranged into a large diagonal matrix by the diag$[\cdot]$ operator. Secondly, utilizing the fact that $\mathbf{U} \otimes_K \mathbf{V}$ is an orthonormal basis, we can write

$$(\mathbf{U} \otimes_K \mathbf{V})\,\mathrm{diag}\left[\mathrm{vec}(\mathbf{H})\right](\mathbf{U} \otimes_K \mathbf{V})^\top\mathbf{x} = \mathbf{y}\,,$$

where we denote $\mathbf{H} := \boldsymbol{\phi}\boldsymbol{\psi}^\top + \rho\mathbf{1}$, with $\mathbf{1}$ being a matrix with all entries equal to 1. Finally, one can express the solution as

$$\mathbf{x} = (\mathbf{U} \otimes_K \mathbf{V})\,\mathrm{diag}\left[\mathrm{vec}(\mathbf{H})\right]^{-1}(\mathbf{U} \otimes_K \mathbf{V})^\top\mathbf{y}\,.$$

Using property (1.9), this can be matricized back to

$$\mathbf{X} = \mathbf{V}(\mathbf{G} \odot \mathbf{U}^\top\mathbf{Y}\mathbf{V})\mathbf{U}^\top\,,$$

where $\mathbf{G}$ is the element-wise inverse of $\mathbf{H}$ and $\odot$ denotes the Hadamard (element-wise) product. Hence we found a solution, which is computable in $\mathcal{O}(K^3)$ operations.

Stationary Equation $(\mathbf{B} + \rho \mathbf{I})\mathbf{x} = \mathbf{y}$

$\mathbf{B}$ is a two-level Toeplitz matrix, and this structure is preserved when a diagonal matrix is added to $\mathbf{B}$, hence we only need to devise a solver for $\mathbf{Bx} = \mathbf{y}$, where $\mathbf{B}$ is positive definite. Even though specialized solvers for this structured linear system exist, provably providing a solution in $\mathcal{O}(K^2 \log^2(K))$, they are not easily accessible, and they are focused on cases when $\mathbf{B}$ is not symmetric. The latter is likely the case because a Preconditioned Conjugate Gradient (PCG) method is the method of choice, when positive definiteness is granted.

We do not describe the conjugate gradient (CG) method here, as it is a classical optimization method. Notably, Shewchuk (1994) provides both rigorous proofs and informal geometrical arguments for the fact that CG converges faster if the eigenvalues of $\mathbf{B}$ are clustered, which can be ensured by preconditioning. One CG step takes $\mathcal{O}(K^2 \log(K))$ operations, and this complexity is retained if a suitable preconditioning is used. Moreover, under mild assumptions and with a convenient preconditioner, the convergence rate of the PCG is super-linear, which means only a constant number of iterations is needed to attain a prescribed accuracy (Chan and Jin, 2007). Even though we cannot guarantee these mild assumptions, the second choice of preconditioning described in Chapter 5 of (Chan and Jin, 2007) was shown to ensure the fixed number of iterations for problems structurally very similar to ours. Hence we use this preconditioning, and provide empirical evidence in Section 2.5.1 that the resulting algorithm performs well.

Summary

In this section, 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 $\mathcal{O}(K^3)$ iterations, the other in $\mathcal{O}(\eta_{pcg} K^2 \log(K))$, where η_{pcg} is the number of the iterations of the inner scheme. In Section 2.5.1, we demonstrate empirically that η_{pcg} does not increase with increasing K, and hence the overall complexity of the algorithm is $\mathcal{O}(\eta_{adi} K^3)$, where η_{adi} is the number of outer iterations. As demonstrated again in Section 2.5.1, η_{adi} also does not depend on K, leading to an overall complexity $\mathcal{O}(K^3)$. Hence we have a tractable inversion algorithm for the separable-plus-stationary model.

Note that stationarity of $\mathbf{B}$ is used at two instances: in the top right-hand side of (2.35), $\mathbf{B}$ needs to be applied fast, and $(\mathbf{B} + \rho \mathbf{I})\mathbf{x} = \mathbf{y}$ needs to be solved fast. Both of these are easy if for example $\mathbf{B}$ is diagonal (instead of stationary), i.e. we also have an efficient inversion algorithm when a separable covariance is observed under heteroscedastic noise.

2.3.4 Ensuring Symmetry and Positive Semi-definiteness

Among other things, the assumption of separability induces extra symmetry. Every covariance C is symmetric in the sense that $c(t,s,t',s') = c(t',s',t,s)$ for any $t,s,t',s' \in [0,1]$. If $c(t,s,t',s') = c_1(t,t')c_2(s,s')$, it is easy to see that it must be

$$c(t,s,t',s') = c(t',s,t,s') = c(t,s',t',s) = c(t',s',t,s), \qquad t,s,t',s' \in [0,1].$$

When we wish to ensure that results of shifted partial tracing are symmetric, we have several options:

(a) symmetrizing the results of shifted partial tracing, for example setting

$$\widehat{A}_1 = \frac{1}{2}\Big[\mathrm{Tr}_1^{\delta}(\widehat{C}_N) + (\mathrm{Tr}_1^{\delta}(\widehat{C}_N))^*\Big],$$

(b) inducing the extra symmetry of the covariance, for example $\widehat{A}_1 = \mathrm{Tr}_1^{\delta}(\widetilde{C}_N)$ with $\widetilde{c}_N(t,s,t',s') = \frac{1}{2}[\widehat{c}_N(t,s,t',s') + \widehat{c}_N(t,s',t',s)]$,

(c) defining shifted partial tracing in a symmetric manner by replacing (2.4) with

$$k^{\delta}(t,s) = \begin{cases} \frac{1}{2}\Big[k(t,s+\delta) + k(t+\delta,s)\Big], & s < 1-\delta, \\ 0, & \text{otherwise,} \end{cases}$$

and developing shifted partial tracing from there, which would ultimately lead to the first formula in (2.3) replaced by

$$k_1(t,t') = \int_0^{1-\delta} \frac{1}{2}\Big[k(t,s,t',s+\delta) + k(t,s+\delta,t',s)\Big]\,\mathrm{d}s$$

These options are equivalent due to the symmetry of $\widehat{C}_N$ and the fact that adjoining commutes with any linear operator, hence also with shifted partial tracing.

Developing our theory as suggested by option (c) above is straightforward, merely lengthening all the calculations. In practice, option (a) is preferable for computational reasons.

Shifted partial tracing (even the symmetrized one) applied to a positive semi-definite (PSD) operator does not necessarily lead to a PSD operator. In the case of the original operator C being separable, it is easy to see that either $\mathrm{Tr}_1^{\delta} C \succeq 0$ or $-\mathrm{Tr}_1^{\delta} C \succeq 0$, so a potential sign flip is enough to ensure PSD. However, $\widehat{C}_N$ is usually not separable even when the original covariance C is. Nonetheless, $\widehat{C}_N$ is still a natural estimator of C and, from our experience, the potential sign flip usually solves the problem. If need be, the eigendecomposition can be calculated and negative eigenvalues set to zero. In the discrete case, this requires $\mathcal{O}(K^3)$ operations and thus it is computationally feasible.

Let us now focus on Toeplitz averaging. Since the argument in (2.17) is symmetric, and since the symmetry is obviously preserved, we only have to discuss positive semi-definiteness. Unfortunately, the argument $\widehat{C}_N - \widehat{A}_1 \tilde{\otimes} \widehat{A}_2$ is not necessarily PSD and thus $\widehat{B}$ may also not be. However, using Bochner's theorem the same way as in Hall and Patil (1994), the positive semi-definite projection of $\widehat{B}$ can be found. In the discrete case, the matricization of $\widehat{\mathbf{B}}$ can be embedded into a two-level circulant matrix with symbol $\mathbf{\Gamma}$ (as in Section 2.3.2). Subsequently, the DFT is applied to $\mathbf{\Gamma}$ to obtain the eigenvalues, negative eigenvalues are set to zero, and and the result is transformed back via the inverse DFT, giving the positive part of $\widehat{\mathbf{B}}$. This procedure requires $\mathcal{O}(K^2 \log K)$ operations when the FFT is used.

2.4 Asymptotic Properties

In this section, we establish asymptotic properties of the proposed estimators both in the case of fully observed and discretely observed data. We do this for a fixed value of the bandwidth parameter before generalizing our results to the case of adaptively chosen bandwidth, as described in Section 2.2.2. The proofs are postponed to the appendix, apart from the asymptotic distribution in the fully observed case, whose proof is straightforward and can be inspected without the general development of shifted partial tracing from Section 2.1.1.

2.4.1 Fully Observed Data

We begin by investigating the asymptotic distribution of our estimators in the fully observed case. There are three features we exploit to this end: linearity of shifted partial tracing, the central limit theorem in the space of trace-class operators (Theorem 1), and the continuous mapping theorem (CMT).

Theorem 3. *Let* $X_1, \ldots, X_N \sim X$ *be a (w.l.o.g. centered) random sample with covariance* (2.1)*, where* B *is stationary and* $\delta^\star$*-banded. Let* $\delta \geq \delta^\star$ *such that* $Tr^\delta(C) \neq 0$*. Let*

$$\sum_{j=1}^{\infty} \left(\mathbb{E}\langle X, e_j \rangle^4 \right)^{1/4} < \infty \tag{2.37}$$

for some orthonormal basis $\{e_j\}_{j=1}^{\infty}$ *in* $\mathcal{L}^2([0,1]^2)$*. Then*

$$\sqrt{N}(\widehat{A}_1 - A_1), \quad \sqrt{N}(\widehat{A}_2 - A_2), \quad \sqrt{N}(\widehat{B} - B)$$

converge to mean zero Gaussian random elements (of the proper trace-normed Banach spaces, as $N \to \infty$*).*

The moment assumption (2.37) ensures that $\sqrt{N}(\widehat{C}_N - C) \xrightarrow{d} Z$, where Z is a mean-zero Gaussian random element in $\mathcal{S}_1(\mathcal{L}^2([0,1]^2))$, i.e. the convergence is in the trace-norm topology (Mas, 2006). Also, it can be seen from the proof of the theorem that the the asymptotic distribution of $\widehat{A}_1$ and $\widehat{A}_2$ remains valid even without the stationarity assumption placed on $\widehat{B}$.

To prove Theorem 3, we need the following auxiliary result.

Lemma 5. *(a) Let $Z \in \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$ be a Gaussian random element. Then $Tr_1^\delta(Z)$ and $Tr_2^\delta(Z)$ are Gaussian random elements of $\mathcal{S}_1(\mathcal{L}^2([0,1]^2))$.*

(b) Let $Z \in \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$ be a Gaussian random element. Then $Ta(Z)$ is a Gaussian random elements of $\mathcal{S}_1(\mathcal{L}^2([0,1]^2))$.

(c) Let $Z \in \mathcal{S}_1(\mathcal{L}^2([0,1]))$ be a Gaussian random element and $F \in \mathcal{S}_1(\mathcal{L}^2([0,1]))$. Then $Z \tilde{\otimes} F$ and $F \tilde{\otimes} Z$ are Gaussian random elements in $\mathcal{S}_1(\mathcal{L}^2([0,1]^2))$.

Proof. Firstly, note that a random element $Z \in \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$ is Gaussian if, for any $G \in \mathcal{S}_\infty(\mathcal{L}^2([0,1]^2))$, $\mathrm{Tr}(GZ)$ is Gaussian (Bosq, 2012).

Secondly, for an operator $F : B_1 \to B_2$, its adjoint $F^* : B_2^* \to B_1^*$ is defined so for any $G \in B_2^*$ we have $F^*G = GF$.

(a) This follows immediately from the above and formula (2.5).

(b) For $\mathrm{Ta} : \mathcal{S}_1(\mathcal{L}^2([0,1]^2)) \to \mathcal{S}_1(\mathcal{L}^2([0,1]^2))$, the adjoint $\mathrm{Ta}^* : \mathcal{S}_\infty(\mathcal{L}^2([0,1]^2)) \to \mathcal{S}_\infty(\mathcal{L}^2([0,1]^2))$ satisfies $\mathrm{Ta}^*(G) = G\,\mathrm{Ta}$ for any $G \in S_\infty(\mathcal{L}^2([0,1]^2))$. Hence $\mathrm{Tr}(G\,\mathrm{Ta}(Z)) = \mathrm{Tr}(\mathrm{Ta}^*(G)Z)$, where $\mathrm{Ta}^*(G) \in S_\infty(\mathcal{L}^2([0,1]^2))$.

(c) This is Proposition 1.2 in the supplement of Aston et al. (2017). A proof can be found there. □

Proof of Theorem 3. We begin with the asymptotic Gaussianity of $\widehat{A}_1$. We have from linearity

$$\sqrt{N}(\widehat{A}_1 - A_1) = \sqrt{N}\Big(\mathrm{Tr}_1^\delta(\widehat{C}_N) - \mathrm{Tr}_1^\delta(C)\Big) = \mathrm{Tr}_1^\delta\Big(\sqrt{N}(\widehat{C}_N - C)\Big) \xrightarrow{d} \mathrm{Tr}_1^\delta(Z),$$

where the convergence follows from the CMT in the Banach space. $\mathrm{Tr}_1^\delta(Z)$ is Gaussian by Lemma 5. Its mean being zero follows from linearity of Tr_1^δ.

The asymptotic Gaussianity of $\widehat{A}_2$ follows in a similar way, but this time the CMT has to be applied using a non-linear function. We have

$$\begin{aligned}\sqrt{N}\left(\widehat{A}_2 - \frac{\mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(C)}\right) &= \sqrt{N}\left(\frac{\mathrm{Tr}_2^\delta(\widehat{C}_N)}{\mathrm{Tr}^\delta(\widehat{C}_N)} \pm \frac{\mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(\widehat{C}_N)} - \frac{\mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(C)}\right)\\ &= \frac{\sqrt{N}}{\mathrm{Tr}^\delta(\widehat{C}_N)}\left(\mathrm{Tr}_2^\delta\Big(\widehat{C}_N - C\Big) + \frac{\mathrm{Tr}^\delta(C)\mathrm{Tr}_2^\delta(C) - \mathrm{Tr}^\delta(\widehat{C}_N)\mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(C)}\right)\\ &= \frac{1}{\mathrm{Tr}^\delta(\widehat{C}_N)}\left(\mathrm{Tr}_2^\delta\Big[\sqrt{N}(\widehat{C}_N - C)\Big] - \frac{\mathrm{Tr}^\delta\Big[\sqrt{N}(\widehat{C}_N - C)\Big]\mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(C)}\right)\\ &\xrightarrow{d} \frac{1}{\mathrm{Tr}^\delta(C)}\left(\mathrm{Tr}_2^\delta(Z) - \frac{\mathrm{Tr}^\delta(Z)\mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(C)}\right),\end{aligned}$$

where we used again the CMT. Since $\mathrm{Tr}_2^\delta(Z)$ is Gaussian again by Lemma 5 and $\mathrm{Tr}^\delta(Z)$ is Gaussian from the definition, the whole limit is Gaussian. The mean is zero from linearity.

Finally, we turn our attention to $\widehat{B}$:

$$\begin{aligned}\sqrt{N}(\widehat{B} - B) &= \sqrt{N}\left(\mathrm{Ta}(\widehat{C}_N - \widehat{A}_1 \tilde{\otimes} \widehat{A}_2) - \mathrm{Ta}(C - A_1 \tilde{\otimes} A_2)\right)\\ &= \mathrm{Ta}\left(\sqrt{N}(\widehat{C}_N - C) - \sqrt{N}(\widehat{A}_1 \tilde{\otimes} \widehat{A}_2 - A_1 \tilde{\otimes} A_2)\right).\end{aligned} \tag{2.38}$$

Recall that in our model it holds $A_1 \tilde{\otimes} A_2 = \frac{\mathrm{Tr}_1^\delta(C) \tilde{\otimes} \mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(C)}$. Hence we have

$$\begin{aligned}\sqrt{N}(\widehat{A}_1 \tilde{\otimes} \widehat{A}_2 - A_1 \tilde{\otimes} A_2) = \sqrt{N}\Bigg(&\frac{\mathrm{Tr}_1^\delta(\widehat{C}_N) \tilde{\otimes} \mathrm{Tr}_2^\delta(\widehat{C}_N)}{\mathrm{Tr}^\delta(\widehat{C}_N)} - \frac{\mathrm{Tr}_1^\delta(C) \tilde{\otimes} \mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(C)}\\ &\pm \frac{\mathrm{Tr}_1^\delta(\widehat{C}_N) \tilde{\otimes} \mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(\widehat{C}_N)} \pm \frac{\mathrm{Tr}_1^\delta(C) \tilde{\otimes} \mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(\widehat{C}_N)}\Bigg)\\ = \frac{\sqrt{N}}{\mathrm{Tr}^\delta(\widehat{C}_N)}\Bigg(&\mathrm{Tr}_1^\delta(\widehat{C}_N) \tilde{\otimes} \mathrm{Tr}_2^\delta\Big[\widehat{C}_N - C\Big]\\ &+ \mathrm{Tr}_1^\delta\Big[\widehat{C}_N - C\Big] \tilde{\otimes} \mathrm{Tr}_2^\delta(C)\\ &- \mathrm{Tr}^\delta\Big[\widehat{C}_N - C\Big]\Big(A_1 \tilde{\otimes} A_2\Big)\Bigg).\end{aligned}$$

Plugging this back to (2.38) and using the CMT again, we obtain

$$\sqrt{N}(\widehat{B} - B) \xrightarrow{d} \mathrm{Ta}\left(Z - \frac{\mathrm{Tr}_1^\delta(C) \tilde{\otimes} \mathrm{Tr}_2^\delta(Z)}{\mathrm{Tr}^\delta(C)} - \frac{\mathrm{Tr}_1^\delta(Z) \tilde{\otimes} \mathrm{Tr}_2^\delta(C)}{\mathrm{Tr}^\delta(C)} + \frac{\mathrm{Tr}^\delta(Z)}{\mathrm{Tr}^\delta(C)}\Big(A_1 \tilde{\otimes} A_2\Big)\right).$$

The right-hand side before Toeplitz averaging is Gaussian again due to the reasons above. And by the previous lemma it remains Gaussian after Toeplitz averaging. □

2.4.2 Discretely Observed Data

Next, we consider the case of discretely measured random fields, potentially subject to additive measurement error contamination. Let $[0,1]^2 = \bigcup_{i=1}^K \bigcup_{j=1}^K I_{i,j}^K$, where $I_{i,j}^K$ is a Cartesian product of two sub-intervals of $[0,1]$ and $I_{i,j}^K \cap I_{i',j'}^K = \emptyset$ for $(i,j) \neq (i',j')$. Assume again that $K_1 = K_2 =: K$ and that $|I_{i,j}^K| = K^{-2}$ for all $i,j = 1,\ldots,K$.

The observations are assumed to be of the form

$$\widetilde{\mathbf{X}}_n^K[i,j] = \mathbf{X}_n^K[i,j] + \mathbf{E}_n^K[i,j]\,, \qquad i = 1,\ldots,K\,,\ s = 1,\ldots,K\,, \tag{2.39}$$

where the matrices $\mathbf{X}_1,\ldots,\mathbf{X}_N \in \mathbb{R}^{K\times K}$ are discretely measured versions of the latent surfaces $X_1,\ldots,X_N \in \mathcal{L}^2([0,1]^2)$, and $\mathbf{E}_n^K$ are measurement errors.

We will consider two types of sampling schemes relating the latent surfaces $X_1,\ldots,X_N \in \mathcal{L}^2([0,1]^2)$ to the discrete data $\mathbf{X}_1,\ldots,\mathbf{X}_N \in \mathbb{R}^{K\times K}$:

(S1) X_n, $n = 1,\ldots,N$, are observed pointwise on a grid, i.e. there exist $t_1^K,\ldots,t_K^K \in [0,1]$ and $s_1^K,\ldots,s_{K_2}^K \in [0,1]$ such that $(t_i^K, s_j^K) \in I_{i,j}^K$

$$\mathbf{X}_n^K[i,j] = X_n(t_i^K, s_j^K)\,, \qquad i = 1,\ldots,K\,,\ j = 1,\ldots,K\,.$$

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

(S2) The average value of X_n on the pixel $I_{i,j}^K$ is observed for every pixel, i.e.

$$\mathbf{X}_n^K[i,j] = \frac{1}{|I_{i,j}^K|}\int_{I_{i,j}^K} X_n(t,s)\,\mathrm{d}t\,\mathrm{d}s\,, \qquad i = 1,\ldots,K\,,\ j = 1,\ldots,K\,.$$

As for the measurement error arrays $\Big(\mathbf{E}_n^K[i,j]\Big)_{i,j=1}^K$, these are assumed to be i.i.d. (with respect to the index n) and uncorrelated with $\mathbf{X}_n$, satisfying the following 4-th order moment conditions:

$$\begin{aligned}
\mathbb{E}\Big(\mathbf{E}_n^K[i,j]\Big) &= 0\,,\\
\mathbb{E}\Big(\mathbf{E}_n^K[i,j]\mathbf{E}_n^K[k,l]\Big) &= \sigma^2 \mathbb{1}_{[i=k,j=l]}\,,\\
\mathbb{E}\Big(\mathbf{E}_n^K[i,j]\mathbf{E}_n^K[k,l]\mathbf{X}_n^K[i',j']\mathbf{E}_n^K[k',l']\Big) &= \mathbb{E}\Big(\mathbf{E}_n^K[i,j]\mathbf{E}_n^K[k,l]\Big)\,\mathbb{E}\Big(\mathbf{X}_n^K[i',j']\mathbf{E}_n^K[k',l']\Big)\,.
\end{aligned}$$

for $i,j,k,l,i',j',k',l' = 1,\ldots,K$ and $n = 1,\ldots,N$. Note that under the sampling scheme (S1), equation (2.39) corresponds to the commonly adopted errors-in-measurements model (Yao et al., 2005a; Zhang and Wang, 2016, and references therein).

We denote by X^K the piecewise constant continuation of $\mathbf{X}^K$, i.e.

$$X^K(t,s) = \sum_{i=1}^K \sum_{j=1}^K \mathbf{X}^K[i,j] \mathbb{1}_{[(t,s)\in I^K_{i,j}]} \,.$$

One can readily verify that pointwise sampling scheme (S1) corresponds to pointwise evaluations of the covariance, i.e. $\mathrm{Var}(X^K) = C^K$, where C^K has kernel

$$c^K(t,s,t',s') = \sum_{i,j,k,l=1}^K c(t_i,s_j,t_k,s_l) \mathbb{1}_{[(t,s)\in I^K_{i,j}]} \mathbb{1}_{[(t',s')\in I^K_{k,l}]} \,,$$

while pixel-wise sampling (scheme S2) corresponds in turn to pixelization of the covariance. Namely, if we denote $g^K_{i,j}(t,s) = K\mathbb{1}_{[(t,s)\in I^K_{i,j}]}$ then we have $\mathrm{Var}(X^K) = C^K$ with

$$X^K = \sum_{i=1}^K \sum_{j=1}^K \langle X, g^K_{i,j}\rangle g^K_{i,j} \,, \qquad C^K = \sum_{i,j,k,l=1}^K \langle C, g^K_{i,j} \otimes g^K_{k,l} \rangle g^K_{i,j} \otimes g^K_{k,l} \tag{2.40}$$

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

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 3, regardless of the sampling scheme.

Corollary 2. *Let* $\mathbf{X}_1,\ldots,\mathbf{X}_N$ *be i.i.d. copies of* $\mathbf{X} \in \mathbb{R}^{K\times K}$*, which has mean zero (w.l.o.g.) and covariance* $\mathbf{C} = \mathbf{A}_1 \tilde{\otimes} \mathbf{A}_2 + \mathbf{B}$ *with* $\mathbf{B}$ *banded by* $d^\star < K$*. Assume that* $\mathbb{E}\|X\|_F^4 < \infty$ *and that there exists* $d \geq d^\star$ *such that* $Tr^d(\mathbf{C}) \neq 0$*. Let* $\widehat{\mathbf{C}}_N = \frac{1}{N}\sum_{n=1}^N \mathbf{X}_n \otimes \mathbf{X}_n$*,* $\widehat{\mathbf{A}}_1 = Tr^d_1(\widehat{\mathbf{C}}_N)$*,* $\widehat{\mathbf{A}}_2 = Tr^d_2(\widehat{\mathbf{C}}_N)/Tr^d(\widehat{\mathbf{C}}_N)$ *and* $\widehat{B} = \widehat{\mathbf{C}}_N - \widehat{\mathbf{A}}_1 \tilde{\otimes} \widehat{\mathbf{A}}_2$*. Then*

$$\sqrt{N}(\widehat{\mathbf{A}}_1 - \mathbf{A}_1), \quad \sqrt{N}(\widehat{\mathbf{A}}_2 - \mathbf{A}_2), \quad \sqrt{N}(\widehat{\mathbf{B}} - \mathbf{B})$$

converge to mean zero Gaussian random elements.

When both N and K diverge, Theorem 3 does not apply, but we can still obtain convergence rates. To this aim, we first ought to clarify how bandedness of B, B^K and $\mathbf{B}^K$ are related. It can be seen that if B is banded by δ, then $\mathbf{B}^K$ is banded by $d_K = \lceil \delta K \rceil$, while B^K is banded by $\delta_K = d_K/K$, which decreases monotonically down to δ for $K \to \infty$. In the following theorem, $\widehat{A}^K_1$ and $\widehat{A}^K_2$ denote piecewise constant continuations of $\widehat{\mathbf{A}}^K_1 = \mathrm{Tr}^{d_K}_1(\widehat{\mathbf{C}}^K_N)$ and $\widehat{\mathbf{A}}^K_2 = \mathrm{Tr}^{d_K}_2(\widehat{\mathbf{C}}^K_N)/\mathrm{Tr}^{d_K}(\widehat{\mathbf{C}}^K_N)$, where $\widehat{\mathbf{C}}^K_N = \frac{1}{N}\sum_{n=1}^N \widetilde{\mathbf{X}}_n \otimes \widetilde{\mathbf{X}}_n$ is the empirical covariance based on the observed (noisy) data (2.39).

Theorem 4. *Let* $X_1,\ldots,X_N$ *be i.i.d. copies of* $X \in \mathcal{L}^2[0,1]^2$*, which has (w.l.o.g. mean zero and) covariance given by* (2.1)*, where the the separable part* $A := A_1 \tilde{\otimes} A_2$ *has kernel* $a(t,s,t',s')$*, which is Lipschitz continuous on* $[0,1]^4$ *with Lipshitz constant* $L > 0$*. Let*

$\mathbb{E}\|X\|^4 < \infty$ *and* $\delta \in [0,1)$ *be such that* B *from* (2.1) *is banded by* δ *and* $Tr^{\delta}(A) \neq 0$*. Let the samples come from* (2.39) *via measurement scheme (S1) or (S2) with* $\mathrm{Var}(\mathbf{E}_n^K[i,j]) \leq \sigma^2 = \mathcal{O}(\sqrt{K})$*. Then we have*

$$\left|\left|\left| \widehat{A}_1^K \tilde{\otimes} \widehat{A}_2^K - A_1 \tilde{\otimes} A_2 \right|\right|\right|_2^2 = \mathcal{O}_P(N^{-1}) + 2K^{-2}L^2, \tag{2.41}$$

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* $\widehat{A}_1^K = \sum_{j\in\mathbb{N}} \widehat{\lambda}_j^K \widehat{e}_j^K \otimes \widehat{e}_j^K$*,* $\widehat{A}_2^K = \sum_{j\in\mathbb{N}} \widehat{\rho}_j^K \widehat{f}_j^K \otimes \widehat{f}_j^K$*,* $A_2 = \sum_{j\in\mathbb{N}} \lambda_j e_j \otimes e_j$*, and* $A_2 = \sum_{j\in\mathbb{N}} \rho_j f_j \otimes f_j$ *are eigendecompositions, then* $|\widehat{\lambda}_i^K \widehat{\rho}_j^K - \lambda_i \rho_j|^2$ *follows the rate given in* (2.41)*, and if the eigensubspace associated with* e_j *is one-dimensional, then also* $\|\widehat{e}_j^K - \mathrm{sign}(\langle \widehat{e}_j^K, e_j\rangle)e_j\|_2^2$ *follows the rate given in* (2.41).

While the proof is postponed to the appendix, we make several comments here.

1. There is a concentration in K due to shifted partial tracing (recall Figure 2.1), hence the variance of the errors is allowed to grow with K as stated in the theorem.
2. The estimators $\widehat{A}_1^K$ and $\widehat{A}_2^K$ are only defined if $\mathrm{Tr}^{d_K}(\widehat{\mathbf{C}}_N^K) \neq 0$. Since $\widehat{\mathbf{C}}_N^K \to \mathbf{C}^K$ for $N \to \infty$ entry-wise apart from the diagonal, we have $\mathrm{Tr}^{d_K}(\widehat{\mathbf{C}}_N^K) \to \mathrm{Tr}^{d_K}(\mathbf{C}^K)$, so we require $\mathrm{Tr}^{d_K}(\mathbf{C}^K) \neq 0$. Due to continuity of the kernel c and the fact that $d_K \to \delta$ for $K \to \infty$, the assumption $\mathrm{Tr}^{\delta}(A) \neq 0$ implies $\mathrm{Tr}^{d_K}(\mathbf{C}^K) \neq 0$ 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.
3. The Lipschitz continuity assumption allows us to bound the bias while the fourth-order moment condition on data allows us to bound the variance. The bulk of the proof has to do with controlling the variance, and doing so uniformly in the grid size.
4. The Lipschitz continuity assumption can be weakened. For example, continuity almost everywhere is sufficient for the bias to convergence to zero, though without an explicit rate in K.
5. Since the roles of A_1 and A_2 are symmetric, one naturally obtains the rates for $\widehat{f}_j^K$ as well.

The rate in (2.41) is valid also in the uniform norm, but uniform rates for the eigenfunctions are a bit trickier to obtain. Standard perturbation bounds cannot be used anymore, and it does not seem possible to separate the effect of the grid size from the effect of the sample size, when dealing with uniform rates for the eigenfunctions. But if we assume e.g. that $K \asymp \sqrt{N}$, the corresponding rate holds for the eigenfunctions as well.

Similar rates of convergence hold also in the uniform norm, but the noise variance is allowed to grow with the grid size only at a slower rate.

Proposition 8. *Under the assumptions of Theorem 4, but with $\sigma^2 = \sqrt[4]{K}$, it holds*

$$\sup_{t,s,t',s'\in[0,1]} |\widehat{a}_1^K(t,t')\widehat{a}_2^K(s,s') - a_1(t,t')a_2(s,s')| = \mathcal{O}_P(N^{-1/2}) + 2K^{-1}L,$$

where the $\mathcal{O}_P(N^{-1})$ term is uniform in K, for all $K \geq K_0$ for a certain $K_0 \in \mathbb{N}$. This rate is also valid for the eigenvalues. And if, furthermore, $K \asymp \sqrt{N}$, we obtain the corresponding rate for the eigenfunctions as well:

$$\|\widehat{e}_j^K - \operatorname{sign}(\langle \widehat{e}_j^K, e_j\rangle)e_j\|_\infty = \mathcal{O}_P(N^{-1/2}) \quad \& \quad \|\widehat{f}_j^K - \operatorname{sign}(\langle \widehat{f}_j^K, f_j\rangle)f_j\|_\infty = \mathcal{O}_P(N^{-1/2}).$$

In case the banded part of the process is also of interest, the same rates can be achieved in the noiseless setting ($\sigma^2 = 0$) under the smoothness assumptions on the banded part of the covariance.

Without the assumption of stationarity on B, i.e. without Toeplitz averaging, one has:

$$\begin{aligned}
\left|\!\left|\!\left| \widehat{B}^K - B \right|\!\right|\!\right|_2 &\leq \left|\!\left|\!\left| \widehat{B}^K - B^K \right|\!\right|\!\right|_2 + \left|\!\left|\!\left| B^K - B \right|\!\right|\!\right|_2 \\
&\leq \left|\!\left|\!\left| \widehat{C}_N^K - \widehat{A}_1^K \tilde{\otimes} \widehat{A}_2^K - (C^K - A_1^K \tilde{\otimes} A_2^K) \right|\!\right|\!\right|_2 + \left|\!\left|\!\left| B^K - B \right|\!\right|\!\right|_2 \\
&\leq \left|\!\left|\!\left| \widehat{C}_N^K - C^K \right|\!\right|\!\right|_2 + \left|\!\left|\!\left| \widehat{A}_1^K \tilde{\otimes} \widehat{A}_2^K - A_1^K \tilde{\otimes} A_2^K \right|\!\right|\!\right|_2 + \left|\!\left|\!\left| B^K - B \right|\!\right|\!\right|_2
\end{aligned}$$

where the separable term can be treated as before and $\left|\!\left|\!\left| \widehat{C}_N^K - C^K \right|\!\right|\!\right|_2$ can be bounded similarly. With the assumption of stationarity, i.e. with Toeplitz averaging, 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 Yao et al. (2005a). We omit the details here. However, we note that full covariance smoothing is obviously not computationally tractable, hence any smoothing should either be applied at the level of data (pre-smoothing) or at the level of the estimated 2D parts of the covariance (post-smoothing). Nonetheless, as exemplified by the previous theorem, the mere presence of white noise errors does not call for smoothing when the target of inference is the separable component, or when an estimator of the covariance including noise is sought, e.g. for the purposes of prediction.

Remark 5. *In the noiseless case, the convergence rates in Theorem 4 are immediately applicable to the special case of a separable model and standard (non-shifted) partial tracing, as used by Aston et al. (2017). In the noisy case, however, shifted partial tracing (with an arbitrarily small shift) is needed to remove the noise. The denoising properties of shifted partial tracing are quite remarkable, as demonstrated by the fact that the noise level is allowed to grow with the sample size in Theorem 4. 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 (non-shifted) partial trace.*

2.4.3 Adaptive Bandwidth

In the previous sections, we have studied asymptotic properties of the estimators obtained by the proposed methodology with a known value of the bandwidth parameter (i.e. the shift used when partially tracing). Here, we provide versions of these results when the bandwidth parameter is unknown and selected adaptively via the procedure described in Section 2.2.2. We begin by providing rates of convergence in the fully observed scenario.

Theorem 5. *Let* $X_1, \ldots, X_N \sim X$ *be a (w.l.o.g. centered) random sample with covariance given by* (2.1)*, where* B *is stationary and* $\delta^\star$*-banded. Let the moment condition* (2.37) *hold for some orthonormal basis* $\{e_j\}_{j=1}^\infty$ *in* $\mathcal{L}^2[0,1]^2$*, and let* $\widehat{\delta}$ *be chosen as in* (2.22) *from the set* Δ *of finite size in which there exists* $\delta \geq \delta^\star$ *such that* $Tr^\delta(C) \neq 0$*. Then* $\left|\!\left|\!\left| \widehat{A}_1(\widehat{\delta}) \tilde{\otimes} \widehat{A}_2(\widehat{\delta}) - A_1 \tilde{\otimes} A_2 \right|\!\right|\!\right|_2^2 = \mathcal{O}_P(N^{-1})$ *and* $\left|\!\left|\!\left| \widehat{B}(\widehat{\delta}) - B \right|\!\right|\!\right|_2^2 = \mathcal{O}_P(N^{-1})$*.*

The adaptively chosen bandwidth itself is not consistent. This is because there is nothing to be consistent for: under the separable-plus-banded model, there is a whole range of valid bandwidths, which are asymptotically indistinguishable. Still, all those bandwidths lead asymptotically to the same estimator, and hence the previous theorem provides consistency of $\widehat{C}(\widehat{\delta})$ for $C(\delta^\star)$ even if $\widehat{\delta}$ itself is not consistent for $\delta^\star$. This reflects that δ is merely a nuisance parameter.

However, if asymptotic distribution is needed, selection among equally well-suited bandwidths is necessary. The latter can be achieved by a slight modification of the bandwidth selection scheme. For $\tau \geq 0$, we define

$$\begin{aligned} \Xi_\tau(\delta) &:= |||C(\delta) - C|||_2^2 + \tau\delta, \\ \widehat{\Xi}_\tau(\delta) &:= \left|\!\left|\!\left| \widehat{C}(\delta) \right|\!\right|\!\right|_2^2 - \frac{2}{N} \sum_{n=1}^{N} \langle X_n, \widehat{C}_{-n}(\delta) X_n \rangle + \tau\delta, \end{aligned}$$

and

$$\widetilde{\delta} := \arg\min_{\delta \in \Delta} \widehat{\Xi}_\tau(\delta). \tag{2.42}$$

A new parameter τ has been introduced into the objective to discriminate between equally good choices of δ. The proposition below shows that under the modified scheme, $\widehat{C}(\widetilde{\delta})$ is asymptotically Gaussian, when $\tau > 0$ is small enough.

Proposition 9. *Let* $X_1, \ldots, X_N \sim X$ *be a (w.l.o.g. centered) random sample with covariance given by* (2.1)*, where* B *is stationary and* $\delta^\star$*-banded. Let the moment condition* (2.37) *hold for some orthonormal basis* $\{e_j\}_{j=1}^\infty$ *in* $\mathcal{L}^2([0,1]^2)$*. Let* $\Delta = \{\delta_1, \ldots, \delta_m\}$ *be such that* $Tr^\delta(C) \neq 0$ *for any* $\delta \in \Delta$ *of which at least one is larger that* $\delta^\star$*. Finally, let* $\widetilde{\delta}$ *be chosen as in* (2.42) *with* $\tau < \min_{\delta \in \Delta, \delta < \delta^\star} |\Xi(\delta) - \Xi(\delta^\star)|$*. Then* $\sqrt{N}(\widehat{A}_1(\widetilde{\delta}) \tilde{\otimes} \widehat{A}_2(\widetilde{\delta}) - A_1 \tilde{\otimes} A_2)$ *and* $\sqrt{N}(\widetilde{B}(\widehat{\delta}) - B)$ *converge to mean zero Gaussian random elements.*

On one hand, τ is a new nuisance parameter that needs to be chosen instead of δ. On the other hand, it is easier to choose it (it just needs to be small enough). Moreover, the previous theorem shows that choosing $\tau = 0$ provides the correct rates of convergence.

Finally, we discuss what happens to the discrete rates in Theorem 4 when the bandwidth is chosen adaptively. While Theorem (4) establishes that the proposed estimation methodology is robust against noise, this is not the case for the bandwidth selection procedure of Section 2.2.2. Here, we will change the bandwidth selection procedure to one that is robust against noise. We should, however, note that this new bandwidth selection procedure should rarely be used in practice. This goes back to whether we see a banded part of the model as a nuisance or as a rough signal to be estimated. With a fixed grid size K, choosing the discrete bandwidth $d = 1$ suggests either presence of noise, or presence of a banded part B with bandwidth smaller than the reciprocal of the grid size $1/K$. There is no way to distinguish between the two options, and a practitioner would hardly care about the difference. Still, the distinction has to be made for the purpose of theory, when the grid size K is allowed to diverge. The development below can be taken simply as a complementary evidence that we truly obtain the correct estimators of the separable part, even under discrete noisy measurements and when the bandwidth is unknown.

For $K \in \mathbb{N}$, $\mathbf{F}, \mathbf{G} \in \mathbb{R}^{K\times K\times K\times K}$, and $F^K, G^K \in \mathcal{S}_2(L^2[0,1]^2)$ the piece-wise constant continuations of $\mathbf{F}$ and $\mathbf{G}$, respectively, we define $\|F^K\|_\star$ via

$$\left|\!\left|\!\left| F^K \right|\!\right|\!\right|_\star^2 = \left|\!\left|\!\left| F^K \right|\!\right|\!\right|_2^2 - \frac{1}{K^2}\| \operatorname{diag}(\mathbf{F})\|_2^2.$$

We also define $\langle \cdot, \cdot \rangle_\star$ as

$$\langle F^K, G^K \rangle_\star = \langle F^K, G^K \rangle - \frac{1}{K^2}\langle \operatorname{diag}(\mathbf{F}), \operatorname{diag}(\mathbf{G})\rangle.$$

Finally, recall that $\widetilde{X}_n^K$ are the discrete noisy samples (or rather piece-wise constant continuations thereof), and define

$$\Xi^K(\delta) := \left|\!\left|\!\left| C^K(\delta) - C^K \right|\!\right|\!\right|_\star^2 \quad \& \quad \widehat{\Xi}^K(\delta) := \left|\!\left|\!\left| \widehat{C}^K(\delta) \right|\!\right|\!\right|_\star^2 - \frac{2}{N}\sum_{n=1}^N \langle \widetilde{X}_n^K, \widehat{C}_{-n}^K(\delta)\widetilde{X}_n^K \rangle_\star,$$

and

$$\widehat{\delta} := \operatorname*{arg\,min}_{\delta\in\Delta} \widehat{\Xi}^K(\delta) \quad \& \quad \delta_\star := \operatorname*{arg\,min}_{\delta\in\Delta} \Xi^K(\delta). \tag{2.43}$$

With these definitions, we are trying to bypass the effect of noise on the bandwidth selection procedure, to obtain an adaptive version of Theorem 4. Since $\|\cdot\|_\star$ is clearly a semi-norm and $\langle \cdot, \cdot \rangle_\star$ is the corresponding semi-inner-product (Conway, 2019), we will be able to combine the continuous-domain result in Theorem 5 with the discrete-domain result in Theorem 4 to obtain the following result.

Theorem 6. *Let* $X_1, \dots, X_N$ *be i.i.d. copies of* $X \in \mathcal{L}^2([0,1]^2)$*, which has (w.l.o.g. mean zero and) covariance given by* (2.1)*, where the the separable part* $A := A_1 \tilde{\otimes} A_2$ *has kernel* $a(t,s,t',s')$*, which is Lipschitz continuous on* $[0,1]^4$ *with a Lipshitz constant* $L > 0$*. Let* $\mathbb{E}\|X\|^4 < \infty$ *and* $\delta^\star \in [0,1)$ *be such that* B *from* (2.1) *is banded by* $\delta^\star$*. Let* Δ *be such that* $Tr^\delta(A) \neq 0$ *for all* $\delta \in \Delta$ *of which at least one is larger than* $\delta^\star$*, and let* $\widehat{\delta}$ *be chosen from* Δ *as in* (2.43)*. Let the samples come from* (2.39) *via measurement scheme (S1) or (S2) with* $\operatorname{Var}(\mathbf{E}_n^K[i,j]) \leq \sigma^2 < \infty$*. Then we have*

$$\left|\left|\left| \widehat{A}_1^K(\widehat{\delta}) \tilde{\otimes} \widehat{A}_2^K(\widehat{\delta}) - A_1 \tilde{\otimes} A_2 \right|\right|\right|_2^2 = \mathcal{O}_P(N^{-1}) + 2K^{-2}L^2, \tag{2.44}$$

where the $\mathcal{O}_P(N^{-1})$ *term is uniform in* K*, for all* $K \geq K_0$ *for a certain* $K_0 \in \mathbb{N}$*.*

Rates for the eigenvalues and eigenfunctions hold as well and can be obtained just as in Theorem 4.

2.5 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 (Aston et al., 2017) or as the nearest Kronecker product (Van Loan and Pitsianis, 1993). We begin with simulated data in Section 2.5.1, where we focus on weakly dependent contamination of separability, and then move on to a real data set in Section 2.5.2, where we find evidence for heteroscedastic white noise contamination.

2.5.1 Simulation Study

The data are generated as follows. Firstly, we create norm-one covariances $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{K \times K}$ and draw $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ independently from the matrix-variate Gaussian distribution with mean zero and covariance $\mathbf{A} = \mathbf{A}_1 \tilde{\otimes} \mathbf{A}_2$, using Theorem 2. Secondly, we draw enough $\mathcal{N}(0,1)$ entries (independent of everything), arrange them on a grid, and perform space-time averaging using a filter $\mathbf{Q} = \left(q_{k,l}\right) \in \mathbb{R}^{d \times d}$ for $d \in \{1, 3, \dots, 19\}$ to obtain a sample $\mathbf{W}_n$ for every $n = 1, \dots, N$. This sample is drawn from a distribution with mean zero and covariance $\mathbf{B} \in \mathbb{R}^{K \times K \times K \times K}$, which is by construction stationary, banded by d, and its entries can be explicitly calculated. We set the sample size $N = 300$ and the grid size $K = 100$, so the discrete bandwidth d corresponds to the continuous bandwidth δ in percentages. Finally, we form our data set $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathbb{R}^{K \times K}$ as

$$\mathbf{X}_n = \mathbf{Y}_n + \sqrt{\tau}\mathbf{W}_n, \quad n = 1, \dots, N, \tag{2.45}$$

where $\tau \in [0,1]$. Thus $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathbb{R}^{K \times K}$ are drawn from a zero-mean distribution with a separable-plus-banded covariance $\mathbf{C} = \mathbf{A}_1 \tilde{\otimes} \mathbf{A}_2 + \tau\mathbf{B}$. Since $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{B}$ are standardized to have the Frobenius norm equal to one, the parameter τ can be understood as signal-to-noise ratio.

Note that our methodology based on shifted partial tracing first estimates the separable part of the model, and subsequently estimates $\mathbf{B}$ using the estimates for the separable part. Therefore the parameter τ of (2.45) controls the difficulty of the estimation problem in a continuous way: a small τ corresponds to a nearly separable model, which is easier to estimate than a highly non-separable model stemming from a larger τ. 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 $d = 0$ leads to an exactly separable model. The third difficulty-governing parameter is naturally the sample size N.

The following methods were used to estimate $\mathbf{C}$:

> SPT-d – shifted partial tracing, i.e. the proposed methodology of Section 2.2, provided with the true bandwidth d;
>
> SPT-CV – shifted partial tracing with δ chosen via cross-validation;
>
> PT – partial tracing, an approach incorrectly assuming separability;
>
> NKP – the nearest Kronecker product, i.e. the solution to (1.23);
>
> ECE – the standard empirical covariance estimator.

The plots also show the *bias* of a separable estimator, calculated as the difference between the true covariance $\mathbf{C}$ and the best separable approximation of $\mathbf{C}$, i.e. the solution to (1.23) with $\widehat{\mathbf{C}}_N$ replaced by $\mathbf{C}$. For several different settings, we calculate the relative estimation error $\|\mathbf{C} - \widehat{\mathbf{C}}\|_F / \|\mathbf{C}\|_F$, where $\widehat{\mathbf{C}}$ is an estimator computed by one of the above-listed methods.

The separable factors $\mathbf{A}_1$ and $\mathbf{A}_2$ are chosen both as either the covariance of the Wiener process or as rank-7 covariances with linearly decaying eigenvalues and shifted Legendre polynomials as the eigenvectors, see Figure 2.3. As will be explained later, the Wiener case is simpler to handle than the Legendre case, because the Wiener covariance decays slower away from the diagonal. For the banded part of the covariance, we choose the filter $\mathbf{Q} = \left(q_{k,l}\right)$ as either $q_{k,l} = (-1)^{|k-l|}$, leading to $\mathbf{B}$ with its symbol depicted in Figure 2.4 (left), or $q_{k,l} = \frac{9}{16}\left(1 - \frac{|k|}{p+1}\right)\left(1 - \frac{|l|}{p+1}\right)$, i.e. the outer product of two Epanechnikov kernels, leading to $\mathbf{B}$ with its symbol depicted in Figure 2.4 (right). Again, the Epanechnikov case will turn out to be easier compared to the other choice of the filter (called the *signed* case).

Figure 2.5 depicts how the estimation error evolves when one of the three difficulty-governing parameters (bandwidth d, signal-to-noise ration τ, and sample size N) varies,

Figure 2.3: The two choices for the separable constituents of the separable-plus-banded model: the Legendre covariance (**left**) and the Wiener covariance (**right**).

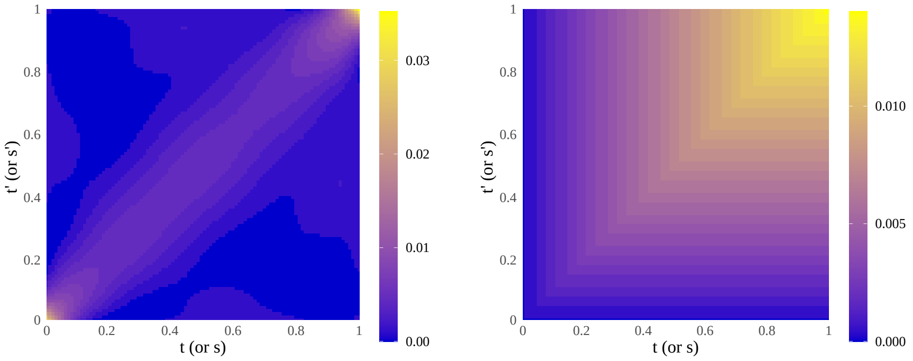
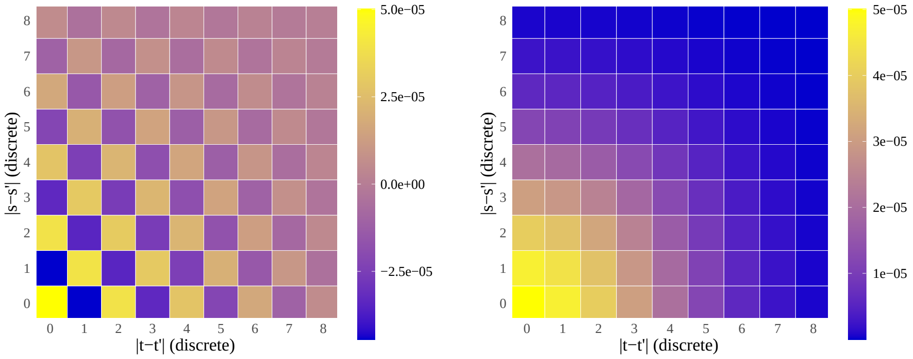


Figure 2.4: The two choices for symbol of the banded part of the separable-plus-banded model: the signed case (**left**) and the Epanechnikov case (**right**).



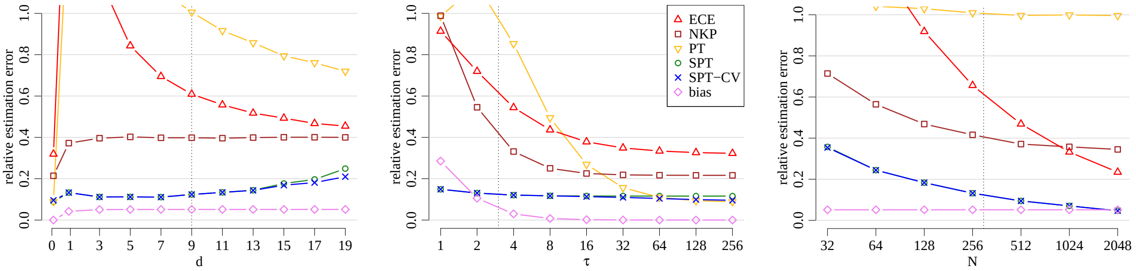
while the remaining two parameters are held fixed at any given plot (at $d = 9$, $\tau = 3$ or $N = 300$) in the Legendre-signed case, which is the most interesting one of the four cases. There are several remarks to be made about the results in Figures 2.5:

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 δ (see the tail end of the left and middle plots in Figure 2.5). This is because the banded part $\mathbf{B}$ decays away from the diagonal, and sometimes choosing a smaller bandwidth than the true one can lead to a better bias-variance trade-off.
3. When the truth is separable (i.e. $d = 0$) or nearly separable (i.e. τ large[1]), partial tracing leads to the best results. In these cases, the bandwidth selection strategy

---

[1] This is because $\mathbf{B}$ is itself separable by construction, so when τ is large – making $\mathbf{A}_1 \tilde{\otimes} \mathbf{A}_2$ negligible – the separable model can lead to a good performance.

Figure 2.5: Estimation errors for several competing methods with changing bandwidth d (**left**), signal to noise ratio τ (**middle**), and sample size N (**right**) in the Legendre-signed scenario. The vertical dotted lines show where every parameter is fixed for the remaining two plots (e.g. for the left plot, it is $\tau = 3$ and $N = 300$).



correctly chooses a very small bandwidth, and hence the performance of SPT-CV matches the one of PT.

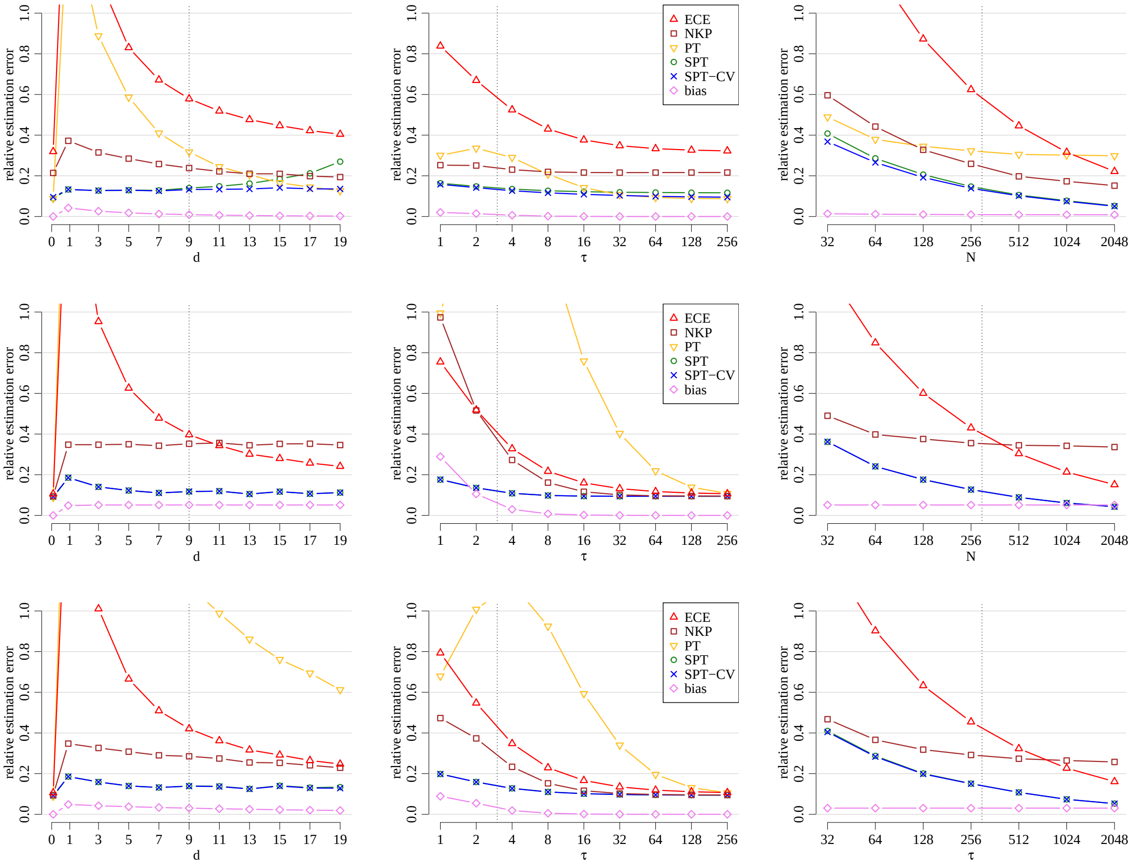
4. Note the extreme rise at the beginning of the error curves belonging to the empirical or the separable estimators in Figure 2.5 (left). While $d = 0$ corresponds to a separable model, $d = 1$ is already quite non-separable. Even though the amount of non-separability (c.f. the *bias* curve) is rather 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 2.5, right). While the amount of non-separability 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. Altogether, we can say that presence of noise strikingly obstructs separable estimation.

Results for the remaining three scenarios (Legendre-Epanechnikov, Wiener-signed, and Wiener-Epanechnikov cases) are shown in Figure 2.6. All the scenarios exhibit qualitatively similar behavior (described above), and the quantitative differences can be attributed to different shapes of the underlying covariances. Still, there are couple of observations worth noting:

1. Firstly, note that the Wiener case seems to be slightly easier than the Legendre case, despite the fact the former covariance is non-differentiable while the latter is analytic. Higher-order smoothness of the covariances does not play a role – this should be expected since our theoretical development does not make any such assumptions. On the other hand, difficulty of the problem is governed by how fast A_1 and A_2 decay away from the diagonal. This is reflected in the theory by the assumption of non-zero shifted traces $\mathrm{Tr}^d(A_1)$ and $\mathrm{Tr}^d(A_2)$. While the shifted traces being non-zero suffices for asymptotic purposes, the finite-sample performance of our methodology is poor if the shifted traces past the true bandwidth are very small.

Figure 2.6: Analogous to Figure 2.5, i.e. relative estimation errors with varying bandwidth d (**left** column), signal-to-noise ratio τ (**middle** column), or sample size N (**right** column). The **top** row corresponds to the Legendre-Epanechnikov scenario, the **middle** row depicts the Brownian-signed scenario, and the **bottom** row captures the Brownian-Epanechnikov case. The black dotted vertical lines show where the active parameter is fixed for the remaining two plots (i.e. $d = 9$, $\tau = 3$, and $N = 300$), i.e. all the plots are roughly the same at the black dotted vertical cuts.



This is quite natural: if **B** covers almost all the mass of the separable component, the latter cannot be estimated reliably. This happens more easily in the Legendre case, because the Legendre covariances are more concentrated around the diagonal compared to the Wiener covariance.

2. Secondly, the signed choice for the banded part seems to make the problem harder than the Epanechnikov choice. While this cannot be visualized very well, one can imagine that the shape of **B** mimics the shape of the separable part better with the Epanechnikov choice. One can also observe this by looking at the bias curves, notice that the bias is generally smaller in the Epanechnikov case compared to the signed case. Hence in the Epanechnikov case, the covariance can be approximated with the assumption of separability much better, making the problem easier.

3. Finally, choosing a smaller bandwidth than the true one can sometimes be beneficial. This happens mainly when a relatively large true bandwidth leads to only a mild amount of non-separability, as happens in the right part of the top-left plot in Figure 2.6. Focusing specifically at $d = 15$ in the top-left plot in Figure 2.6, we see an instance where performance of the proposed methodology with adaptively chosen bandwidth outperforms both the separable model and the separable-plus-banded model with the true (oracle) choice of the bandwidth. This is because, as suggested by Figure 2.4 (right), the effective bandwidth is in this case smaller than the true bandwidth (because the symbol of $\mathbf{B}$ decays fast away from the diagonal), while not completely ignoring the banded part is still beneficial.

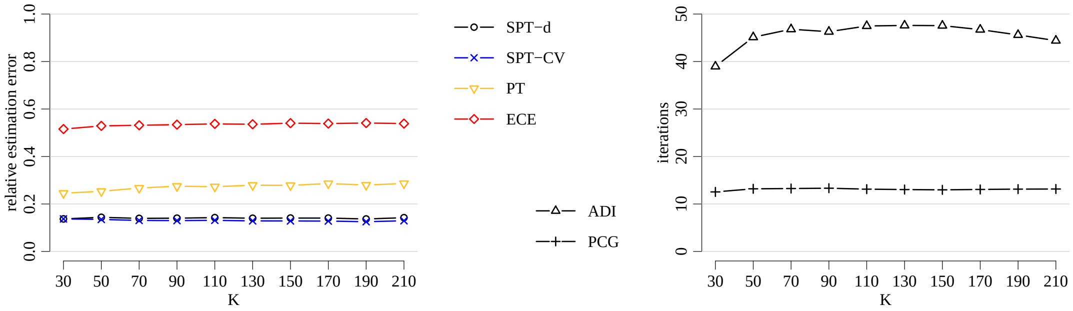
In the remainder of this section, we examine the functional nature of our problem, behavior of the ADI algorithm of Section 2.3.3 designed to solve inverse problems involving the covariance, and the number of iterations needed by the algorithm to converge. We simulate data as described before in the Legendre-Epanechnikov case, but now we vary the grid size $K \in \{10(2j+1); j = 1, \ldots, 10\}$, fix δ at 10 % (i.e. $d = K/10$), and we keep $\tau = 3$ and $N = 300$ for all the grid sizes.

Let $\widehat{\mathbf{C}} = \widehat{\mathbf{A}}_1 \tilde{\otimes} \widehat{\mathbf{A}}_2 + \widehat{\mathbf{B}}$ denote the estimator obtained by shifted partial tracing. $\widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2$ and $\widehat{\mathbf{B}}$ are subsequently projected onto positive semi-definite matrices (as described in Section 2.3.4). Also, a ridge regularization of size 10^{-5} is added to $\widehat{\mathbf{C}}$. Note that this is not necessary, because $\widehat{\mathbf{B}}$ is positive definite, and thus the problem is well defined even without any ridge regularization. However, the performance of ADI method heavily depends on the condition number of the system matrix, and 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 $\mathbf{Y} = \widehat{\mathbf{C}}\mathbf{X}$. Subsequently, the ADI algorithm is called on the inverse problem $\widehat{\mathbf{C}}\mathbf{X} = \mathbf{Y}$ with $\widehat{\mathbf{C}}$ 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 desired precision. Instead, we report estimation errors and number of iterations needed by the ADI scheme in Figure 2.7.

As suggested by our theoretical results, the relative estimation error does not depend on the grid size (see Figure 2.7, 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 2.7, right). This suggests super-linear convergence of the algorithm.

Figure 2.7: Left: Estimation errors for several competing methods relative to Oracle (i.e. higher curve corresponds to better performance) 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 2.3.3.



2.5.2 Data Analysis: Mortality Rates

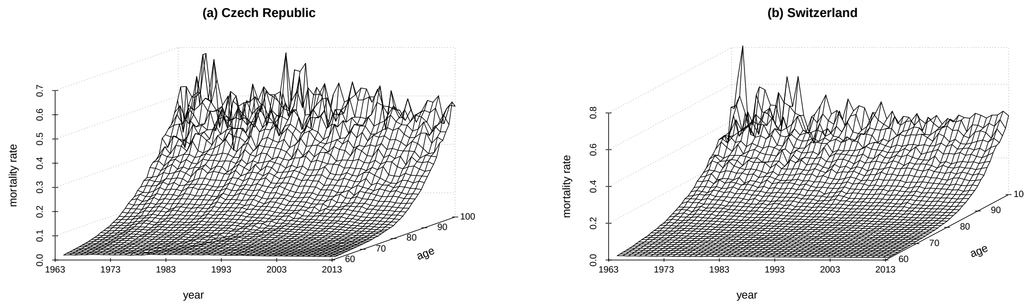
In this section, we analyse a data set $\mathbf{X} \in \mathbb{R}^{N \times K_1 \times K_2}$, where $\mathbf{X}[n, k_1, k_2]$ denotes the mortality rate for the n-th country, on the k_1-th calendar year and for subjects of age k_2. We consider the same set of 32 countries as Chen and Müller (2012) and Chen et al. (2017a), 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 $\mathbf{X} \in \mathbb{R}^{32 \times 50 \times 40}$. 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 2.8 shows a visualization of the raw data for two sample countries. The underlying continuous surfaces for different countries are assumed to be i.i.d. functional observations. 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 conducted by Chen and Müller (2012). Mortality surfaces were also considered by Chen et al. (2017a), who – presumably motivated by the work of Aston et al. (2017) and aiming for computational efficiency – calculated the “marginal kernels” $\mathrm{Tr}_1(\widehat{\mathbf{C}}_N)$ and $\mathrm{Tr}_2(\widehat{\mathbf{C}}_N)$, found the leading eigenfunctions of these marginal kernels, say $\{\widehat{\boldsymbol{\phi}}_i\}_{i=1}^I$ and $\{\widehat{\boldsymbol{\psi}}_j\}_{j=1}^J$, and used the tensor product approximation

$$\widehat{\mathbf{C}}_N \approx \sum_{i=1}^{I} \sum_{j=1}^{J} \widehat{\gamma}_{ij} (\widehat{\boldsymbol{\phi}}_i \otimes \widehat{\boldsymbol{\psi}}_j) \otimes (\widehat{\boldsymbol{\phi}}_i \otimes \widehat{\boldsymbol{\psi}}_j)\,, \tag{2.46}$$

where $\widehat{\gamma}_{ij} = \langle \widehat{\mathbf{C}}_N, (\widehat{\boldsymbol{\phi}}_i \otimes \widehat{\boldsymbol{\psi}}_j) \otimes (\widehat{\boldsymbol{\phi}}_i \otimes \widehat{\boldsymbol{\psi}}_j) \rangle$, which we note can be calculated fast. 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 C is not separable (Lynch and Chen, 2018).

Figure 2.8: Raw mortality rate surfaces for the Czech Republic and Switzerland.


(a) Czech Republic

(b) Switzerland

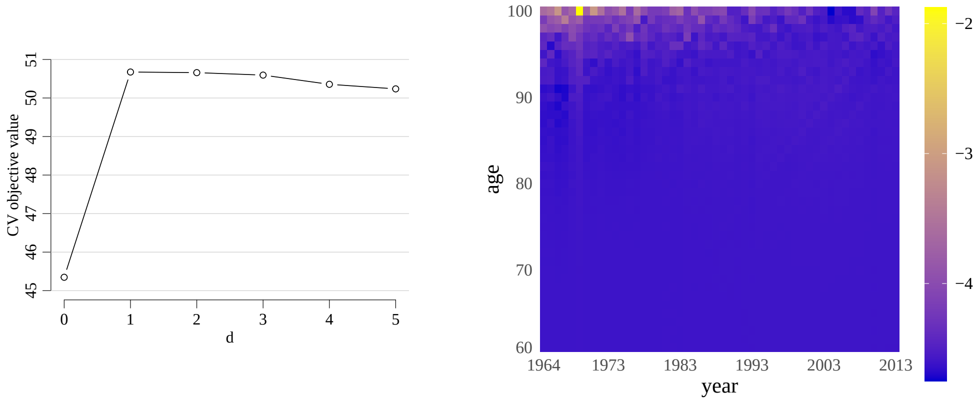
Compared to Chen and Müller (2012) or Chen et al. (2017a), we consider the mortality data with a slightly larger span of calendar years (the maximal span in which no data are missing). Our aim here is not to provide a novel analysis of the mortality dataset, but merely to illustrate the usefulness of shifted partial tracing.

Firstly, when investigating the sample curves in Figure 2.8, 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 decreases fast with increasing age. To verify whether the (most likely heteroscedastic) noise is white, we utilize the CV procedure. We do not assume stationarity, but we set the the estimator of the banded part (2.15) to zero outside of the current bandwidth in every step. The CV objective (2.22) is maximized at $\widehat{d} = 1$. We plot the objective curve in Figure 2.9, providing a strong evidence for the presence of noise. Since $\widehat{d} = 1$, we are in the separable-plus-noise regime, which is computationally feasible. Figure 2.9 also shows a heatmap of the estimated variance (or rather its logarithm, for visualisation purposes) of the noise depending on the location. The heatmap is in alignment with the conjecture that the noise variance is increasing with age.

Secondly, we compare the spectra of the marginal kernels $\mathrm{Tr}_1(\widehat{\mathbf{C}}_N)$ and $\mathrm{Tr}_2(\widehat{\mathbf{C}}_N)$ to their shifted counterparts $\mathrm{Tr}_1^1(\widehat{\mathbf{C}}_N)$ and $\mathrm{Tr}_2^1(\widehat{\mathbf{C}}_N)$. When partial tracing is used to obtain the marginal kernels, one has to keep 16 and 4 eigenfunctions, respectively, to capture 90 % of the variance in both dimensions. When shifted partial tracing is used instead, one instead needs to retain only 4 and 2 eigenfunctions, respectively, to capture 90 % of the variance in both dimensions. Hence shifted partial tracing offers a more parsimonious representation.

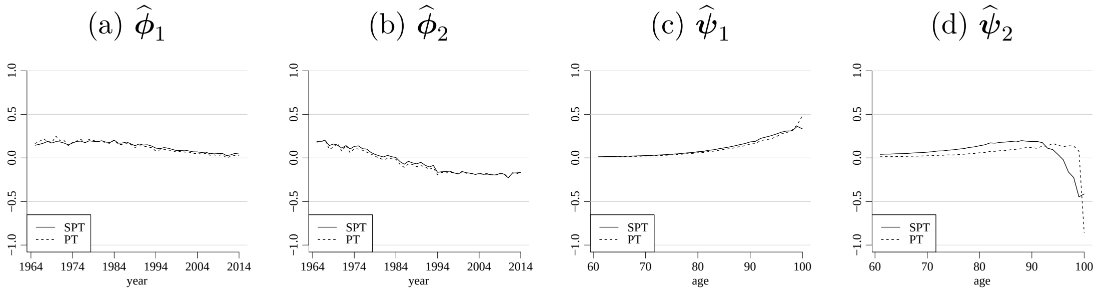
Thirdly, the empirical bootstrap test of Aston et al. (2017) with 4 and 2 marginal eigenfunctions (which seems to be the most reasonable choice, also used by Lynch and Chen, 2018) leads to a borderline p-value of 0.06. In comparison, the p-value for testing the separable-plus-banded model via the test described in Section 2.2.3 is over 0.4, suggesting that the separable-plus-banded model cannot be rejected for this data set.

Figure 2.9: Cross-validation objective for the mortality data (**left**) and log-heatmap of the heteroscedastic white noise variance (**right**).



Finally, keeping only 2 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 2.10. The first eigenfunctions in both dimensions capture the overall trend: $\widehat{\boldsymbol{\phi}}_1$ captures the decreasing variance in calendar years (the first dimension) and $\widehat{\boldsymbol{\psi}}_1$ the increasing variance in age (the second dimension). The second eigenfunction in the first dimension $\widehat{\boldsymbol{\phi}}_2$ 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 is too subtle to be visible by eye in the raw data plotted in Figure 2.8. Finally, the second eigenfunction in the second dimension $\widehat{\boldsymbol{\psi}}_2$ contrasts the old age (around 85) and oldest-old age (post 90) mortalities. However, this is only the case if shifted partial tracing is used. The eigenfunction $\widehat{\boldsymbol{\psi}}_2$ obtained from $\mathrm{Tr}_2(\widehat{\mathbf{C}}_N)$ 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 by Chen and Müller (2012) based on a different, computationally much more demanding methodology (see Park and Staicu, 2015, for a discussion).

Figure 2.10: First two eigenfunctions of the marginal kernels obtained by partial tracing (PT) and shifted partial tracing (SPT).


(a) $\widehat{\boldsymbol{\phi}}_1$ (b) $\widehat{\boldsymbol{\phi}}_2$ (c) $\widehat{\boldsymbol{\psi}}_1$ (d) $\widehat{\boldsymbol{\psi}}_2$

3 Separable Component Decomposition

In this chapter, we introduce and study a Hilbert-Schmidt operator decomposition allowing for a parsimonious representation, efficient estimation, and tractable manipulation of a random surface's covariance. Truncating this decomposition can be viewed as a natural generalization of separability. Since the decomposition itself applies to any covariance, the level of generality offered by the truncated decomposition is vast.

Let us view a covariance C as a Hilbert-Schmidt operator, i.e. we work in the Hilbert-Schmidt topology instead of the trace-class topology of the previous chapter. But more importantly, we utilize some of the isomorphisms between Hilbert-Schmidt operator spaces and product Hilbert Spaces discussed in Section 1.2. The following four spaces are isometrically isomorphic, and thus the covariance C of a random element $X \in \mathcal{H}_1 \otimes \mathcal{H}_2$ can be regarded as an element of any of these, each of which leads to different perspectives on potential decompositions:

$$\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2 \simeq \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2) \simeq \mathcal{S}_2(\mathcal{H}_2 \otimes \mathcal{H}_2, \mathcal{H}_1 \otimes \mathcal{H}_1) \simeq \mathcal{S}_2(\mathcal{H}_1) \otimes \mathcal{S}_2(\mathcal{H}_2).$$

If we consider $C \in \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$, we can write its eigendecomposition as

$$C = \sum_{j=1}^{\infty} \lambda_j g_j \otimes g_j, \tag{3.1}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots$ are the eigenvalues and $\{g_j\} \subset \mathcal{H}_1 \otimes \mathcal{H}_2$ are the eigenvectors, forming an ONB of $\mathcal{H}_1 \otimes \mathcal{H}_2$. On the other hand, if we consider $\mathcal{C} \in \mathcal{S}_2(\mathcal{H}_2 \otimes \mathcal{H}_2, \mathcal{H}_1 \otimes \mathcal{H}_1)$, we can write its singular value decomposition (SVD) as

$$\mathcal{C} = \sum_{j=1}^{\infty} \sigma_j e_j \otimes_2 f_j, \tag{3.2}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$ are the singular values, and $\{e_j\} \subset \mathcal{H}_1 \otimes \mathcal{H}_1$ and $\{f_j\} \subset \mathcal{H}_2 \otimes \mathcal{H}_2$ are the (left and right) singular vectors, forming ONBs of $\mathcal{H}_1 \otimes \mathcal{H}_1$ and $\mathcal{H}_2 \otimes \mathcal{H}_2$, respectively. Note that $\mathcal{C}$ is not self-adjoint in this case; $\{e_j\}$ are the eigenvectors of $\mathcal{C}\mathcal{C}^\top$, $\{f_j\}$ are the eigenvectors of $\mathcal{C}^\top\mathcal{C}$, and $\{\sigma_j^2\}$ are eigenvalues of both $\mathcal{C}\mathcal{C}^\top$ and $\mathcal{C}^\top\mathcal{C}$. We deliberately write $\mathcal{C}$ instead of C whenever the covariance is understood as something else than a self-adjoint element of $\mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$.

If we consider $\mathcal{C} \in \mathcal{S}_2(\mathcal{H}_1) \otimes \mathcal{S}_2(\mathcal{H}_2)$, the decomposition (3.2) can be re-expressed as

$$\mathcal{C} = \sum_{r=1}^{\infty} \sigma_r A_r \otimes B_r, \tag{3.3}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$ are the same as before, and $\{A_j\} \subset \mathcal{S}_2(\mathcal{H}_1)$ and $\{B_j\} \subset \mathcal{S}_2(\mathcal{H}_2)$ are isomorphic to $\{e_j\}$ and $\{f_j\}$, respectively. Finally, the expression (3.3) can be written down for (the self-adjoint version of) $C \in \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$ using the symbol defined in (1.7) as

$$C = \sum_{r=1}^{\infty} \sigma_r A_r \tilde{\otimes} B_r. \tag{3.4}$$

We will refer to (3.4) as the *separable component decomposition* (SCD) of C. On the level of kernels, the SCD corresponds to the following decomposition:

$$c(t, s, t', s') = \sum_{r=1}^{\infty} \sigma_r a_r(t, t') b_r(s, s').$$

We will also refer to the σ_r's as the *separable component scores.* One can verify using (1.8) that the eigendecomposition (3.1) and the SCD (3.4) are two different decompositions of the same element $C \in \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$. If all but $R \in \mathbb{N}$ separable component scores in (3.4) are zero, we say that the *degree-of-separability* (DoS) of C is R and write $\mathrm{DoS}(C) = R$. In this case, we also say in short that C is R-separable. If C does not necessarily have a finite degree-of-separability, but we truncate the series at level R yielding $C_R := \sum_{j=1}^{R} \sigma_j A_j \tilde{\otimes} B_j$ for some $R \in \mathbb{N}$, we call C_R the best R-separable approximation of C, because

$$C_R = \arg\min_{G} |||C - G|||_2^2 \quad \text{s.t.} \quad \mathrm{DoS}(G) \leq R. \tag{3.5}$$

It may be tempting to find an analogy between the degree-of-separability and the rank of an operator, which is defined as the number of non-zero eigenvalues in (3.1). But, as should be clear from Lemma 1, there is no simple relationship between these two. In particular, C can be of infinite rank even if it is 1-separable. However, C has degree-of-separability $R = 1$ if and only if C is separable according to Definition 1. In that case, we simply call C separable instead of 1-separable.

Remark 6. *The separable component decomposition generalizes separability on an arbitrary domain in a conceptually similar way to how the nearest Kronecker product*

(Van Loan and Pitsianis, 1993) generalizes to the sum of Kronecker products *(e.g., the PhD thesis of N. Pitsianis, 1997, Cornell University) or the* Kronecker product singular value decomposition *(Van Loan, 2000, Section 6) or the* Kronecker sum decomposition *(Tsiligkaridis and Hero, 2013) on the domain of matrices. When working with general Hilbert spaces, however, the Kronecker product and the associated matricizations and vectorizations of multi-dimensional objects (such as the covariance) do not apply and must be circumvented. Moreover, the general framework offers more simplicity and versatility, as will be seen later.*

3.1 Power Iteration Method

With the aim of constructing the separable component decomposition at the level of generality presented in the previous section, we first review calculation of eigenvalues and eigenvectors of matrices. Generally, the eigendecomposition of a symmetric matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$ can only by approximated numerically, and one of the basic methods for calculation of the leading eigenvector is the power iteration method described by the following recurrence relation:

$$\mathbf{v}^{(k+1)} = \frac{\mathbf{M}\mathbf{v}^{(k)}}{\|\mathbf{M}\mathbf{v}^{(k)}\|_2}, \quad k = 0, 1, \ldots,$$

where $\mathbf{v}^{(0)}$ is an initial guess. Provided $\mathbf{M}$ is diagonalizable, the leading eigenvalue is unique, and $\mathbf{v}^{(0)}$ is not orthogonal to the leading eigenvector, the sequence $\{\mathbf{v}^{(k)}\}_{k=1}^{\infty}$ converges to the leading eigenvector linearly with rate given by the spacing between the eigenvalues (cf. Van Loan and Golub, 1983). Once the leading eigenvector $\mathbf{v}_1$ is found, the leading eigenvalue is obtained as $\lambda_1 := \mathbf{v}_1^\top \mathbf{M}\mathbf{v}_1$, and the subsequent eigenvector is found via power iteration applied to the deflated matrix $\mathbf{M} - \lambda_1 \mathbf{v}_1 \mathbf{v}_1^\top$. The procedure resumes similarly until the desired number of leading eigenvalue-eigenvector pairs are found.

Assume now that $\mathbf{M} \in \mathbb{R}^{m \times n}$ and we are interested in finding its singular value decomposition (SVD). Since the right singular vectors of $\mathbf{M}$ are eigenvectors of $\mathbf{M}^\top\mathbf{M}$, they can be found via the power iteration method applied to $\mathbf{M}^\top\mathbf{M}$. Similarly the left singular vectors can be found by decomposing $\mathbf{M}\mathbf{M}^\top$. In practice, neither of the matrix products is formed, and the power iteration is carried out instead by alternating between the two following sequences for $k = 1, 2, \ldots$:

$$\mathbf{u}^{(k+1)} = \frac{\mathbf{M}\mathbf{v}^{(k)}}{\|\mathbf{M}\mathbf{v}^{(k)}\|_2}, \quad \mathbf{v}^{(k+1)} = \frac{\mathbf{M}^\top\mathbf{u}^{(k+1)}}{\|\mathbf{M}^\top\mathbf{u}^{(k+1)}\|_2},$$

which is equivalent to alternation between the two power iteration schemes on $\mathbf{M}^\top\mathbf{M}$ and $\mathbf{M}\mathbf{M}^\top$, respectively.

One can also view the power iteration method as an alternating least squares (ALS) algorithm. Due to the Eckart-Young-Minsky theorem, the leading principal subspace of $\mathbf{M}$ is the solution to the following least squares problem:

$$\operatorname*{arg\,min}_{\mathbf{u}\in\mathbb{R}^m,\mathbf{v}\in\mathbb{R}^n} \|\mathbf{M}-\mathbf{u}\mathbf{v}^\top\|_F^2.$$

The ALS algorithm fixes $\mathbf{v}$ and solves for $\mathbf{u}$, which is immediately chosen as $\mathbf{M}\mathbf{v}$, and then fixes $\mathbf{u}$ and sets $\mathbf{v}=\mathbf{M}^\top\mathbf{u}$. The two steps are iterated until convergence. It is clear that this corresponds to the power iteration method, once standardization is incorporated.

The reason not to explicitly form the matrix products $\mathbf{M}^\top\mathbf{M}$ and $\mathbf{M}\mathbf{M}^\top$ is the following. If $m \gg n$ (or vice versa), one of the matrix products will be much larger than $\mathbf{M}$ itself. For the same reason (Jolliffe, 1986), the eigenvectors of the sample covariance matrix $\widehat{\mathbf{C}}_N = \frac{1}{N}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$, where $\mathbf{X}\in\mathbb{R}^{N\times p}$ is the data matrix with the N observed vectors of size p in its rows and $\widetilde{\mathbf{X}}$ is its column-centered version, are calculated via the SVD of $\widetilde{\mathbf{X}}$ instead of the eigendecomposition of $\widehat{\mathbf{C}}_N$. Namely, in the case of the empirical covariance, the power iteration can be performed at the level of the data. This is particularly useful in high-dimensional statistics, where $p \gg N$, or when the observations live on multi-dimensional domains, as is the case in the following example.

Example 2. *Let $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_N\in\mathbb{R}^{K\times K}$ be independent realizations of a random matrix-valued variable $\mathbf{X}\in\mathbb{R}^{K\times K}$ with a zero mean. We want to estimate the modes of variation of $\mathbf{X}$, i.e. to calculate the eigensurfaces of $\widehat{\mathbf{C}}_N = \frac{1}{N}\sum_{n=1}^N \mathbf{X}_n\otimes\mathbf{X}_n$. If $\widehat{\mathbf{C}}_N \in R^{K\times K\times K\times K}$ was explicitly formed, a single step of the power iteration method would take $\mathcal{O}(K^4)$ operations. On the other hand, note that if $\mathbf{V}_1^{(k)}$ is the k-th step approximation of the leading eigensurface $\mathbf{V}_1$, then we have*

$$\widehat{\mathbf{C}}_N\mathbf{V}_1^{(k)} = \frac{1}{N}\sum_{n=1}^N \langle \mathbf{X}_n, \mathbf{V}_1^{(k)}\rangle \mathbf{X}_n, \tag{3.6}$$

which can be calculated instead in $\mathcal{O}(NK^2)$ operations. The difference is considerable already if $K\approx N$. Moreover, the same is true for the memory requirements.

Notice the slight ambiguity in the previous example, especially in the left-hand side of equation (3.6). Since $\widehat{\mathbf{C}}_N$ is a tensor of order 4, it is not immediately clear how it should be applied to a matrix $\mathbf{V}_1^{(k)}$, and whether this leads to the right-hand side of (3.6). One could vectorize the observations (i.e. define $\mathbf{x}_n := \mathrm{vec}(\mathbf{X}_n)\in\mathbb{R}^{K^2}$, $n=1,\ldots,N$) and work with the vectors instead. Then $\widehat{\mathbf{C}}_N$ would be matricized into $\mathrm{mat}(\widehat{\mathbf{C}}_N)\in\mathbb{R}^{K^2\times K^2}$, $\mathbf{V}_1^{(k)}$ would be replaced by a vector $\mathbf{v}_1^{(k)}$, and equation (3.6) would turn into

$$\mathrm{mat}(\widehat{\mathbf{C}}_N)\mathbf{v}_1^{(k)} = \frac{1}{N}\sum_{n=1}^N \mathbf{x}_n\mathbf{x}_n^\top\mathbf{v}_1^{(k)}.$$

The right-hand side of the previous formula becomes the right-hand side of (3.6) after matricization.

In the following section, we develop an operator, which will allow us to generalize the power iteration method to both continuous and multi-dimensional domains, without the need to ever vectorize or matricize the objects at hand. This development is more general, compared to the one in Section 1.7.2.

3.2 Partial Inner Product

Recall the tensor product operators defined in equation (1.3). The symbols $\otimes_1$ and $\otimes_2$ themselves can be understood as mappings:

$$\begin{aligned}\otimes_1 &: [\mathcal{H}_1 \times \mathcal{H}_2] \times \mathcal{H}_1 \to \mathcal{H}_2,\\ \otimes_2 &: [\mathcal{H}_1 \times \mathcal{H}_2] \times \mathcal{H}_2 \to \mathcal{H}_1.\end{aligned}$$

We develop the partial inner products T_1 and T_2 by extending the definition of the tensor product operators $\otimes_1$ and $\otimes_2$ from the Cartesian product space $\mathcal{H}_1 \times \mathcal{H}_2$ in the previous equations to the richer outer product space $\mathcal{H}_1 \otimes \mathcal{H}_2$. This is straightforward in principle, because the finite linear combinations of the elements of $\mathcal{H}_1 \times \mathcal{H}_2$ are by definition dense in $\mathcal{H}_1 \otimes \mathcal{H}_2$.

Definition 13. *Let $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$. The partial inner products are the two unique bi-linear operators $T_1 : \mathcal{H} \times \mathcal{H}_1 \to \mathcal{H}_2$ and $T_2 : \mathcal{H} \times \mathcal{H}_2 \to \mathcal{H}_1$ defined by*

$$\begin{aligned}T_1(x \otimes y, e) &= (x \otimes_1 y)e, \quad x, e \in \mathcal{H}_1,\ y \in \mathcal{H}_2,\\ T_2(x \otimes y, f) &= (x \otimes_2 y)f, \quad x \in \mathcal{H}_1,\ y, f \in \mathcal{H}_2.\end{aligned}$$

The definition includes the claim of uniqueness of the partial inner products, which needs to be discussed. Consider a fixed element G from the set (1.6), which is dense in $\mathcal{H}_1 \otimes \mathcal{H}_2$, i.e. let $G = \sum_{j=1}^m x_j \otimes y_j$. Then, for $e \in \mathcal{H}_1$ and $f \in \mathcal{H}_2$, we have

$$\langle e, T_2(G, f)\rangle_{\mathcal{H}_1} = \Big\langle e, \sum_{j=1}^m \langle y_j, f\rangle_{\mathcal{H}_2} x_j \Big\rangle_{\mathcal{H}_1} = \sum_{j=1}^m \langle x_j, e\rangle_{\mathcal{H}_1} \langle y_j, f\rangle_{\mathcal{H}_2} = \sum_{j=1}^m \langle x_j \otimes y_j, e \otimes f\rangle_{\mathcal{H}} = \langle G, e \otimes f\rangle_{\mathcal{H}}.$$

Choosing $T_2(G, f)$ for e in the previous equation, we obtain from the Cauchy-Schwarz inequality that

$$\|T_2(G,f)\|^2_{\mathcal{H}_1} = \langle T_2(G,f), T_2(G,f)\rangle_{\mathcal{H}_1} = \langle G, T_2(G,f) \otimes f\rangle_{\mathcal{H}} \le \|G\|_{\mathcal{H}} \|T_2(G,f)\|_{\mathcal{H}_1} \|f\|_{\mathcal{H}_2}. \tag{3.7}$$

Hence $\|T_2(G,f)\|_{\mathcal{H}_1} \le \|G\|_{\mathcal{H}} \|f\|_{\mathcal{H}_2}$. Therefore, $T_2(\cdot, f)$ can be continuously extended from the set (1.6) to the whole space $\mathcal{H}$, and the uniqueness holds true.

Example 3. *Let* $\mathcal{H}_1 = \mathbb{R}^m$*,* $\mathcal{H}_2 = \mathbb{R}^n$*,* $u \in \mathbb{R}^m$*, and* $v, y \in \mathbb{R}^n$*. Let us denote* $\mathbf{G} = u \otimes v = uv^\top \in \mathbb{R}^{m\times n} = \mathbb{R}^m \otimes \mathbb{R}^n$*. By definition, we have*

$$T_2(\mathbf{G}, y) = \langle v, y\rangle u = v^\top y u = \mathbf{G}y.$$

Since matrix-vector multiplication is a bi-linear operator, it follows from the uniqueness proven above that the partial inner product is nothing else (with this particular choice of spaces) than matrix-vector multiplication. Thus $T_2(\mathbf{G}, y) = \mathbf{G}y$ *holds for any* $\mathbf{G} \in \mathbb{R}^{m\times n}$*. Similarly, for* $x \in \mathcal{H}_1$*, we have* $T_1(\mathbf{G}, x) = \mathbf{G}^\top x$*.*

We now show that the partial inner product has an explicit integral representation on any $\mathcal{L}^2$ space.

Proposition 10. *Let* $\mathcal{H}_1 = \mathcal{L}^2(E_1, \mathcal{E}_1, \mu_1)$*,* $\mathcal{H}_2 = \mathcal{L}^2(E_2, \mathcal{E}_2, \mu_2)$*, and* $g \in \mathcal{H}_1 \otimes \mathcal{H}_2$*,* $v \in \mathcal{H}_2$*. If we denote* $u = T_2(g, v)$*, then*

$$u(t) = \int_{E_2} g(t, s) v(s) \,\mathrm{d}\mu_2(s). \tag{3.8}$$

Proof. Since g is an $\mathcal{L}^2$ kernel, there is a Hilbert-Schmidt operator $G : \mathcal{H}_2 \to \mathcal{H}_1$ with the singular value decomposition $G = \sum \sigma_j e_j \otimes_2 f_j$ and kernel given by $g(t, s) = \sum \sigma_j e_j(t) f_j(s)$ (note that the sum only converges in the $\mathcal{L}^2$ sense, not uniformly). By isometry, from the SVD we have $g = \sum \sigma_j e_j \otimes f_j$, so

$$u = T_2(g, v) = \sum_{j=1}^{\infty} \sigma_j T_2(e_j \otimes f_j, v) = \sum_{j=1}^{\infty} \sigma_j \langle f_j, v\rangle e_j = \sum_{j=1}^{\infty} \sigma_j e_j \int_{E_2} f_j(s) v(s) \,\mathrm{d}\mu_2(s)$$

and hence by Fubini's theorem:

$$u(t) = \sum_{j=1}^{\infty} \sigma_j e_j(t) \int_{E_2} f_j(s) v(s) \,\mathrm{d}\mu_2(s) = \int_{E_2} \left[\sum_{j=1}^{\infty} \sigma_j e_j(t) f_j(s)\right] v(s) \,\mathrm{d}\mu_2(s)$$

from which the claim follows. □

Note that the proposition covers Example 3 for a suitable choice of the Hilbert spaces. From the computational perspective, the four-dimensional discrete case is the most important one. Hence we state it as the following corollary.

Corollary 3. *Let* $K_1, K_2 \in \mathbb{N}$*,* $\mathcal{H}_1 = \mathbb{R}^{K_1\times K_1}$*,* $\mathcal{H}_2 = \mathbb{R}^{K_2\times K_2}$*. Let* $\mathbf{G} \in \mathcal{H}_1 \otimes \mathcal{H}_2 =: \mathcal{H}$ *and* $\mathbf{V} \in \mathcal{H}_2$*. If we denote* $\mathbf{U} = T_2(\mathbf{G}, \mathbf{V})$*, we have*

$$\mathbf{U}[i, j] = \sum_{k=1}^{K_2} \sum_{l=1}^{K_2} \mathbf{G}[i, j, k, l] \mathbf{V}[k, l], \quad \forall i, j = 1, \ldots, K_1. \tag{3.9}$$

Remark 7. *Definition 13 is more general than the respective definitions provided by Bagchi and Dette (2020) and Dette et al. (2020). Still, the partial inner product can be defined to work with arbitrary dimensions. For* $J \subset \{1, \dots, d\}$*, let*

$$\mathcal{H} = \mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_d = \bigotimes_{j=1}^{d} \mathcal{H}_j, \quad \mathcal{H}_J := \bigotimes_{j \in J} \mathcal{H}_j, \quad \mathcal{H}_{-J} := \bigotimes_{j \notin J} \mathcal{H}_j.$$

We can define $T_J : \mathcal{H} \times \mathcal{H}_J \to \mathcal{H}_{-J}$ *to be the unique bi-linear operator such that*

$$T_J(X \otimes Y, A) = \langle X, A \rangle_{\mathcal{H}_J} Y, \quad \forall A, X \in \mathcal{H}_J, \forall Y \in \mathcal{H}_{-J}.$$

Note that $\mathcal{H}_1 \otimes \mathcal{H}_2$ *is isomorphic to* $\mathcal{H}_2 \otimes \mathcal{H}_1$ *with the isomorphism given by* $\Phi(x \otimes y) = y \otimes x$, $\forall x \in \mathcal{H}_1, y \in \mathcal{H}_2$*, and the same holds for products of multiple spaces. Hence we can always permute dimensions as we wish. Thus we can w.l.o.g. assume that* $J = \{1, \dots, d'\}$ *for some* $d' < d$*. Proving uniqueness is now the same as it was in the case of Definition 13, and Proposition 10 can be generalized to multiple dimensions as well.*

We can now go back to Example 2 and see that the tensor-matrix product in equation (3.6) *can be written as* $T_{\{3,4\}}(\widehat{\mathbf{C}}_N, \mathbf{V}_1^{(k)})$*. However, this level of generality will not be needed. We will stick to Definition 13, and write the tensor-matrix product in* (3.6) *as* $T_2(\widehat{\mathbf{C}}_N, \mathbf{V}_1^{(k)})$ *with a proper choice of the Hilbert spaces for Definition 13.*

The following lemma explores some basic properties of the partial inner product.

Lemma 6. *Let* $C \in \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$*,* $W_1 \in \mathcal{S}_2(\mathcal{H}_1)$ *and* $W_2 \in \mathcal{S}_2(\mathcal{H}_2)$*. Then the following claims hold.*

(a) If C *is separable, i.e.* $C = A \tilde{\otimes} B$*, then* $T_2(C, W_2) = \langle B, W_2 \rangle A$ *and* $T_1(C, W_1) = \langle A, W_1 \rangle B$*.*

(b) If C, W_1 *and* W_2 *are positive semi-definite (resp. self-adjoint), then* $T_2(C, W_2)$ *and* $T_1(C, W_1)$ *are also positive semi-definite (resp. self-adjoint).*

(c) If $\mathcal{H}_1 = \mathcal{L}^2(E_1, \mathcal{E}_1, \mu_1)$ *and* $\mathcal{H}_2 = \mathcal{L}^2(E_2, \mathcal{E}_2, \mu_2)$ *and the kernels of* C*,* W_1 *and* W_2 *are non-negative (resp. stationary, resp. banded), then the kernels of* $T_2(C, W_2)$ *and* $T_1(C, W_1)$ *are also non-negative (resp. stationary, resp. banded).*

Proof. The first claim follows directly from the definition of the partial inner product.

In the remaining two claims, we want to show that if both C and the respective weighting W_1 or W_2 have a certain property, then the partial inner product will retain that property.

In the case of self-adjointness or stationarity, the claim follows immediately from the definition because the set of all Hilbert-Schmidt operators having one of these properties is a closed linear subspace of a space of Hilbert-Schmidt operators, and hence it is itself a Hilbert space. Thus we can constrain ourselves to only work on such a subspace and the claim follows directly from validity of Definition 13.

As for positive semi-definitness, consider the eigendecompositions $C = \sum \lambda_j g_j \otimes g_j$ and $W_2 = \sum \alpha_j h_j \otimes h_j$. Then we have

$$T_2(C, W_2) = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \lambda_j \alpha_i T_2(g_j \otimes g_j, h_i \otimes h_i) = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \lambda_j \alpha_i T_2(g_j, h_i) \otimes T_2(g_j, h_i),$$

where the last equality can be verified on rank-1 elements g_j. Thus $T_2(C, W_2)$ is a weighted sum of quadratic forms with weights given by the non-negative eigenvalues of C and W_2. As such, $T_2(C, W_2)$ must be positive semi-definite.

Finally, both bandedness and non-negativity of the kernel can by seen immediately from the integral representation given by Proposition 10. $\square$

Part 1 of Lemma 6 exemplifies why operators T_1 and T_2 are called partial inner products. One can also see this directly from Definition 13. If the covariance is not exactly separable, the partial inner product is at the basis of the algorithm for finding a separable proxy to the covariance. The necessity to choose (correctly scaled) weights W_1 and W_2 is bypassed via an iterative procedure, which can be understood as a generalization of the power iteration method.

3.2.1 Generalized Power Iteration

Proposition 11. *For $C \in \mathcal{H}_1 \otimes \mathcal{H}_2$, let $C = \sum_{j=1}^{\infty} \sigma_j A_j \otimes B_j$ be a decomposition such that $|\sigma_1| > |\sigma_2| \geq |\sigma_3| \geq \dots$ and $\{A_j\}$ is an ONB in $\mathcal{H}_1$ and $\{B_j\}$ is an ONB in $\mathcal{H}_2$. Let $V^{(0)} \in \mathcal{H}_2$ be such that $\|V^{(0)}\| = 1$ and $\langle B_1, V^{(0)} \rangle > 0$. Then the sequences $\{U^{(k)}\}$ and $\{V^{(k)}\}$ formed via the recurrence relation*

$$U^{(k+1)} = \frac{T_2(C, V^{(k)})}{\|T_2(C, V^{(k)})\|}, \quad V^{(k+1)} = \frac{T_1(C, U^{(k+1)})}{\|T_1(C, U^{(k+1)})\|}$$

converge to A_1 and B_1, respectively. The convergence speed is linear with the rate given by the spacing between σ_1 and σ_2.

Proof. Let $C_1 := \sum_{j=1}^{\infty} \sigma_j^2 A_j \otimes A_j$ and $C_2 := \sum_{j=1}^{\infty} \sigma_j^2 B_j \otimes B_j$. We will abuse the notation slightly and denote for any $k \in \mathbb{N}$

$$C_1^k := \sum_{j=1}^{\infty} \sigma_j^{2k} A_j \otimes A_j \quad \& \quad C_2^k := \sum_{j=1}^{\infty} \sigma_j^{2k} B_j \otimes B_j.$$

Note that if C was an operator, it would hold that $C_1 = CC^*$ and $C_2 = C^*C$, while C_1^k and C_2^k would denote the powers as usual. However, we aim for a more general statement, forcing us to view C as an element of a product space rather than an operator, so the "powers" serve just as a notational convenience in this proof. Also, the proportionality sign is used here to avoid the necessity of writing down the scaling constants for unit norm elements.

From the recurrence relation, and the definition of the partial inner product, we have

$$\begin{aligned} V^{(1)} \propto T_1\Big(C, T_2(C, V^{(0)})\Big) &= T_1\left(\sum_{j=1}^{\infty} \sigma_j A_j \otimes B_j, \sum_{i=1}^{\infty} \sigma_i \langle B_i, V^{(0)}\rangle A_i\right) \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \sigma_i \sigma_j \langle B_i, V^{(0)}\rangle T_1(A_j \otimes B_j, A_i). \end{aligned}$$

Since $T_1(A_j \otimes B_j, A_i) = \langle A_j, A_i\rangle B_j = \mathbb{1}_{[i=j]} B_j$, we have

$$V^{(1)} \propto \sum_{j=1}^{\infty} \sigma_j^2 \langle B_i, V^{(0)}\rangle B_j = T_2(C_2, V^{(0)}).$$

By the same token we have $V^{(2)} \propto T_2(C_2, V^{(1)})$, from which we obtain similarly

$$\begin{aligned} V^{(2)} &\propto T_2\left(\sum_{j=1}^{\infty} \sigma_j^2 B_j \otimes B_j, \sum_{i=1}^{\infty} \sigma_i^2 \langle B_i, V^{(0)}\rangle B_i\right) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \sigma_i^2 \sigma_j^2 \langle B_i, V^{(0)}\rangle T_2(B_j \otimes B_j, B_i) \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \sigma_i^2 \sigma_j^2 \langle B_i, V^{(0)}\rangle \langle B_j, B_i\rangle B_j = \sum_{j=1}^{\infty} \sigma_j^4 \langle B_j, V^{(0)}\rangle B_j = T_2\left(\sum_{j=1}^{\infty} \sigma_j^4 B_j \otimes B_j, V^{(0)}\right) \\ &= T_2(C_2^2, V^{(0)}). \end{aligned}$$

By induction, we have $V^{(k)} \propto T_2(C_2^k, V^{(0)})$ for $k = 1, 2, \ldots$.

Now we express the starting point $V^{(0)}$ in terms of the ONB $\{B_j\}$. Let $V^{(0)} = \sum_{j=1}^{\infty} \beta_j B_j$, where we have $\beta_j = \langle B_j, V^{(0)}\rangle$. Then we can express

$$\begin{aligned} V^{(k)} \propto T_2(C_2^2, V^{(k)}) &= T_2\left(\sum_{j=1}^{\infty} \sigma_j^{2k} B_j \otimes B_j, \sum_{i=1}^{\infty} \beta_i B_i\right) \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \sigma_j^{2k} \beta_i \langle B_i, B_j\rangle B_j = \sum_{j=1}^{\infty} \sigma_j^{2k} \beta_j B_j. \end{aligned}$$

Hence we have an explicit formula for the k-th step:

$$V^{(k)} = \frac{B_1 + R^{(k)}}{\|B_1 + R^{(k)}\|},$$

where $R^{(k)} = \sum_{j=2}^{\infty} \left(\frac{\sigma_j}{\sigma_1}\right)^{2k} \frac{\beta_j}{\beta_1} B_j$.

It remains to show that $\|R^{(k)}\| \to 0$ for $k \to \infty$. Due to the decreasing ordering of the scores $\{\sigma_j\}$, we have

$$\begin{aligned}\|R^{(k)}\|^2 &= \sum_{j=2}^{\infty} \left(\frac{\sigma_j}{\sigma_1}\right)^{4k} \frac{\beta_j^2}{\beta_1^2} \leq \left(\frac{\sigma_2}{\sigma_1}\right)^{4k-2} \sum_{j=2}^{\infty} \left(\frac{\sigma_j}{\sigma_1}\right)^{2} \frac{\beta_j^2}{\beta_1^2} \\ &= \left(\frac{\sigma_2}{\sigma_1}\right)^{4k-2} \frac{1}{\sigma_1^2\beta_1^2} \sum_{j=1}^{\infty} \sigma_j^2 \beta_j \leq \left(\frac{\sigma_2}{\sigma_1}\right)^{4k-2} \frac{1}{\sigma_1^2\beta_1^2} \|C\|,\end{aligned}$$

where in the last inequality we used that $|\beta_j| \leq 1$. Since $\sigma_1 > \sigma_j$ for $j \geq 2$, we see that the remainder $R^{(k)}$ goes to zero.

The proof of the statement concerning the sequence $\{U^{(k)}\}$ follows the same steps. □

The assumption $\langle B_1, V^{(0)} \rangle > 0$ in the previous proposition can be weakened to only $\langle B_1, V^{(0)} \rangle \neq 0$. In that case, the sequences do not necessarily converge, but all limit points span the appropriate spaces. The proof is similar, except that some care has to be taken at the level of the signs. The sign ambiguity is caused by the fact that the separable components are (even in the case of non-zero spacing between the scores $\{\sigma_j\}$) unique only up to the sign.

Example 4. *In the previous proposition, let $\mathcal{H}_1 = \mathcal{S}_2(\mathcal{H}_1')$ and $\mathcal{H}_2 = \mathcal{S}_2(\mathcal{H}_2')$ for some Hilbert spaces $\mathcal{H}_1', \mathcal{H}_2'$. Then C is the covariance operator of a random element $X \in \mathcal{H}_1' \otimes \mathcal{H}_2'$, and the previous proposition shows that the separable proxy to C, i.e. a solution to*

$$\operatorname*{arg\,min}_{A \in \mathcal{H}_1' \otimes \mathcal{H}_1', B \in \mathcal{H}_2' \otimes \mathcal{H}_2'} \|C - A \,\tilde{\otimes}\, B\|_2^2, \tag{3.10}$$

can be found via the power iteration method, consisting of a series of partial inner products.

Let us take in the previous proposition $\mathcal{H}_1 = \mathcal{H}_1' \otimes \mathcal{H}_2'$ and $\mathcal{H}_2 = \mathcal{H}_1' \otimes \mathcal{H}_2'$ for some Hilbert spaces $\mathcal{H}_1', \mathcal{H}_2'$. Then C is still (isometric to) the covariance operator of a random element $X \in \mathcal{H}_1' \otimes \mathcal{H}_2'$, only this time viewed as an element of a different, probably more natural space. Hence the SVD in Proposition 11 is in fact the eigendecomposition here. The previous proposition shows that the leading eigenvalue-eigenvector pair, i.e. the solution to

$$\operatorname*{arg\,min}_{e_1 \in \mathcal{H}_1' \otimes \mathcal{H}_2'} \|C - e_j \otimes e_j\|_2^2$$

can be found via the power iteration method, consisting of a series of partial inner products.

Table 3.1 Power iteration method for finding the leading separable components on a general Hilbert space.

```
**Input** $C \in \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$, initial guesses $A_1, \ldots, A_R \in \mathcal{H}_1 \otimes \mathcal{H}_1$
**for** r=1,...,R
    $\widetilde{C} := C - \sum_{j=1}^{r-1} \sigma_j A_j \tilde{\otimes} B_j$
    **repeat**
        $B_r := T_1(\widetilde{C}, A_r)$
        $B_r := B_r / \|B_r\|$
        $A_r := T_2(\widetilde{C}, B_r)$
        $\sigma_r := \|A_r\|$
        $A_r := A_r / \sigma_r$
    **until convergence**
**end for**
**Output** $\sigma_1, \ldots, \sigma_R$, $A_1, \ldots, A_R$, $B_1, \ldots B_R$
```

As shown above, the power iteration method can be performed in an arbitrary Hilbert space, and it can be used to find the best separable approximation of a covariance operator. In this chapter, we expand our attention beyond separability (3.10), to the solution of (3.5) with $R \in \mathbb{N}$, i.e. searching for the best R-separable approximation. This optimisation problem can be solved via Algorithm 3.1, which contains subsequent search for R separable components, deflating the covariance matrix for every previously found component, and standardizing the components to have norm one.

Remark 8. 1. *In the case of* $\mathcal{H}_1 = \mathbb{R}^m$ *and* $\mathcal{H}_2 = \mathbb{R}^n$, *a similar problem was studied by Van Loan and Pitsianis (1993), who showed that the solution to the optimization problem*

$$\underset{\mathbf{A} \in \mathbb{R}^{m \times m}, \mathbf{B} \in \mathbb{R}^{n \times n}}{\arg\min} \|\mathbf{C} - \mathbf{A} \tilde{\otimes} \mathbf{B}\|_F^2 \tag{3.11}$$

can be found as the leading singular subspace of a suitable permutation of $\mathrm{mat}(\mathbf{C})$, *i.e. the matricization of* $\mathbf{C}$ *(see Van Loan and Golub, 1983). When this leading singular subspace is found via power iteration, a single step is given by the partial inner product and the algorithm provided as Framework 2 by Van Loan and Pitsianis (1993) corresponds to Algorithm 3.1 for* $R = 1$. *Indeed, it is straightforward to verify that the element-wise formulas provided by Van Loan and Pitsianis (1993) correspond to* (3.9) *in the case of* $\mathcal{H}_1 = \mathbb{R}^m$ *and* $\mathcal{H}_2 = \mathbb{R}^n$.

2. *A notable portion of the paper of Van Loan and Pitsianis (1993) was devoted to proving that if* $C \in \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$ *has a certain property (one of those discussed in*

Lemma 6), then the leading eigenvectors A_1 and B_1 retain that property (Van Loan and Pitsianis, 1993, only discuss this in the discrete case). Owing to the generality of our partial inner product definition, this can be argued quite simply on the algorithmic basis. By Proposition 11, the power iteration method converges to A_1 and B_1 regardless of the starting point (as long as the starting point is not orthogonal to the solution). Consider the starting point satisfying the property in question. Then by Lemma 6, the algorithm will never leave the closed subset defined by the property in question (note that e.g. positive semi-definitness does not characterize a closed subspace, *but it still designates a closed* subset*), and hence the limit of the power iteration method will also have this property.*

3. *Bagchi and Dette (2020) claim it is hard to find the minimum of* (3.10) *in a general Hilbert space, and hence they settle on a procedure which can be translated as stopping the power iteration method after just a single iteration. However, the true minimizer of* (3.10) *is, in fact, obtainable via power iteration, and we focus on this minimizer.*

3.3 Estimation

Let $X_1, \dots, X_N$ be i.i.d. elements in $\mathcal{H}_1 \otimes \mathcal{H}_2$, and $R \in \mathbb{N}$ be given (the choice of R will be discussed later). We propose the following estimator for the covariance operator:

$$\widehat{C}_{R,N} = \arg\min_{G} \left|\left|\left| \widehat{C}_N - G \right|\right|\right|_2^2 \quad \text{s.t.} \quad \mathrm{DoS}(G) \leq R, \tag{3.12}$$

where $\widehat{C}_N = N^{-1} \sum_{n=1}^N (X_n - \bar{X}) \otimes (X_n - \bar{X})$ is the empirical covariance. The estimator $\widehat{C}_{R,N}$ is the best R-separable approximation to the empirical covariance $\widehat{C}_N$. This leads to an estimator of the form

$$\widehat{C}_{R,N} = \sum_{r=1}^{R} \widehat{\sigma}_r \widehat{A}_r \tilde{\otimes} \widehat{B}_r, \tag{3.13}$$

where $\widehat{\sigma}_1, \dots, \widehat{\sigma}_R$, $\widehat{A}_1, \dots, \widehat{A}_r$, $\widehat{B}_R, \dots \widehat{B}_R$ are the output of Algorithm 3.1 applied to the empirical covariance estimator $\widehat{C}_N$ as the input.

Now, assume the data are observed as matrices, i.e. $\mathbf{X}_1, \dots, \mathbf{X}_N \in \mathbb{R}^{K_1 \times K_2}$. Then the covariance $\mathbf{C}$ and the estimator $\widehat{\mathbf{C}}_{R,N}$ of (3.13) are also discrete, namely $\mathbf{C}, \widehat{\mathbf{C}}_{R,N} \in \mathbb{R}^{K_1 \times K_2 \times K_1 \times K_2}$. Recall that we use boldface to emphasize discrete objects, and assume for simplicity $K_1 = K_2 =: K$. In the theoretical development of Section 3.5, we differentiate between multivariate data and functional data observed on a dense grid (of size $K \times K$), the latter being our primary interest. But for now, the distinction is immaterial: we only assume we are in the discrete case to exemplify the computational benefits of the partial inner product.

Table 3.2: Time and memory complexities of computing the empirical covariance estimator and a separable estimator when N surfaces are observed discretely on a $K \times K$ grid.

Method	Memory complexity	Time complexity		
		Estimation	Application	Inversion
empirical	$\mathcal{O}(K^4)$	$\mathcal{O}(NK^4)$	$\mathcal{O}(K^4)$	$\mathcal{O}(K^6)$
separable	$\mathcal{O}(K^2)$	$\mathcal{O}(NK^3)$	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$

The key observation here is that the partial inner product operations required in Algorithm 3.1 can be carried out directly on the level of the data, without the need to explicitly form or store the empirical covariance estimator $\widehat{C}_N$. In the discrete case, for example, it is straightforward to verify from Corollary 3 that

$$\begin{aligned} T_1(\widehat{\mathbf{C}}_N, \mathbf{A}) &= \frac{1}{N}\sum_{n=1}^{N} T_1(\mathbf{X}_n \otimes \mathbf{X}_n, \mathbf{A}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{X}_n^\top \mathbf{A}\mathbf{X}_n, \\ T_2(\widehat{\mathbf{C}}_N, \mathbf{B}) &= \frac{1}{N}\sum_{n=1}^{N} T_2(\mathbf{X}_n \otimes \mathbf{X}_n, \mathbf{B}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{X}_n \mathbf{B}\mathbf{X}_n^\top. \end{aligned} \tag{3.14}$$

The immense popularity of the separability assumption stems from the computational savings it entails. These are captured in Table 3.2. Notice the reduced estimation complexity of the separable model. Since the convergence rate of the power iteration method is linear, and a single iteration can be evaluated efficiently on the level of data due to (3.14), our approach can also be used to estimate the R-separable model with the same efficiency (times R, of course). Moreover, we will show in Section 3.4 that it is possible to apply and (numerically) invert an R-separable covariance (3.13) with the same computational costs as for a separable model. Hence the complexities reported in Table 3.2 for a separable model are also valid for an R-separable approximation.

Remark 9. *To the best of our knowledge, the only approach to efficient estimation of a separable model, which reduces the estimation complexity to that reported in Table 3.2, is partial tracing of Aston et al. (2017). Partial tracing achieves this complexity by considering only some of the available raw covariances (i.e. some of the the cross-products of two sampled entries on the same surface). In contrast, our approach uses all the available raw covariances composing the empirical estimator. The computational savings are achieved by reducing the number of necessary operations via the formulas in* (3.14)*. Moreover, our estimation procedure facilitates the search for an approximation (either R-separable or entirely separable), while the partial tracing estimator lacks any optimality properties, and assumes separability as a model. This can be said also for other approaches built upon partial tracing, such as the weakly separable model (Lynch and Chen, 2018).*

Table 3.3 Constructing the R-separable estimator from discrete measurements.

```
Input $\mathbf{X}_1,\dots,\mathbf{X}_N \in \mathbb{R}^{K_1\times K_2}$, initial guesses $\mathbf{A}_1,\dots,\mathbf{A}_R \in \mathbb{R}^{K_1\times K_2}$
for r=1,...,R
    repeat
        $\mathbf{B}_r = \frac{1}{N}\sum_{n=1}^N(\mathbf{X}_n-\bar{\mathbf{X}})^\top\mathbf{A}_r(\mathbf{X}_n-\bar{\mathbf{X}}) - \sum_{j=1}^{r-1}\sigma_j\langle\mathbf{A}_r,\mathbf{A}_j\rangle\mathbf{A}_j$
        $\mathbf{B}_r = \mathbf{B}_r/\|\mathbf{B}_r\|$
        $\mathbf{A}_r = \frac{1}{N}\sum_{n=1}^N(\mathbf{X}_n-\bar{\mathbf{X}})^\top\mathbf{B}_r(\mathbf{X}_n-\bar{\mathbf{X}}) - \sum_{j=1}^{r-1}\sigma_j\langle\mathbf{B}_r,\mathbf{B}_j\rangle\mathbf{B}_j$
        $\sigma_r = \|\mathbf{A}_r\|$
        $\mathbf{A}_r = \mathbf{A}_r/\sigma_r$
    until convergence
end for
Output $\sigma_1,\dots,\sigma_R$, $\mathbf{A}_1,\dots,\mathbf{A}_R$, $\mathbf{B}_1,\dots\mathbf{B}_R$
```

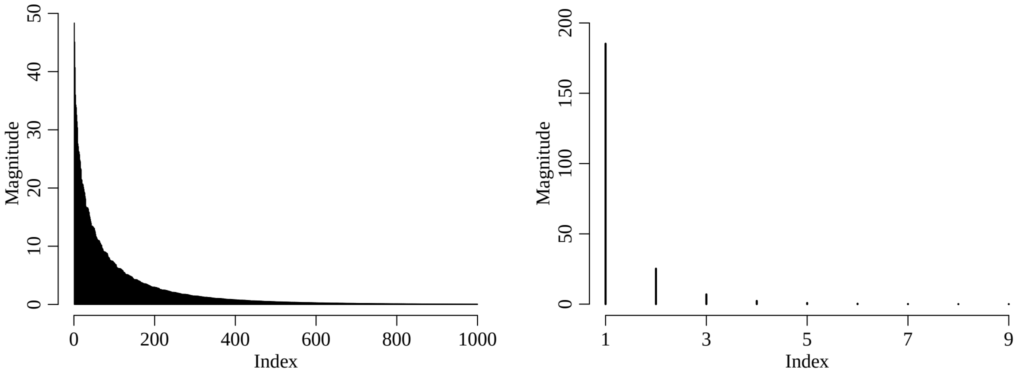
Comparing equations (3.14) to (3.6), one can notice that computing a single separable component of C is slightly more expensive than computing a single eigenvalue-eigenvector pair, by a factor of the grid size K. This modest computational overheard is paid in hope that the (approximate) degree-of-separability of C is smaller than the (approximate) rank of C, leading to statistical savings, and potentially also computational savings, depending on the grid size. The following example provides an illustration.

Example 5. *Consider the Irish Wind data set of Haslett and Raftery (1989). The data set was modeled with a separable covariance structure at first, before Gneiting (2002) and Gneiting et al. (2006) argued that separability has hardly justifiable practical consequences for this data set. Later, separability was formally rejected by a statistical test in (Bagchi and Dette, 2020). A non-separable parametric covariance model was developed specifically for the Irish Wind by Gneiting (2002). We consider the fitted parametric covariance model as the ground truth C (see Section 3.6.1 for a full specification). We plot the eigenvalues of C (evaluated on a 50×50 grid in $[0,20]^2$ domain) and the separable component scores of C (evaluated on the same grid) in Figure 3.1. While C is clearly not low-rank (Figure 3.1, left), it is approximately of very low degree-of-separability (Figure 3.1, right). We will return to this particular covariance in Section 3.6.1 and show that choices $R=2$ or $R=3$ lead to very good approximations of C.*

3.3.1 Degree-of-Separability Selection

Given $R\in\mathbb{N}$, we have demonstrated how to construct and invert an R-separable proxy of the covariance, based on i.i.d. observations. It now remains to discuss how to choose the degree-of-separability R. Recall that we do not assume that the covariance in question

Figure 3.1: Eigenvalues (**left**) and separable component scores (**right**) of the covariance from Example 5.



is R-separable per se, so there is no “correct" choice of R. In this context, R can be seen as governing the effective number of parameters being estimated from the data (the “degrees of freedom"), serving in effect as a regularization parameter. So, one can seek a positive integer R that minimizes the mean squared error

$$\mathbb{E}\left|\left|\left|\widehat{\mathbf{C}}_{R,N} - \mathbf{C}\right|\right|\right|_2^2. \tag{3.15}$$

The underlying bias-variance trade-off is precisely the reason why small values of R can lead to an improved mean squared error compared to the empirical covariance estimator. Intuitively, with increasing R, the estimator $\widehat{C}_{R,N}$ has increasing number of degrees of freedom, and more observations are needed to estimate it reliably. This will be also reflected in our theory (Section 3.5).

To determine an empirical surrogate of (3.15) and devise a cross-validation (CV) strategy similar to the one in the previous chapter. Note first that

$$\mathbb{E}\left|\left|\left|\widehat{\mathbf{C}}_{R,N} - \mathbf{C}\right|\right|\right|_2^2 = \mathbb{E}\left|\left|\left|\widehat{\mathbf{C}}_{R,N}\right|\right|\right|_2^2 - 2\mathbb{E}\langle\widehat{\mathbf{C}}_{R,N}, \mathbf{C}\rangle + |||\mathbf{C}|||_2^2.$$

Of the three terms on the right-hand side of the previous equation, the final term does not depend on R and hence it does not affect the minimization. The first term can be unbiasedly estimated by $\left|\left|\left|\widehat{\mathbf{C}}_{R,N}\right|\right|\right|_2^2$, and it remains to estimate the middle term. Denote by $\widehat{\mathbf{C}}_{R,N-1}^{(j)}$ the R-separable estimator constructed excluding the j-th datum $\mathbf{X}_j$. Due to independence between samples, we have

$$\mathbb{E}\langle\mathbf{X}_j, \widehat{\mathbf{C}}_{R,N-1}^{(j)}\mathbf{X}_j\rangle = \mathbb{E}\langle\widehat{\mathbf{C}}_{R,N-1}^{(j)}, \mathbf{X}_j \otimes \mathbf{X}_j\rangle = \langle\mathbb{E}\widehat{\mathbf{C}}_{R,N-1}^{(j)}, \mathbb{E}\mathbf{X}_j \otimes \mathbf{X}_j\rangle = \mathbb{E}\langle\widehat{\mathbf{C}}_{R,N}, C\rangle,$$

and hence the quantity $N^{-1}\sum_{j=1}^N\langle\mathbf{X}_j, \widehat{\mathbf{C}}_{R,N-1}^{(j)}\mathbf{X}_j\rangle$ is an unbiased estimator of the expected inner product between the truth and the estimator: $\mathbb{E}\langle\widehat{\mathbf{C}}_{R,N}, \mathbf{C}\rangle$. In summary, a CV strategy is to choose R as

$$\underset{R}{\arg\min} \quad \left\{ \left|\!\left|\!\left| \widehat{\mathbf{C}}_{R,N} \right|\!\right|\!\right|_2^2 - \frac{2}{N} \sum_{j=1}^{N} \langle \mathbf{X}_j, \widehat{\mathbf{C}}_{R,N-1}^{(j)} \mathbf{X}_j \rangle \right\}. \tag{3.16}$$

This procedure corresponds to leave-one-out CV, which is computationally prohibitive. In practice, we perform a 10-fold CV (see e.g. Murphy, 2012).

The choice (3.16) is inspired by analogy to the classical cross-validated bandwidth selection scheme in kernel density estimation (Wand and Jones, 1994). The typical CV scheme for PCA (e.g. Jolliffe, 1986) is based on finding a low-dimensional plane fitting the data cloud well, the low-dimensional plane being tied to the principal components. Such a scheme is not applicable here, because it degenerates: the first separable component alone might very well span the whole ambient space, thus projecting a datum on a subspace generated by a varying number of leading separable components will not be informative.

The CV scheme requires fitting the covariance repeatedly for different values of R. We fit the covariance for a maximal value of the degree-of-separability we are interested in or can hope to estimate reliably with our number of observations. The theoretical development of the following section can provide some guidance for this. Then, we can fit the covariance for this degree-of-separability only, and use a subset of the obtained decomposition for any smaller degree-of-separability. This still has to be done multiple times when cross-validating. Hence for very large data sets, a visual inspection is recommended: fitting the covariance once using a maximal relevant value of the degree-of-separability, one can then visualize the scores in the form of a scree plot (see Figure 3.3), and choose the degree based on this plot. We provide an example of this approach in Section 3.6.1, while a very detailed discussion on scree plots is given by Jolliffe (1986). We note that the interpretation of the j-th singular value σ_j as the additional variance, which is explained by considering the j-th separable term as opposed to only first $j-1$ terms, is still valid here. Hence, rule-of-thumb choices based on scree plots, such as requiring that the total variance explained by the chosen R components must be at least 95%, can be used as well.

3.4 Inversion

In this section, we are interested in solving a linear system

$$\widehat{\mathbf{C}}_{R,N} \mathbf{X} = \mathbf{Y}, \tag{3.17}$$

where $\widehat{\mathbf{C}}_{R,N} \in \mathbb{R}^{K \times K \times K \times K}$ is our R-separable estimator of equation (3.13). This linear system needs to be solved for the purposes of prediction or kriging, among other tasks.

The linear system (3.17) can be naively solved in $\mathcal{O}(K^6)$ operations, which is required for a general $\mathbf{C}$. However, if we had $R = 1$, i.e. $\widehat{\mathbf{C}}_{R,N}$ was separable, the system would be solveable in just $\mathcal{O}(K^3)$ operations. These huge computational savings are one of the main reasons for the immense popularity of the separability assumption. In the case of $R > 1$, we will not be able to find an explicit solution, but the iterative algorithm we develop here will be substantially faster than the $\mathcal{O}(K^6)$ operations needed for a general $\mathbf{C}$.

The crucial observation here is that $\widehat{\mathbf{C}}_{R,N}$ can be applied in $\mathcal{O}(K^3)$ operations:

$$\widehat{\mathbf{C}}_{R,N}\mathbf{X} = \sum_{r=1}^{R} \widehat{\sigma}_r \widehat{\mathbf{A}}_r \mathbf{X} \widehat{\mathbf{B}}_r. \tag{3.18}$$

To this end, it can be verified that $(\mathbf{A} \tilde{\otimes} \mathbf{B})\mathbf{X} = \mathbf{A}\mathbf{X}\mathbf{B}$ for $\mathbf{B}$ self-adjoint either directly from the definitions, or from the connection with the Kronecker product (cf. Remark 1). So, we can rewrite the linear system (3.17) as

$$\widehat{\sigma}_1 \widehat{\mathbf{A}}_1 \mathbf{X} \widehat{\mathbf{B}}_1 + \ldots + \widehat{\sigma}_R \widehat{\mathbf{A}}_R \mathbf{X} \widehat{\mathbf{B}}_R = \mathbf{Y}.$$

A system of this form is called a *linear matrix equation* and it has been extensively studied in the field of numerical linear algebra (see Palitta and Simoncini, 2020, and the numerous references therein). Even though there exist provably convergent specialized solvers (e.g. Xie et al., 2009), the simple preconditioned conjugate gradient (PCG) algorithm (Shewchuk, 1994) is the method of choice when $\widehat{\mathbf{C}}_{R,N}$ is symmetric and positive semi-definite, which is exactly our case. The conjugate gradient (CG) method works by iteratively applying the left hand side of the equation to a gradually updated vector of residuals. In theory, there can be up to $\mathcal{O}(K^2)$ iterations needed, which would lead to $\mathcal{O}(K^5)$ complexity. In practice, the algorithm is stopped much earlier; how early depends on the properties of the spectrum of the left hand side, which can be improved via preconditioning. We refer the reader to Shewchuk (1994) for the definitive exposition of CG.

The usage of a preconditioner $\mathbf{P} = \mathbf{V}\mathbf{V}^\top$ can be thought of as applying the standard conjugate gradient to the system $\widetilde{\mathbf{C}}\widetilde{\mathbf{X}} = \widetilde{\mathbf{Y}}$ with $\widetilde{\mathbf{C}} = \mathbf{V}^{-1}\widehat{\mathbf{C}}_{R,N}(\mathbf{V}^{-1})^\top$, $\widetilde{\mathbf{X}} = \mathbf{V}^\top\mathbf{X}$ and $\widetilde{\mathbf{Y}} = \mathbf{V}^{-1}\mathbf{Y}$. In our case, $\mathbf{P} = \widehat{\sigma}_1 \widehat{\mathbf{A}}_1 \tilde{\otimes} \widehat{\mathbf{B}}_1$ is a natural preconditioner, whose square-root $\mathbf{V}$ can be both obtained and applied easily due to Lemma 1.

This preconditioner is chosen because $\widehat{\sigma}_1 \widehat{\mathbf{A}}_1 \tilde{\otimes} \widehat{\mathbf{B}}_1$ is the leading term in $\widehat{\mathbf{C}}_{R,N}$. The more dominant the term is in $\widehat{\mathbf{C}}_{R,N}$, the flatter the spectrum of $\mathbf{V}^{-1}\widehat{\mathbf{C}}_{R,N}(\mathbf{V}^{-1})^\top$ is, and the fewer iterations are needed. This is a manifestation of a certain statistical-computational trade-off. The R-separable model can be, in theory, used to fit any covariance C, when R is taken large enough. However, the more dominant the leading (separable) term is, the better computational properties we have, see Section 3.6.4.

It remains to address the existence of a solution to the linear system (3.17). Note that we cannot guarantee that $\widehat{\mathbf{C}}_{R,N}$ is positive semi-definite. Lemma 6 says that $\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{B}}_1$ are positive semi-definite, and they will typically be positive definite for N sufficiently large (when $\widehat{\mathbf{C}}_N$ is positive definite). But $\widehat{\mathbf{C}}_N - \widehat{\mathbf{A}}_1 \tilde{\otimes} \widehat{\mathbf{B}}_1$ is necessarily indefinite, and hence we cannot say anything about the remaining terms. However, $\widehat{\mathbf{C}}_{R,N}$ is a consistent estimator of the true $\mathbf{C}$ for sufficiently large values of R (see Section 3.5 for a discussion on the rate of convergence of this estimator depending on R). So, for a large enough sample size and appropriate values of R, $\widehat{\mathbf{C}}_{R,N}$ cannot be far away from positive semi-definiteness. To eliminate practical anomalies, we will *positivize* the estimator. Doing this is also computationally feasible, as discussed below.

Due to (3.18), the power iteration method can be used to find the leading eigenvalue $\lambda_{\max}$ of $\widehat{\mathbf{C}}_{R,N}$ in $\mathcal{O}(K^3)$ operations. We can then find the smallest eigenvalue $\lambda_{\min}$ of $\widehat{\mathbf{C}}_{R,N}$ by applying the power iteration method to $\lambda_{\max}\mathbf{I} - \widehat{\mathbf{C}}_{R,N}$, where $\mathbf{I} \in \mathbb{R}^{K1 \times K_2 \times K_1 \times K_2}$ is the identity. Note that this is only a simple proof of concept, in practice we recommend to used a specialized solver to find the minimum eigenvalue, e.g. the one of Wu and Stathopoulos (2014). Subsequently, if $\lambda_{\min} < 0$, we can perturb $\widehat{\mathbf{C}}_{R,N}$ to obtain its positive semi-definite version:

$$\widehat{\mathbf{C}}^{+}_{R,N} = \widehat{\mathbf{C}}_{R,N} + (\epsilon - \lambda_{\min})\mathbf{I}, \tag{3.19}$$

where $\epsilon \geq 0$ is a potential further regularization.

The positivized estimator $\widehat{\mathbf{C}}^{+}_{R,N}$ is $(R+1)$-separable, so it can still be approached in the same spirit, with one exception. If the inverse problem is ill-conditioned and regularization is used, the preconditioner discussed above is no longer effective, since $\mathbf{A}_1$ or $\mathbf{B}_1$ may not be invertible. In this case, we use the preconditioner $\mathbf{P} = \widehat{\sigma}_1 \widehat{\mathbf{A}}_1 \tilde{\otimes} \widehat{\mathbf{B}}_1 + (\epsilon - \lambda_{\min})\mathbf{I}$, whose eigenvectors are still given by Lemma 1 and eigenvalues are simply inflated by $\epsilon - \lambda_{\min}$. This is preconditioning via the discrete Stein's equation, see Section 2.3.3.

The effectiveness of the proposed inversion algorithm is demonstrated in Section 3.6.1. We stress here that the potential need to regularize an estimator of a low degree-of-separability arises in a very different way from the necessity to regularize a (more typical) low-rank estimator. When the truncated eigendecomposition is used as an estimator, the spectrum is by construction singular and regularization is thus necessary for the purposes of prediction. Contrarily, when an estimator of low degree-of-separability is used, the spectrum of the estimator mimics that of C more closely. If C itself is well-conditioned, there may be no need to regularize, regardless of what degree-of-separability R is used as a cut-off. However, in the case of functional data observed on a dense grid, regularization may be necessary due to the spectral decay of C itself.

3.4.1 Prediction

As an important application, we use the R-separable estimator $\widehat{\mathbf{C}}_{R,N} = \sum_{r=1}^{R} \widehat{\sigma}_r \widehat{\mathbf{A}}_r \tilde{\otimes} \widehat{\mathbf{B}}_r$ to predict the missing values of a datum $\mathbf{X} \in \mathbb{R}^{K_1 \times K_2}$. In the case of a random vector, say $\mathrm{vec}(\mathbf{X})$, such that

$$\mathrm{vec}(\mathbf{X}) = \Big(\mathrm{vec}(\mathbf{X})_1, \mathrm{vec}(\mathbf{X})_2\Big)^\top \sim \left(0, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}\right),$$

where $\mathrm{vec}(\mathbf{X})_1$ is missing and $\mathrm{vec}(\mathbf{X})_2$ is observed, the best linear unbiased predictor (BLUP) of $\mathrm{vec}(\mathbf{X})_1$ given $\mathrm{vec}(\mathbf{X})_2$ is given by

$$\widehat{\mathrm{vec}(\mathbf{X})}_1 = \Sigma_{12}\Sigma_{22}^{-1}\mathrm{vec}(\mathbf{X})_2. \tag{3.20}$$

The goal here is to show that this BLUP is calculable within the set computational limits, which prevents us from naively vectorizing $\mathbf{X}$, as above, and using the matricization of $\widehat{\mathbf{C}}_{R,N}$ in place of Σ.

Assume initially that an element $\mathbf{X}$ is observed up to whole columns indexed by the set I and whole rows indexed by the set J. We can assume w.l.o.g. that $I = \{1, \ldots, m_1\}$ and $J = \{1, \ldots, m_2\}$ (otherwise we can permute the rows and columns to make it so). Denote the observed submatrix of $\mathbf{X}$ as $\mathbf{X}_{\mathrm{obs}}$. If the covariance of $\mathbf{X}$ is R-separable, so is the covariance of $\mathbf{X}_{\mathrm{obs}}$, specifically

$$\mathrm{Cov}(\mathbf{X}_{\mathrm{obs}}) = \sum_{r=1}^{R} \sigma_r \mathbf{A}_{r,22} \tilde{\otimes} \mathbf{B}_{r,22},$$

where $\mathbf{A}_{r,22}$ (resp. $\mathbf{B}_{r,22}$) are the bottom-right sub-matrices of $\mathbf{A}_r$ (resp. $\mathbf{B}_r$) of appropriate dimensions. Hence the inversion algorithm discussed above can be used to efficiently calculate $\Sigma_{22}^{-1}\mathrm{vec}(\mathbf{X})_2$ in equation (3.20). It remains to apply the cross-covariance Σ_{12} to this element. It is tedious (though possible) to write down this application explicitly using the structure of $\widehat{\mathbf{C}}_{R,N}$. Fortunately, it is not necessary, because $\mathbf{z} = \Sigma_{12}\mathbf{y}$ can be calculated as

$$\begin{pmatrix} \mathbf{z} \\ \star \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} 0 \\ \mathbf{y} \end{pmatrix}, \tag{3.21}$$

so we can apply the entire $\widehat{\mathbf{C}}_{R,N}$ to $\Sigma_{22}^{-1}\mathrm{vec}(\mathbf{X})_2$ enlarged to the appropriate dimensions by the suitable adjunction of zeros.

If an arbitrary pattern Ω in $\mathbf{X}$ is missing (i.e. Ω is a bivariate index set), we make use of the previous trick also when calculating $\Sigma_{22}^{-1}\mathrm{vec}(\mathbf{X})_2$. The PCG algorithm discussed above only requires a fast application of Σ_{22}. This can be achieved by applying the entire estimator $\widehat{\mathbf{C}}_{R,N}$ to $\widetilde{\mathbf{X}}$, where

$$\widetilde{\mathbf{X}}[i,j] = \begin{cases} \mathbf{X}[i,j] & \text{for } (i,j) \in \Omega, \\ 0 & \text{for } (i,j) \notin \Omega. \end{cases}$$

The same trick can be used to apply the cross-covariance to the solution of the inverse problem. Hence the BLUP can be calculated efficiently for an arbitrary missing pattern in $\mathbf{X}$.

3.5 Asymptotic Properties

3.5.1 Complete Observations

In this section we establish the consistency of our estimator and derive its rate of convergence. We separately consider three cases: the case of fully observed data, the case of data observed discretely on a grid, and the case of irregular observations including measurement errors. All proofs are postponed to the appendix.

In the fully observed case, we consider our sample to consist of i.i.d. observations $X_1, \ldots, X_N \sim X$, where X is a random element on $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$. Recall that our estimator $\widehat{C}_{R,N}$ given in (3.13) is the best R-separable approximation of the sample covariance matrix $\widehat{C}_N$. Under the assumption of finite fourth moment, we get the following rate of convergence for our estimator.

Theorem 7. *Let $X_1, \ldots, X_N \sim X$ be a collection of i.i.d. random elements of $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ with $\mathbb{E}(\|X\|^4) < \infty$. Denote the SCD of the covariance of X by $C = \sum_{i=1}^{\infty} \sigma_i A_i \tilde{\otimes} B_i$, with $\sigma_1 > \cdots > \sigma_R > \sigma_{R+1} \geq \cdots \geq 0$. Define*

$$\alpha_i = \min\{\sigma_{i-1}^2 - \sigma_i^2, \sigma_i^2 - \sigma_{i+1}^2\}$$

for $i = 1, \ldots, R$, and let $a_R = |||C|||_2 \sum_{i=1}^{R} (\sigma_i / \alpha_i)$. Then,

$$\left|\left|\left| \widehat{C}_{R,N} - C \right|\right|\right|_2 \leq \sqrt{\sum_{i=R+1}^{\infty} \sigma_i^2} + \mathcal{O}_{\mathbb{P}}\left(\frac{a_R}{\sqrt{N}} \right).$$

The first term $\sqrt{\sum_{i=R+1}^{\infty} \sigma_i^2}$ can be viewed as the *bias* of our estimator, which appears because we estimate an R-separable approximation of a general C. Since C is a Hilbert-Schmidt operator, this term converges to 0 as R increases. If C is actually R-separable then this term equals zero. The second term signifies the estimation error of the R-separable approximation and can be thought of as the *variance*. As is the case in PCA, the variance depends on the *spectral gap* α_i of the covariance operator (Hsing and Eubank, 2015). Note, however, that this is not the usual spectral gap, but rather the spectral gap for the sequence of squared separable component scores $(\sigma_i^2)_{i\geq 1}$. This is due to the fact that we are estimating the leading separable components of the covariance operator, rather than its principal components (Hsing and Eubank, 2015, Section 5.2).

The derived rate clearly shows the bias-variance trade-off. While the bias term is a decreasing function of R, generally the variance term (governed by a_R) is an increasing function of R. This emphasizes the need to choose an appropriate R in practice. The actual trade-off depends on the decay of the separable component scores. In particular, if the scores decay slowly, then we can estimate a relatively large number of components in the SCD, but the estimation error will be high. On the other hand, when we have a fast decay in the scores, we can estimate a rather small number of components, but with much better precision. In practice, we expect only a few scores to be significant and C to have a relatively low degree-of-separability. The theorem shows that in such situations, our estimator enjoys a convergence rate of $\mathcal{O}_{\mathbb{P}}(N^{-1/2})$, which is the same as that of the empirical estimator.

Theorem 7 shows that the convergence rate of the estimator depends on the decay of the separable component scores, and it also gives us convergence rates when we allow $R = R_N$ to increase as a function of N. For that we need to assume some structure on the decay of the separable component scores. For instance, assume that the separable component scores of C are convex, in the sense that the linearly interpolated scree plot $x \mapsto \sigma_x$ is convex, where $\sigma_x = (\lceil x \rceil - x)\sigma_{\lfloor x \rfloor} + (x - \lfloor x \rfloor)\sigma_{\lceil x \rceil}$ whenever x is not an integer (Jirak, 2016). Then it also holds that $x \mapsto \sigma_x^2$ is convex. Now, following Jirak (2016), we get that for $j > k$,

$$k\sigma_k \geq j\sigma_j \text{ and } \sigma_k - \sigma_j \geq (1 - k/j)\sigma_k.$$

Also, because of the convexity of $x \mapsto \sigma_x^2$, it follows that $\sigma_{i-1}^2 - \sigma_i^2 \geq \sigma_i^2 - \sigma_{i+1}^2$ for every i, showing

$$\alpha_i = \min\{\sigma_{i-1}^2 - \sigma_i^2, \sigma_i^2 - \sigma_{i+1}^2\} = \sigma_i^2 - \sigma_{i+1}^2.$$

So,

$$\frac{\sigma_i}{\alpha_i} = \frac{\sigma_i}{\sigma_i^2 - \sigma_{i+1}^2} = \frac{\sigma_i}{(\sigma_i - \sigma_{i+1})(\sigma_i + \sigma_{i+1})} \geq \frac{1}{2\sigma_1}\frac{\sigma_i}{(\sigma_i - \sigma_{i+1})}.$$

Now, for $0 < x < 1$, $(1-x)^{-1} > 1 + x$. Since $\sigma_i > \sigma_{i+1}$ for $i = 1, \ldots, R$, this shows that

$$\sum_{i=1}^{R} \frac{\sigma_i}{\sigma_i - \sigma_{i+1}} = \sum_{i=1}^{R} \frac{1}{1 - \frac{\sigma_{i+1}}{\sigma_i}} > \sum_{i=1}^{R} \left(1 + \frac{\sigma_{i+1}}{\sigma_i}\right) > R.$$

So, $a_R \gtrsim R$. Again, since $x \mapsto \sigma_x^2$ is convex, $\sigma_i^2 - \sigma_{i+1}^2 \geq \sigma_i^2/(i+1)$. Using this we get

$$\begin{aligned} a_R &\propto \sum_{i=1}^{R} \frac{\sigma_i}{\sigma_i^2 - \sigma_{i+1}^2} \leq \sum_{i=1}^{R} \frac{\sigma_i^2}{\sigma_i(\sigma_i^2 - \sigma_{i+1}^2)} \leq \sum_{i=1}^{R} \frac{i+1}{\sigma_i} \leq \frac{1}{R\sigma_R} \sum_{i=1}^{R} i(i+1) \\ &= \frac{1}{\sigma_R} \frac{(R+1)(R+2)}{3} \asymp \frac{R^2}{\sigma_R}, \end{aligned}$$

where we have used $i\sigma_i \geq R\sigma_R$ on the fourth step. It follows that, $a_R = \mathcal{O}(R^2/\sigma_R)$. Also, following Jirak (2016, Eq. (7.25)), we have $\sum_{i>R} \sigma_i^2 \leq (R+1)\sigma_R^2$. Thus, from Theorem 7,

$$\left\| \widehat{C}_{R,N} - C \right\|_2 = \mathcal{O}(\sqrt{R}\sigma_R) + \mathcal{O}_{\mathbb{P}}\left(\frac{R^2}{\sigma_R \sqrt{N}} \right).$$

On the other hand, $R\sigma_R \leq \sigma_1$ implies that $\sqrt{R}\sigma_R \leq \sigma_1/\sqrt{R}$. This finally shows that

$$\left\| \widehat{C}_{R,N} - C \right\|_2 = \mathcal{O}\left(\frac{1}{\sqrt{R}} \right) + \mathcal{O}_{\mathbb{P}}\left(\frac{R^2}{\sigma_R \sqrt{N}} \right).$$

So, for consistency, we need $R = R_N \to \infty$ as $N \to \infty$ while $\sigma_{R_N}^{-1} R_N^2 = o(\sqrt{N})$. The optimal rate of R is obtained by solving $\sigma_{R_N}^{-4} R_N^3 \asymp N$. Clearly, the rate of decay of the separable component scores plays an important role in determining the *admissible* and *optimal* rates of R. For instance, if the scores have an exponential decay, i.e., $\sigma_R \sim R^{-\tau}$ with $\tau > 1$, we need $R_N = o(N^{1/(2\tau+4)})$ for consistency. The optimal rate is achieved by taking $R_N \asymp N^{1/(4\tau+3)}$, which gives

$$\left\| \widehat{C}_{R,N} - C \right\|_2 = \mathcal{O}_{\mathbb{P}}\left(N^{-\frac{2\tau-1}{4\tau+3}} \right).$$

The derived rates show a trade-off between the number of estimated components and the error. While the optimal rate for R_N is a decreasing function of τ, the rate of convergence of the error is an increasing function. In particular, if the scores decay slowly (i.e., τ is close to 1), then we can estimate a relatively large number of components in the SCD, but this will likely not lead to a lower estimation error (since the scores which are cut off are still substantial). On the other hand, when we have a fast decay in the scores (i.e., τ is large), we can estimate a rather small number of components in the SCD, but with much better precision.

Similar rates can be derived assuming polynomial decay of the scores, i.e. when $\sigma_R \sim R^\tau \rho^{-R}$ with $0 < \rho < 1, \tau \in \mathbb{R}$. In this case, consistency is achieved when $R_N^{2-\tau} \rho^{R_N} = o(\sqrt{N})$, while to obtain the optimal rate, one needs to solve $R_N^{3-4\tau} \rho^{4R_N} \asymp N$. Thus, in the case of polynomial decay of the scores (which is considerably slower than the exponential decay), we cannot expect to reliably estimate more than $\log N$ many components in the SCD.

Remark 10. *The rates that we have derived are genuinely nonparametric, in the sense* *that we have not assumed any structure whatsoever on the true covariance. We have only assumed that X has finite fourth moment, which is standard in the literature for covariance estimation. We can further relax that condition if we assume that C is actually R-separable. From the proof of the theorem, it is easy to see that if we assume* $\mathrm{DoS(C)} \le \mathrm{R}$*, then our estimator is consistent under the very mild condition of* $\mathbb{E}(\|X\|^2) < \infty$*.*

3.5.2 Discrete Observations on a Grid

In practice, the data are observed and manipulated discretely, and we now develop asymptotic theory in that context. Specifically, assume that $X = \{X(t,s) : t \in \mathcal{T}, s \in \mathcal{S}\}$ is a random field taking values in $\mathcal{H} = \mathcal{L}_2(\mathcal{T} \times \mathcal{S})$, where $\mathcal{T}$ and $\mathcal{S}$ are compact sets. To simplify notation, we assume w.l.o.g. that $\mathcal{T} = \mathcal{S} = [0,1]$.

We observe the data at $K_1 K_2$ regular grid points or *pixels*. Let $\{T_1^{K_1}, \dots, T_{K_1}^{K_1}\}$ and $\{S_1^{K_2}, \dots, S_{K_2}^{K_2}\}$ denote regular partitions of $[0,1]$ of lengths $1/K_1$ and $1/K_2$, respectively. We denote by $I_{i,j}^K = T_i^{K_1} \times S_j^{K_2}$ the (i,j)-th pixel for $i = 1, \dots, K_1, j = 1, \dots, K_2$. The pixels are non-overlapping (i.e., $I_{i,j}^K \cap I_{i',j'}^K = \emptyset$ for $(i,j) \neq (i',j')$) and have the same volume $|I_{i,j}^K| = 1/(K_1 K_2)$. For each surface $X_n, n = 1, \dots, N$, we make one measurement at each of the pixels. Note that we can represent the measurements for the n-th surface by a matrix $\mathbf{X}_n^K \in \mathbb{R}^{K_1 \times K_2}$, where $\mathbf{X}_n^K[i,j]$ is the measurement at $I_{i,j}^K$ for $i = 1, \dots, K_1, j = 1, \dots, K_2$.

We consider again two different sampling schemes, which relate the latent surfaces $X_1, \dots, X_N$ to the discrete observations $\mathbf{X}_1^K, \dots, \mathbf{X}_N^K$.

(S1) Pointwise evaluation within each pixel, i.e.,

$$\mathbf{X}_n^K[i,j] = X_n(t_i^{K_1}, s_j^{K_2}), i = 1, \dots, K_1, j = 1, \dots, K_2,$$

where $(t_i^{K_1}, s_j^{K_2}) \in I_{i,j}^K$ are spatio-temporal locations. Note that the square integrability of X is not sufficient for such pointwise evaluations to be meaningful, so we will assume that X has continuous sample paths in this case.

(S2) Averaged measurements over each pixel, i.e.,

$$\mathbf{X}_n^K[i,j] = \frac{1}{|I_{i,j}^K|} \iint_{I_{i,j}^K} X_n(t,s) dt ds, i = 1, \dots, K_1, j = 1, \dots, K_2.$$

In both scenarios, we denote by X_n^K the *pixelated version* of X_n,

$$X_n^K(t,s) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \mathbf{X}_n^K[i,j] \, \mathbf{1}\{(t,s) \in I_{i,j}^K\}, n = 1, \dots, N.$$

The corresponding pixelated version of X is denoted by X^K. Under our assumptions, X^K is zero-mean with covariance $C^K = \mathbb{E}(X^K \otimes X^K)$. It can be easily verified that under the pointwise measurement scheme (S1), C^K is the integral operator with kernel

$$c^K(t,s,t',s') = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} c(t_i^{K_1}, s_j^{K_2}, t_k^{K_1}, s_l^{K_2})\,\mathbf{1}\{(t,s)\in I_{i,j}^K, (t',s')\in I_{k,l}^K\}.$$

For the averaged measurement scheme (S2), we can represent X^K as

$$X^K = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \left\langle X, g_{i,j}^K \right\rangle g_{i,j}^K,$$

where $g_{i,j}^K(t,s) = |I_{i,j}^K|^{-1/2}\mathbf{1}\{(t,s)\in I_{i,j}^K\}$, from which it immediately follows that

$$C^K = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} \left\langle C, g_{i,j}^K \otimes g_{k,l}^K \right\rangle g_{i,j}^K \otimes g_{k,l}^K.$$

In the discrete observation scenario, our estimator is the best R-separable approximation of $\widehat{C}_N^K$, the empirical covariance of $X_1^K, \ldots, X_N^K$. Note that $\widehat{C}_N^K$ is the pixel-wise continuation of $\widehat{\mathbf{C}}_N^K$, the empirical covariance of $\mathbf{X}_1^K, \ldots, \mathbf{X}_N^K$. To derive the rate of convergence in this scenario, we need some *continuity* assumption relating the random field X and its pixelated version X^K. The following theorem gives the rate of convergence of the estimator when the true covariance is *Lipschitz continuous.*

Theorem 8. *Let $X_1, \ldots, X_N \sim X$ be a collection of random surfaces on $[0,1]^2$, where the covariance of X has SCD $C = \sum_{i=1}^\infty \sigma_i A_i \tilde{\otimes} B_i$, with $\sigma_1 > \cdots > \sigma_R > \sigma_{R+1} \geq \cdots \geq 0$. Further assume that the kernel $c(t,s,t',s')$ of C is L-Lipschitz continuous on $[0,1]^4$. Suppose that one of the following holds.*

1. *X has almost surely continuous sample paths and $\mathbf{X}_1^K, \ldots, \mathbf{X}_N^K$ are obtained from $X_1, \ldots, X_N$ under the measurement scheme (S1).*
2. *$\mathbf{X}_1^K, \ldots, \mathbf{X}_N^K$ are obtained from $X_1, \ldots, X_N$ under the measurement scheme (S2).*

If $\mathbb{E}(\|X\|_2^4) < \infty$, then

$$\left|\!\left|\!\left| \widehat{C}_{R,N}^K - C \right|\!\right|\!\right|_2 = \sqrt{\sum_{i=R+1}^\infty \sigma_i^2} + \mathcal{O}_{\mathbb{P}}\left(\frac{a_R}{\sqrt{N}}\right) + \left(16a_R + \sqrt{2}\right)L\sqrt{\frac{1}{K_1^2} + \frac{1}{K_2^2}} + \frac{8\sqrt{2}L^2}{|\!|\!|C|\!|\!|_2}\left(\frac{1}{K_1^2} + \frac{1}{K_2^2}\right)a_R,$$

where the $\mathcal{O}_{\mathbb{P}}$ term is uniform in K_1, K_2 and $a_R = |\!|\!|C|\!|\!|_2 \sum_{i=1}^R (\sigma_i/\alpha_i)$ is as in Theorem 7.

The theorem shows that in the case of discretely observed data, we get the same rate of convergence as in the fully observed case, plus additional terms reflecting the estimation error of R components at finite resolution. We assumed Lipschitz continuity to quantifiably control those error terms and derive rates of convergence, but the condition is not necessary if we merely seek consistency, which can be established assuming continuity alone.

3.5.3 Unbalanced Design

The method that we have described so far is suitable for surfaces observed on the same regular grid. However, it is possible to have irregular and uneven observations. Next, we describe a way to adapt our proposed method to this setup. Suppose that for the n-th random surface $X_n = (X_n(t,s) : t \in T, s \in S)$, we make K_n measurements at bivariate locations $(t_1, s_1), \dots, (t_{K_n}, s_{K_n})$. To fix ideas, we consider the *point-wise evaluation with additive noise* model

$$Y_{ni} = X_n(t_{ni}, s_{ni}) + E_{ni}, \qquad i = 1, \dots, K_n, n = 1, \dots, N, \tag{3.22}$$

where E_{ni}'s are i.i.d. with mean zero and variance σ^2, independent of X_n. To tackle this situation, we take the classical approach, where we first smooth the observations and then apply our method on the smoothed surfaces (e.g., Ramsay and Silverman, 2005, 2007). To be precise, based on the measurements for the n-th surface, first we construct its *smoothed* or functional version

$$\widetilde{X}_n = \mathcal{S}(Y_{n1}, \dots, Y_{nK_n}), \qquad n = 1, \dots, N, \tag{3.23}$$

where $\mathcal{S}$ is a *smoothing operator* that takes as input the discrete measurements and outputs a function. Then, we apply our methodology to the surfaces $\widetilde{X}_1, \dots, \widetilde{X}_N$, i.e., use the estimator

$$\widetilde{C}_{R,N} = \arg\min_G \left|\!\left|\!\left| \widetilde{C}_N - G \right|\!\right|\!\right|_2^2 \quad \text{s.t.} \quad \mathrm{DoS}(G) \le R,$$

where $\widetilde{C}_N$ is the empirical covariance based on $\widetilde{X}_1, \dots, \widetilde{X}_N$. The rate of convergence of this estimator depends crucially on the smoother, as shown in the following theorem.

Theorem 9. *Let $X_1, \dots, X_N \sim X$ be a collection of i.i.d. random elements of $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ with $\mathbb{E}(\|X\|^4) < \infty$, and the covariance of X has separable expansion $C = \sum_{i=1}^{\infty} \sigma_i A_i \tilde{\otimes} B_i$, with $\sigma_1 > \dots > \sigma_R > \sigma_{R+1} \ge \dots \ge 0$. Suppose that the surfaces are observed discretely as* (3.22) *and let $\widetilde{X}_n$ denote the smoothed surface from* (3.23). *Let $\widehat{C}_N$ and $\widetilde{C}_N$ be the empirical covariance operators based on $X_1, \dots, X_N$ and $\widetilde{X}_1, \dots, \widetilde{X}_N$,* *respectively, and suppose that* $\left|\!\left|\!\left| \widehat{C}_N - \widetilde{C}_N \right|\!\right|\!\right|_2 = \mathcal{O}_{\mathbb{P}}(b_N)$. *Then,*

$$\left|\!\left|\!\left| \widetilde{C}_{R,N} - C \right|\!\right|\!\right|_2 \leq \left(\sum_{i=R+1}^{\infty} \sigma_i^2 \right)^{1/2} + \mathcal{O}_{\mathbb{P}}\left(a_R b_N \right),$$

where $a_R = |||C|||_2 \sum_{i=1}^{R} (\sigma_i/\alpha_i)$ *is as in Theorem 7.*

The theorem shows that the rate of convergence of the estimator in this case remains similar to the case of fully observed surfaces except $N^{-1/2}$ is replaced by b_N, the rate of convergence of $\left|\!\left|\!\left| \widehat{C}_N - \widetilde{C}_N \right|\!\right|\!\right|_2$. Note that b_N depends on the interplay between (i) the denseness of the measurements (via $K_1, \ldots, K_N$), (ii) the degree of noise (via the noise variance σ^2), (iii) the smoother used, and (iv) the smoothness of the underlying surfaces $X_1, \ldots, X_N$, and in certain scenarios is equal to the optimal rate of $N^{-1/2}$ (e.g., Hall and Hosseini-Nasab, 2006, Theorem 3).

In practice, one does not construct the whole functions $\widetilde{X}_1, \ldots, \widetilde{X}_N$, but rather evaluate them at a fixed number of points. It is possible to develop the asymptotic properties of our estimator in this setup similar to Theorem 8. But we avoid doing that for the sake of brevity.

3.6 Simulation Study

In this section, we examine the finite-sample behavior of the R-separable estimator. Section 3.6.1 discusses in detail the parametric covariance used in Example 5, displaying the bias-variance trade-off controlled by the choice of R both in the case of estimation and in the case of prediction, which can be calculated fast using the inversion algorithm of Section 3.4. Section 3.6.2 focusses on a non-parametric covariance, constructed in a way such that it is substantially non-separable, leading to more significant reductions in estimation and prediction errors already for smaller sample sizes. Section 3.6.3 aims to demonstrate the practical consequences of our theory, in particular of Theorem 8. This is achieved by introducing a non-parametric covariance allowing for a perfect control of the separable component scores. Finally, Section 3.6.4 probes in detail performance of the proposed inversion algorithm. It shows that the number of iterations of the inversion algorithm does not increase with increasing grid size, when adequate regularization is used. It also demonstrates the effectiveness of the chosen preconditioning, showing that the number of iterations is smaller when the leading term in the separable component decomposition is more dominant, i.e. when the covariance is closer to being separable.

3.6.1 Parametric Covariance

We first explore the behavior of the proposed methodology in the parametric covariance setting of Example 5. The kernel is given by

$$c(t,s,t',s') = \frac{\sigma^2}{(a^2|t-t'|^{2\alpha}+1)^{\tau}} \exp\left(\frac{b^2|s-s'|^{2\gamma}}{(a^2|t-t'|^{2\alpha}+1)^{\beta\gamma}} \right).$$

This covariance was introduced by Gneiting (2002). Among the various parameters, a, resp. b, control the temporal, resp. spatial, domain scaling, and $\beta \in [0,1]$ controls the departure from separability with $\beta = 0$ corresponding to a separable model. We fix $\beta = 0.7$, which seems to be as a rather high degree of non-separability given the range of β, but one should note that non-separability is rather small for this model regardless of the choice of β (Genton, 2007). The remaining parameters are set $\alpha = \gamma = \sigma^2 = \tau = 1$, and the scaling parameters are set as $a = b = 20$ on the $[0,1]$ interval, which corresponds to considering the domain as $[0,20]$ interval with $a = b = 1$. We stretch the domain this way in order to strengthen the non-separability of the model. Even though the covariance is stationary, this fact is completely ignored. None of the methods presented in this paper make use of stationarity, and the reported results are not affected by it.

We discretize this kernel as before to obtain the ground truth $\mathbf{C} \in \mathbb{R}^{K\times K\times K\times K}$, whose eigenvalues and separable component scores are plotted in Figure 3.1. Note that $\mathbf{C}$ is *not* R-separable for any R, but it is well-aproximated by R-separable cut-offs for small values of R already.

We fix $K = 50$ and generate N observations $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbb{R}^{K\times K}$ as independent zero-mean matrix-variate Gaussians with covariance $\mathbf{C}$. Then we fit the estimator $\widehat{\mathbf{C}}_{R,N}$ using the data and calculate the relative Frobenius error defined as

$$\|\widehat{\mathbf{C}}_{R,N} - \mathbf{C}\|_F \Big/ \|\mathbf{C}\|_F. \tag{3.24}$$

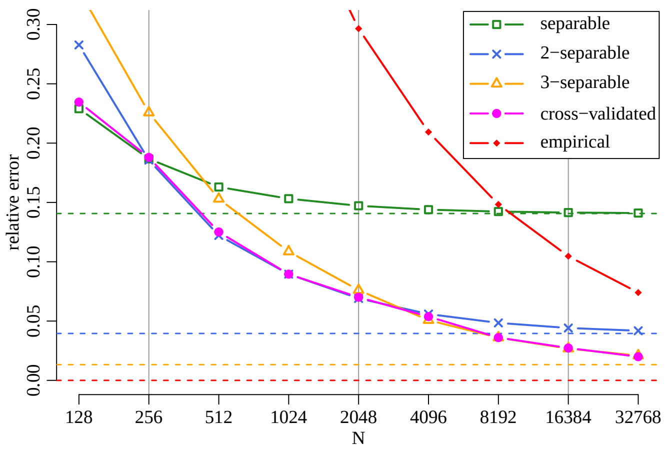
This is done for different values of R and N, and the reported results are averages over one hundred independent simulation runs.

Figure 3.2 shows how the relative error evolves as a function of N for a few fixed values of R. According to Theorem 7, the relative error converges as $N \to \infty$ to

$$\sqrt{\sum_{r=R+1}^{\infty} \sigma_r^2} \Big/ \sqrt{\sum_{r=1}^{\infty} \sigma_r^2}, \tag{3.25}$$

which can be seen as the *bias*. This is the minimal achievable error by means of an R-separable approximation, even if we knew $\mathbf{C}$, and it is depicted by a dashed horizontal line (an asymptote) in Figure 3.2 for every considered R. As expected, for $R = 1$ the relative error converges fast to its asymptote (i.e. the *variance* converges to zero), which is

Figure 3.2: Error for different values of the degree-of-separability R and for the empirical covariance $\widehat{\mathbf{C}}_N$ decreases with different speed and approaches different asymptotes. Dashed horizontal lines show the asymptotes for error curves of corresponding color. Grey vertical lines depict the sample sizes for which average scree plots are shown in Figure 3.3. When the degree-of-separability R is automatically chosen via the CV scheme from Section 3.3.1, the resulting error curve forms a lower envelope of the error curves with fixed degrees-of-separability.



higher than the asymptotes for higher values of R (higher values of R introduce a smaller *bias*). However, the speed of convergence to these lower asymptotes is substantially slower (variance goes to zero more slowly). Hence a higher choice of R does not necessarily lead to a smaller error. For example, for $N = 256$, the choice $R = 2$ is optimal (being only slightly better than a separable model with $R = 1$), while choosing $R = 3$ is in fact worse than choosing the separable model. The only unbiased estimator we consider is the empirical estimator $\widehat{\mathbf{C}}_N$, which is substantially worse than any of the R-separable estimators. Even though $\widehat{\mathbf{C}}_N$ is the only estimator among those considered, whose error will eventually converge to 0 for $N \to \infty$, Figure 3.2 shows that N would have to be extremely large for the empirical estimator to beat the R-separable estimators with reasonably chosen degree-of-separability R.

In practice, we naturally do not know $\mathbf{C}$, and hence cannot choose R to optimally balance bias and variance as a visual inspection of Figure 3.2 might allow one to. Instead, we need to use the cross-validation strategy described in Section 3.3.1. Figure 3.2 also shows the relative errors achieved with a cross-validated choice of R. Cross-validation appears to work rather well, with the cross-validated error curve forming almost a lower envelope of the curves for $R = 1, 2, 3$, i.e. leading to an error that is always near optimal.

For very large problems, where CV may become prohibitive, one may prefer to visually inspect the estimated separable component scores and decide a suitable degree-of-separability by hand based on a scree plot. Hence we show in Figure 3.3 such scree plots for 3 different values of N (those depicted by grey vertical lines in Figure 3.2; we encourage the reader to compare the two plots). For $N = 256$, we would likely pick $R = 1$

Figure 3.3: Scree plots for 3 different values of N averaged over the 10 independent simulation runs.

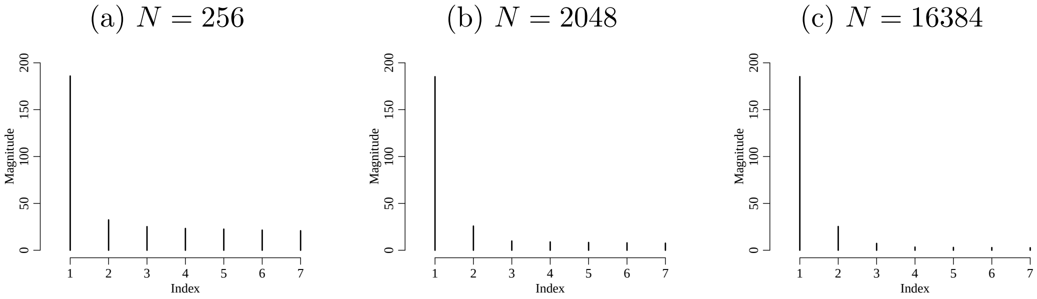


Table 3.4: Relative prediction error in the parametric covariance setup for the R-separable estimators ($R = 1, 2, 3$) and the empirical covariance. Minima for a fixed N (given column) are depicted in bold.

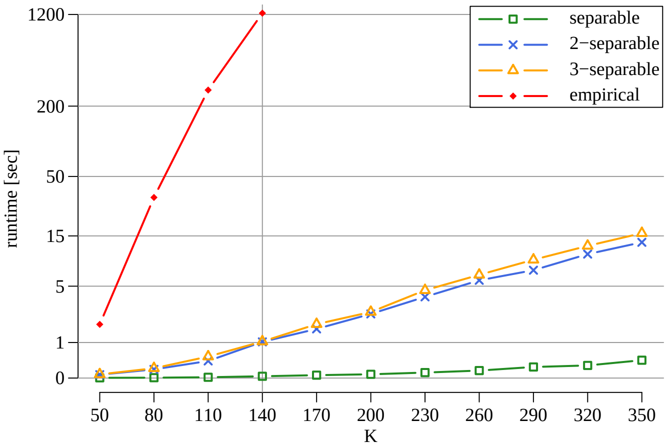
N	128	256	512	1024	2048	4096	8192	16384	32768
$R = 1$	**0.356**	0.354	0.354	0.353	0.349	0.354	0.350	0.352	0.353
$R = 2$	0.363	**0.349**	**0.345**	0.342	0.337	0.341	0.338	0.340	0.340
$R = 3$	0.422	0.352	0.346	**0.341**	**0.335**	**0.338**	**0.335**	**0.336**	**0.335**
EMP	1.012	0.970	0.866	0.674	0.473	0.396	0.361	0.348	0.342

or $R = 2$ based on Figure 3.3 (a), these two choices leading to roughly the same errors. For $N = 2048$, one would likely choose $R = 2$, leading to the optimal error. Finally, in the case of Figure 3.3 (c), one would likely choose $R = 3$ based on the visual inspection and comparing the sample size N to the grid size K, leading to the optimal error.

When the task is prediction rather than estimation, we can also benefit from including terms beyond separability. With the parametric ground-truth considered here, the gains are rather small, hence we show different prediction errors in form of a table, see Table 3.4. The prediction errors are calculated from additional test samples, after the training sample (constructed as above) was used to estimate the covariance. The final row and the final column of every observation in the test sample (of size 100) are predicted using the fitted covariance and the remainder of the given observation, i.e. we perform one-step ahead prediction both in space and time at the same time. We use a small amount of ridge regularization for all the competing methods in this chapter.

While there are only small differences between the relative prediction errors, there are notable differences in runtime. In particular, prediction with the empirical covariance estimator is computationally demanding, cf. Table 3.2. To demonstrate this in practice, Figure 3.4 shows the runtime of a single prediction task, run on a personal laptop with Windows 10 (64-bit) operating system, Intel Core i7-7700HQ (2.8 GHz) processor, 16 GB RAM, and R version 3.6.3 (R Core Team, 2020). The memory complexity of constructing the empirical covariance restricts us to modest grid size (up to $K = 140$, we ran out of

Figure 3.4: Runtimes of calculating a single prediction error (i.e. performing a single prediction task after the covariance has been already estimated) with a separable covariance ($R = 1$), R-separable covariance with $R = 2$ and $R = 3$, and an unstructured (here the empirical) covariance.



memory at $K = 170$). Moreover, the runtimes explode for the empirical covariance. At the edge of feasibility ($K = 140$), our inversion algorithm with $R = 3$ runs about 1200 times faster compared to the empirical covariance, while it runs only 29 times slower compared to the separable model. It will be shown empirically in Section 3.6.4 that the number of iterations of our inversion algorithm does not depend on the grid size and hence the theoretical complexity of prediction is the same for the separable model as for the R-separable model. Nonetheless, the cost in practice is slightly higher due to the iterative nature and some overhead calculations (e.g. the preconditioning).

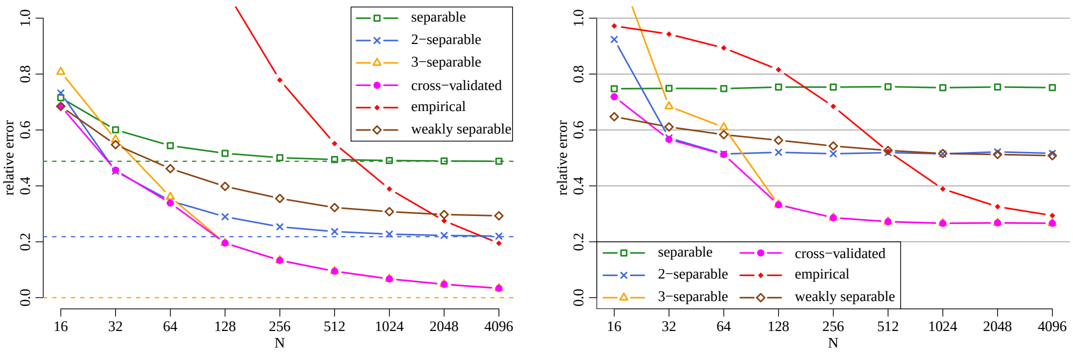
3.6.2 Weakly Separable Covariance

Here we consider a non-parametric ground-truth, constructed as follows. Let $\{\phi_1, \ldots, \phi_{50}\}$ denote the first fifty functions of the trigonometric basis, and let $J_1 = \{1, 4, \ldots, 49\}$, $J_2 = \{2, 5, \ldots, 50\}$ and $J_3 = \{3, 6, \ldots, 48\}$ be three index sets. Covariances $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ are constructed to have norm one, power decay of the eigenvalues, and the trigonometric system as their eigenbasis. However, the leading eigenfunctions in $\mathbf{A}_r$ are those trigonometric functions ϕ_l with indices in J_r, $r = 1, \ldots, 3$. $\mathbf{B}_1$, $\mathbf{B}_2$ and $\mathbf{B}_3$ are chosen in the same way, and the resulting covariance is chosen as

$$\mathbf{C} = \sum_{r=1}^{3} \sigma_r \mathbf{A}_r \tilde{\otimes} \mathbf{B}_r$$

with $\sigma_1 = 8$, $\sigma_2 = 4$ and $\sigma_3 = 2$. Note that these are not the separable component scores, since $\mathbf{A}_r$'s (as well as $\mathbf{B}_r$'s) are not orthogonal, and the same is true for $\mathbf{B}_r$'s. In other words, the previous equation is *not* a separable component decomposition. Still, the covariance is 3-separable, i.e. it is a superposition of three separable terms. Moreover,

Figure 3.5: Relative estimation (**left**) and prediction (**right**) errors depending on sample size N depending on sample size N. Considered estimators are the separable estimator ($R = 1$), R-separable estimators with $R = 2$ and $R = 3$, R-separable estimator with cross-validated R, the weakly separable estimator, and the empirical covariance estimator. Straight horizontal lines (asymptotes) show the bias of R-separable estimators for $R = 1, 2, 3$.



all the separable terms have the same eigenfunctions, which are outer products of the trigonometric functions. Hence the covariance is by construction weakly separable (Lynch and Chen, 2018).

We compare different estimators just as in the previous section, with the addition of the weakly separable estimator of Lynch and Chen (2018). Since the codes for the weakly separable estimator are only available in Matlab, we use our own R implementation. A weakly separable estimator is obtained in two steps. The product eigenbasis is estimated from the data via partial tracing of Aston et al. (2017), and the eigenvalues are estimated in the subsequent step. Lynch and Chen (2018) propose to enforce low-rankness by only retaining a part of the basis, such that at least 95% of variance is explained in both dimensions. We follow this suggestion.

Figure 3.5 shows the relative estimation and prediction errors achieved by different estimators. The results are qualitatively similar to those in the previous section. Reductions in prediction errors achieved by considering an R-separable model with $R > 1$ are more profound here, while the runtimes are virtually the same to those reported in Figure 3.4. Compared to the previous section, the ground-truth covariance here is by construction weakly separable, and hence we also compare our methodology against the weakly separable model of Lynch and Chen (2018). As displayed in Figures 3.5, the weakly separable model does better than the separable model or the empirical estimator, but is outperformed eventually by the proposed R-separable model with a suitably chosen $R > 1$, e.g. by cross-validation.

3.6.3 Superposition of Independent Separable Processes

In this section, we consider randomly generated 4-separable covariances, i.e.

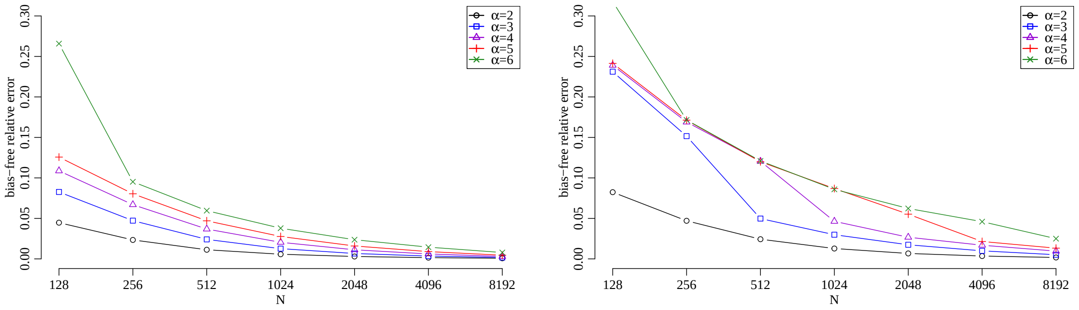
$$\mathbf{C} = \sum_{r=1}^{4} \sigma_r \mathbf{A}_r \tilde{\otimes} \mathbf{B}_r. \tag{3.26}$$

Our goal is to show how the constant a_R of Theorem 7, which is related to the decay of the separable component scores, affects the convergence speed, i.e. the "variance" part of the error. To this end, we generate a random orthonormal basis, and split this basis into four sets of vectors, say $\mathcal{D}_1, \ldots, \mathcal{D}_4$. To generate $\mathbf{A}_r$ for $r = 1, \ldots, 4$, we use the vectors from $\mathcal{D}_r$ as the eigenvectors, while the non-zero eigenvalues are set as $|\mathcal{D}_r|, \ldots, 1$. Hence $\mathbf{A}_r$ is singular with random eigenvectors and the eigenvalues, which are non-zero, are linearly decaying. The procedure is the same for $\mathbf{B}_1, \ldots, \mathbf{B}_4$. Note that the exact form of the covariances $\mathbf{A}$ and $\mathbf{B}$ is quite arbitrary and is not affecting the results heavily. However, it is not easy to come up with a generation procedure for $\mathbf{C}$, which allows for a perfect control of its separable component scores. One basically needs $\mathbf{A}_1, \ldots, \mathbf{A}_4$ to be positive semi-definite and orthogonal at the same time. And the only way to achieve this is to have $\mathbf{A}_1, \ldots, \mathbf{A}_4$ low-rank with orthogonal eigenspaces. If this was not the case, equation (3.26) would not define a separable component decomposition, and σ_r's (which we are going to choose) would not be the separable component scores. We choose $\sigma_r = \alpha^{R-r}$, $r = 1, \ldots, R$, for different values of α. Hence we have different polynomial decays for the scores. Higher α's correspond to faster decays and consequently to a higher value of a_R from Theorem 7. Thus we expect a slower convergence for higher values of α.

Figure 3.6 shows bias-free relative estimation errors for $\widehat{R} = 2$ and $\widehat{R} = 3$ (i.e. for a wrongly chosen degree-of-separability, since the truth is $R = 4$). The sample size N is varied and the grid size is fixed again as $K = 50$. To be able to visually compare the speed of convergence, we removed the bias (3.25) from the relative estimation errors. For example, for $\widehat{R} = 3$, the bias is proportional to σ_4 (note that $\mathbf{C}$ is standardized to have norm equal to one). However, σ_4 varies for different α's, so we opt to remove σ_4 from the error corresponding to all the α's, in order for the curves in Figure 3.6 to depict only the variance converging to zero. As expected, the convergence is faster for smaller α's corresponding to a slower decay of the separable component scores.

One can also notice certain transitions in Figure 3.6. For $\widehat{R} = 2$ and $\alpha = 6$, the drop in error between sample sizes $N = 128$ and $N = 256$ clearly stands out in the figure. This is because when $\alpha = 6$, the scores decay so rapidly that the sample size $N = 128$ is not enough for the second separable component to be estimated reliably. The situation is similar to Figure 3.2, but since here the bias is subtracted from the error, choosing a higher $\widehat{R}$ than we can afford to estimate is even more striking. A similar behavior can be observed for multiple curves in Figure 3.6 (right), when $\widehat{R} = 3$. For example, one can observe an "elbow" at $N = 512$ for the relatively slow decay of $\alpha = 3$. This "elbow" is

Figure 3.6: Relative estimation error curves for $\widehat{R} = 2$ (**left**) and $\widehat{R} = 3$ (**right**) when the true degree-of-separability is $R = 4$. The reported errors are bias-free; bias was subtracted to make apparent the different speed of convergence of the variance to zero.



present because for smaller sample sizes, a smaller value of $\widehat{R}$ would have been better. From $N = 512$ onwards, all 3 separable components are estimated reliably and the variance decays rather slowly and smoothly. This “elbow” exists for $\alpha = 4, 5$ as well, but manifests “later" in terms of N. Finally, for $\alpha = 6$, one can actually observe 3 different modes of convergence in Figure 3.6 (b): before $N = 256$, the degree-of-separability is overestimated by 2; between $N = 256$ and $N = 4096$, it is overestimated by 1; and it seems that for a larger N, the curve would finally enter the slowly converging mode.

3.6.4 Random Covariance

In this section, we are interested in the numerical performance of the inversion algorithm of Section 3.4. To emulate more realistic inversion problems, we first simulate data with the ground truth $\mathbf{C}$ specified below, then find an R-separable estimate $\widehat{\mathbf{C}}_{R,N}$, construct its positivized version $\widehat{\mathbf{C}}^{+}_{R,N}$, calculate $\mathbf{Y} = \widehat{\mathbf{C}}^{+}_{R,N}\mathbf{X}$ for a randomly generated $\mathbf{X} \in \mathbb{R}^{K \times K}$, and then use the inversion algorithm to recover $\mathbf{X}$ from the knowledge of $\widehat{\mathbf{C}}^{+}_{R,N}$ and $\mathbf{Y}$. The number of observations and the degree-of-separability are fixed now as $N = 500$ and $R = 5$ (both true and used for estimation), while the grid size K varies.

The ground truth covariance is given as as

$$\mathbf{C} = \sum_{r=1}^{R} \sigma_r \mathbf{A}_r \,\tilde{\otimes}\, \mathbf{B}_r.$$

Since we do not require a special control of the separable component scores, we have complete freedom in the choice of $\mathbf{A}_r$'s, $\mathbf{B}_r$'s and σ_r's. We set $\mathbf{A}_1$ and $\mathbf{B}_1$ both as the covariance of Brownian motion, standardized to have Hilbert-Schmidt norm equal to one. Since we keep all the σ's equal to one, ordering of the covariance is immaterial. For a fixed $r = 2, \dots, R$, we generate $\mathbf{A}_r$ as follows (the procedure for $\mathbf{B}_r$ is again the same). We have a pre-specified list of functions, including polynomials of low order, trigonometric functions, and a B-spline basis. We choose a random number of these functions (evaluated on a grid), complement them with random vectors to span the whole space, and orthogonalize this collection to obtain an eigenbasis. The eigenvalues are chosen of to have a power decay with a randomly selected base. This procedure leads to a visually smooth covariances, which cannot be orthogonal to each other, but their collinearity varies quite randomly. As a consequence, we have R-separable covariance with some random decay of the separable component scores.

Figure 3.7 (left) shows how the number of iterations required by the PCG inversion algorithm evolves as the grid size K increases for different values of the regularizer ϵ used to positivize the estimator, see (3.19), leading to three fixed condition numbers $\kappa = 10, 10^2, 10^3$. The results are again averages over one hundred independent Monte Carlo runs. For a fixed condition number κ, we always find ϵ such that the condition number of $\mathbf{C}^{+}_{R,N}$ is exactly κ. We want to control the condition number of the left-hand side matrix because it generally captures the difficulty of the inversion problem (Van Loan and Golub, 1983). In the case of PCG, the number of iterations is expected to grow roughly as the square-root of the condition number. As seen in Figure 3.7 (left), the number of iterations needed for convergence depends on the condition number in the expected manner, while the grid size K does not affect the required number of iterations. This fact allows us to claim that the computational complexity of the inversion algorithm is $\mathcal{O}(K^3)$, i.e. the same as for the separable model.

Remark 11. *We ran the PCG algorithm with a relatively stringent tolerance* 10^{-10} *(i.e. stopping the algorithm only when two subsequent iterates are closer than* 10^{-10} *in the Frobenius norm). The maximum recovery error across all simulation runs and all setups of the parameters was* $3 \cdot 10^{-10}$*. Hence there is no doubt that the inversion algorithm performs as intended.*

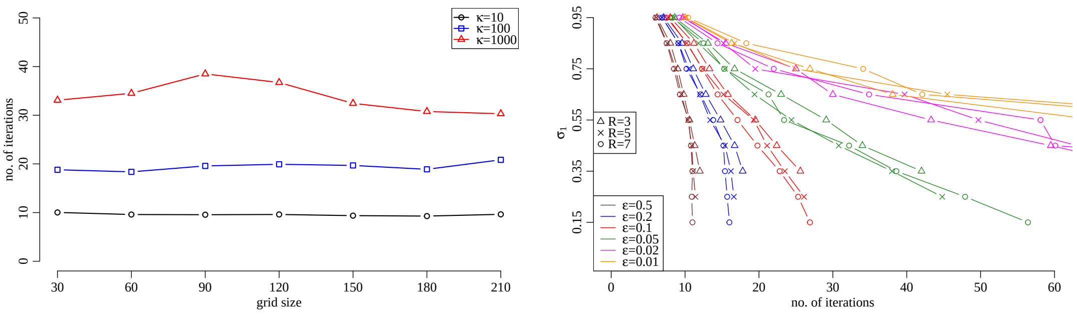
Finally, we explore the claim that a nearly separable model leads to milder computational costs than a highly non-separable model. We take $\widehat{\mathbf{C}}_{R,N}$ estimated with $N = 500$ as above with different values of $R = 3, 5, 7$, and we change its scores $\widehat{\sigma}_r$, $r = 1, \ldots, R$. Firstly, we fix $\widetilde{\sigma}_1 \in \{0.15, 0.25, \ldots, 0.95\} \cap \{\sigma; \sigma \geq 1/R\}$. Then, we generate $\widetilde{\sigma}_r$ for $r = 2, \ldots, R$ as a random variable uniformly distributed on the interval

$$\left(\max\left(0, 1 - \sum_{j=1}^{r-1} \widetilde{\sigma}_j - (R-r)\widetilde{\sigma}_1 \right), \min\left(\widetilde{\sigma}_1, 1 - \sum_{j=1}^{r-1} \widetilde{\sigma}_j \right) \right).$$

This leads to a collection of scores which are smaller than or equal to $\widetilde{\sigma}_1$ and they sum up (together with $\widetilde{\sigma}_1$) to one. Lastly, we set

$$\widetilde{\mathbf{C}}^{\epsilon}_{R,N} = \sum_{r=1}^{R} \widetilde{\sigma}_r \widehat{\mathbf{A}}_r \widetilde{\otimes} \widehat{\mathbf{B}}_r + \epsilon \mathbf{I}.$$

Figure 3.7: Left: Number of iterations required by the PCG algorithm does not increase with increasing grid size K when the condition number κ is held fixed, $\kappa \in \{10, 10^2, 10^3\}$. **Right:** The smaller the leading score $\widetilde{\sigma}_1$ is (relatively to other scores), the higher number of iterations is required by the PCG algorithm. Different regularization constants ϵ are distinguished by different colors, while different degrees-of-separability are distinguished by different symbols. Smaller degrees-of-separability prevent σ_1 from being too small, leading to naturally shorter curves, but otherwise the degree-of-separability is not affecting the results very much.



This time, we do not look for ϵ in a way such that the condition number of $\widetilde{\mathbf{C}}^{\epsilon}_{R,N}$ is fixed, because we want to explore how the size of $\widetilde{\sigma}_1$ affects the number of iterations. However, a part of this effect is how $\widetilde{\sigma}_1$ affects the condition number, so we just standardize with several different (but fixed) values of ϵ. Finally, we generate a random $\mathbf{X} \in \mathbb{R}^{K \times K}$, calculate $\mathbf{Y} = \widetilde{\mathbf{C}}^{\epsilon}_{R,N}\mathbf{X}$, and use the inversion algorithm to recover $\mathbf{X}$ from the knowledge of $\widetilde{\mathbf{C}}^{\epsilon}_{R,N}$ and $\mathbf{Y}$. Remark 11 applies here as well.

The results are plotted in Figure 3.7 (right). As expected, the better regularized problems with a larger ϵ generally require a smaller number of iterations. But more importantly, the number of iterations increases with decreasing $\widetilde{\sigma}_1$. This effect is milder for large ϵ, but more severe for smaller regularization constants. This means that unless $\mathbf{C}$ is well-posed (i.e. with a relatively large smallest eigenvalue $\lambda_{\min}$), we have to pay extra costs for very substantial departures from separability (i.e. when the largest score σ_1 is not much larger than the other scores).

3.7 Data Analysis: Classification of EEG Signals

Electroencephalography (EEG) is dominantly utilized for screening and diagnosis of various mental disorders, such as epilepsy or autism. It is an effective monitoring procedure for studying various brain activities. Since the responsive capacity of the brain is severely affected by alcoholism, EEG can also be used for detection and diagnosis of alcoholism. EEG signals are acquired using numerous electrodes placed on the scalp, and each electrode produces a time series of measurements, sampled over a specific interval. Since the time series produced by EEG electrodes are known to be non-stationary, it is natural to consider a block of measurements as a random surface, with time being one dimension and space (electrode placement) being the second dimension. We consider detection of alcoholism as a classification problem based on random surfaces, which are event-related EEG signals. We work with a data set of 77 alcoholic and 45 control subjects, which is freely available from University of California, Irvine machine learning database[1]. Each subject was repeatedly exposed to either a single stimulus (Condition 1) or two matching stimuli (Condition 2) or two non-matching stimuli (Condition 3). There was a total of 120 of these 1-second-long trials sampled at 256 Hz, but many were discarded right after the acquisition due to artifacts such as blinking. We discarded one of the control-group subjects, since this was a clear outlier with only 19 successful trials. After that, we have calculated averages of 10 (the maximum possible number for the remaining subjects) random available trials per condition, resulting in a data set $X_1, \ldots, X_{121} \in \mathbb{R}^{256 \times 64}$ of $N = 121$ subjects, $K_1 = 256$ time points and $K_2 = 64$ spatial locations for each of the three conditions. Also, the class membership variables $Y_1, \ldots, Y_{121} \in \{0, 1\}$ (control or alcoholic) for all the subjects are available.

Classification of subjects into their respective classes (control versus alcoholic) using the EEG data set described above has been conducted many times before – a search through Google Scholar reveals dozens of papers published this year only. These attempts differ in many aspects, e.g. in how the data are sub-sampled, pre-processed or filtered, or which features are selected, and which type of classifier is used. For example, Prabhakar and Rajaguru (2020) compare over 50 different classification approaches with their accuracy varying between 80 % and 99 %. However, our goal is *not* to build a competitive classifier. It is merely to demonstrate how covariance estimation beyond separability can improve classification accuracy. For this purpose, we consider the functional linear discriminant analysis (fLDA) classifier (Baíllo et al., 2011). To the best of our knowledge, fLDA classifier has not been used before with the EEG data set.

Specifically, we utilize the centroid classifier of Delaigle et al. (2012), see also Kraus and Stefanucci (2019) for an elegant exposition. Assuming that the control group has a Gaussian distribution with mean $\mathbf{m}_0$ and covariance $\mathbf{C}$ while the alcoholic group has a Gaussian distribution with mean $\mathbf{m}_1$ and covariance $\mathbf{C}$, the optimal classifier (i.e. a predictor of Y given the EEG measurements $\mathbf{X}$) based on a one-dimensional projection is given by

$$\widehat{Y} := \mathbb{1}_{[\langle \mathbf{X} - \mathbf{m}_0, \mathbf{v} \rangle > \langle \mathbf{X} - \mathbf{m}_1, \mathbf{v} \rangle]}, \tag{3.27}$$

where $\mathbf{v}$ is a solution to the linear problem involving the covariance, namely

$$\mathbf{C}\mathbf{v} = \mathbf{m}_1 - \mathbf{m}_0, \tag{3.28}$$

provided this solution exists. If the solution does not exist, neither does the optimal centroid classifier. Regardless, however, $\mathbf{m}_0$, $\mathbf{m}_1$ and $\mathbf{C}$ are unknown in practice and have to be estimated from the data. The estimator of $\mathbf{C}$ is typically regularized by adding

---

[1] https://archive.ics.uci.edu/ml/datasets/eeg+database, downloaded on 14 May 2021.

Table 3.5: Out-of-sample cross-validated classification accuracy (i.e. the ratio between correctly classified subjects and all subjects) for the fLDA classifier with the separable ($R = 1$) or R-separable ($R = 2$) estimator of the covariance, and for three different data sets given by different conditions.

Degree-of-separability	Condition 1	Condition 2	Condition 3
R=1	78 %	79 %	77 %
R=2	90 %	84 %	91 %

a ridge (Baíllo et al., 2011; Kraus and Stefanucci, 2019) so the inverse problem can be solved and an estimator of $\mathbf{v}$, denoted $\widehat{\mathbf{v}}$, is obtained. Then, one obtains the classifier by plugging in $\widehat{\mathbf{v}}$ as well as $\widehat{\mathbf{m}_0}$ and $\widehat{\mathbf{m}_1}$ into (3.27).

While $\mathbf{m}_0$ and $\mathbf{m}_1$ can be easily estimated as the empirical means of the respective classes, estimation of the covariance poses an issue here. Even though the empirical covariance $\widehat{\mathbf{C}}_N \in \mathbb{R}^{256\times 64\times 256\times 64}$ can be evaluated in principle, we cannot expect it to be a good estimator, given the evidence in the previous section. Moreover, and more importantly, the inverse problem (3.28) is not computationally feasible with $\mathbf{C}$ estimated empirically.

Instead, one may use the separable estimator or the proposed R-separable estimator for $\mathbf{C}$. In this particular case, the cross-validation strategy of section 3.3.1 suggests the choice $R = 2$ for all three data sets. Then, the inverse algorithm of Section 3.4 can be used to solve (3.28) efficiently, and the classifier is obtained easily.

Table 3.5 demonstrates the gains acquired by estimation beyond separability. To compare the resulting two classifiers (one for the separable estimator and other for the 2-separable estimator of the covariance), we split the data set for every condition into folds $\mathcal{F}_k$, $k = 1, \dots, 24$ of size 5. Let $Y_{k,j}$, $k = 1, \dots, 24$, $j = 1, \dots, 5$, denote the class membership of the j-th observation in k-th fold, and $\widehat{Y}_{k,j}^{(-k)}(R, \epsilon)$ denote the predicted class membership of $Y_{k,j}$ obtained by the fLDA classifier trained solely on folds $\mathcal{F}_{k'}$, $k' \neq k$, using the R-separable estimator of the covariance and ϵI as the ridge regularizer. Out-of-sample cross-validated classification accuracy for a given classifier is then calculated as

$$\mathrm{ACC}(R, \epsilon) = \sum_{k=1}^{24} \sum_{j=1}^{5} \left| \widehat{Y}_{k,j}^{(-k)}(R, \epsilon) - Y_{k,j} \right| \Big/ 121.$$

The maximum accuracies over a grid of ridge constants ϵ are reported in Table 3.5. For every condition, the proposed R-separable estimator with $R = 2$ (which is the degree-of-separability suggested by cross-validation) clearly outperforms the separable alternative of $R = 1$.

4 Separability under Sparse Measurements

In this chapter, we discuss how to estimate a separable covariance when the surfaces are observed only on a small number of locations, which vary across the surfaces and are burdened with measurement errors. Examples of sparsely observed random surfaces include longitudinal studies (where only a part of a functional profile is measured at each visit, e.g. Lopez et al., 2020), geolocalized data (Yarger et al., 2020; Zhang and Li, 2020; Wang et al., 2020), or financial data (Fengler, 2009; Kearney et al., 2018). To the best of our knowledge, a procedure for non-parametric estimation of a separable covariance has not yet been established in the literature for the sparse sampling regime.

We have already seen in the previous chapters (in the dense sampling regime) that the assumption of separability reduces the computational burden of working on a multi-dimensional domain. At the same time, separability often amounts to oversimplification. However, while wrongfully assuming it introduces a bias, separability may still lead to improved estimation due to the bias-variance trade-off. This was observed in Chapter 3.

Sparse data are generally associated with higher computational complexity (extra costs associated with smoothing) as well as higher statistical complexity (variance is inflated due to sparse measurements burdened by noise). Hence the variance stemming from sparse measurements and noise contamination is often of a larger magnitude, sanctioning separability as a means to achieve a better bias-variance trade-off. Moreover, the faster computation and lower storage requirements, which separability entails, are pronounced in the sparse regime.

The goal of this chapter is to leverage the separability assumption to reduce complexity of covariance estimation down to that of mean estimation when working under the sparse regime. The method of choice for sparsely observed functional data on one-dimensional domains is the PACE approach (Yao et al., 2005a), which is based on kernel regression smoothers. A naive generalization of PACE to a two-dimensional domain would entail a computationally infeasible local linear smoothing step the in four-dimensional space. Instead, we demonstrate how to make careful use of separability to collapse the four-dimensional smoother into several two-dimensional surface smoothers. Consequently, the estimation of the mean, covariance, and noise level all become of similar computational complexity.

We illustrate the benefits of our approach on a qualitative analysis of implied volatility surfaces corresponding to call options. Here, one surface corresponds to a fixed asset (e.g. a stock), for which the right to buy it for an agreed-upon *strike* price (first dimension) at a future *time to expiration* (second dimension) is traded. The value of implied volatility at given strike and time to expiration is derived directly from observed market data (the option prices), by the well-known and commonly used Black-Scholes formula (Black and Scholes, 1973; Merton, 1973). The implied volatilities are preferred over the option prices, since they are dimensionless quantities, allow for a direct comparison of different assets, and are well familiar to practitioners. Since the options are traded only for a finite number of strikes and times to expiration, which vary across different surfaces, the observed data consist of a sparse ensemble. The interpolation of such an ensemble is a typical objective in financial mathematics as the latent implied volatility surfaces are of interest for other tasks such as prediction (forecasting) of option prices (Hull, 2006). The common practice is to interpolate or smooth the measurements for every surface independently. For example, Cont and Da Fonseca (2002) utilized pre-smoothing by local polynomial regression, evaluating an individual smoother for every single surface. This approach may, however, pose issues when the available sparse measurements for a given surface are concentrated only on a subset of the domain, which is often the case. Once the predicted surfaces are fed into a subsequent predictive models, the naively extrapolated parts of the surface are given the same weight as the more reliable interpolated parts, which can naturally hinder the resulting prediction quality. Instead, we advocate for the idea of "borrowing strength" for the purpose of predicting the latent surfaces via best linear unbiased prediction using the information from the entire data set, which also allows for uncertainty quantification (Yao et al., 2005a). Under separability, we only need to use two-dimensional surface smoothing, which is the case of the pre-smoothing approach as well. At the same time, the proposed methodology outperforms the pre-smoothing approach in terms of prediction error.

4.1 Model and Observation Scheme

We assume the existence of i.i.d. latent surfaces $X_n \in \mathcal{L}^2([0,1]^2)$, $n = 1, \dots, N$, which are mean-square continuous with continuous sample paths. We denote the mean function as $\mu = \mu(t,s)$, where

$$\mu(t,s) = \mathbb{E}X_1(t,s), \qquad t,s \in [0,1],$$

and the covariance kernel as $c = c(t,s,t',s')$, where

$$c(t,s,t',s') = \mathbb{E}\left[\Big(X_1(t,s) - \mu(t,s)\Big)\Big(X_1(t',s') - \mu(t',s')\Big)\right], \qquad t,s,t',s' \in [0,1].$$

Recall that we think of the first dimension as being temporal, denoted by variable t, and the second dimension as being spatial, denoted by variable s, though this convention is only made for the purposes of presentation.

The crucial assumption in this chapter is that of separability of the covariance:

(A1) The covariance kernel of the random surfaces $X_1, \dots, X_N$ satisfies

$$c(t, s, t', s') = a(t, t')b(s, s'), \qquad t, s, t', s' \in [0, 1], \tag{4.1}$$

for some purely temporal covariance $a = a(t, t')$ and some purely spatial covariance $b = b(s, s')$.

The process $X \in \mathcal{L}^2([0, 1]^2)$ is separable if it is, for example, an outer product of two independent univariate processes (a purely temporal one and a purely spatial one). In that case, apart from the covariance, the mean function also separates into a product of a purely temporal and a purely spatial functions. However, a process can have a separable covariance even when it is not itself separable (Rougier, 2017), for example the mean function may not be separable. We do not assume separability of the latent process itself, we only assume separability of its covariance in the sense of (4.1).

We work under the sparse sampling regime, where every surface is observed only at a finite number of irregularly distributed locations, and those measurements are corrupted by independent additive errors. For the n-th latent surface X_n, the number of measurements M_n as well as the locations of the measurements $\{(t_{nm}, s_{nm}) \mid m = 1, \dots, M_n\} \subset [0, 1]^2$ are considered random, and the observations are given by the errors-in-measurements model (Yao et al., 2005a; Li and Hsing, 2010; Zhang and Wang, 2016):

$$Y_{nm} = X_n(t_{nm}, s_{nm}) + \varepsilon_{nm}, \quad m = 1, \dots, M_n,\ n = 1, \dots, N, \tag{4.2}$$

where ε_{nm} are i.i.d. with $\mathbb{E}\,\varepsilon_{nm} = 0$ and $\mathrm{Var}(\varepsilon_{nm}) = \sigma^2 > 0$ being the noise level.

4.2 Motivation

In this section, we assume for simplicity that the mean is zero, and provide a heuristic description of how one might estimate the separable covariance (4.1). Note that we have

$$\mathrm{Cov}(Y_{nm}, Y_{nm'}) = a(t_{nm}, t_{nm'})b(s_{nm}, s_{nm'}) + \sigma^2 \mathbb{1}_{[m=m']}$$

for $n = 1, \dots, N$ and $m, m' = 1, \dots, M_n$.

Consider the *raw covariances* $G_{nmm'} := Y_{nm}Y_{nm'}$. Ignoring the assumption of separability, one could attempt to lift the PACE approach (Yao et al., 2005a) up to higher dimensions.

This amounts to plotting the raw covariances as a scatter plot with a four-dimensional domain, discarding the diagonal covariances burdened by noise, and using a surface smoother to obtain an estimator of the covariance. Specifically, this amounts to setting $\widehat{c} = \widehat{\gamma}_0$ with $\widehat{\gamma}_0 = \widehat{\gamma}_0(t, s, t', s')$ given for every fixed $(t, s', t', s') \in [0, 1]^4$ as

$$\begin{aligned}\underset{\gamma_0,\gamma_1,\gamma_2,\gamma_3,\gamma_4}{\arg\min} \sum_{n=1}^{N} \sum_{m=1}^{M_n} \sum_{m'=1}^{M_n} \mathcal{K}\left(\frac{t - t_{nm}}{h_1}\right) \mathcal{K}\left(\frac{s - s_{nm}}{h_2}\right) \mathcal{K}\left(\frac{t' - t_{nm'}}{h_3}\right) \mathcal{K}\left(\frac{s' - s_{nm'}}{h_4}\right) \cdot \\ \cdot \Big[G_{nmm'} - \gamma_0 - \gamma_1(t - t_{nm}) - \gamma_2(s - s_{nm}) \\ - \gamma_3(t' - t_{nm'}) - \gamma_4(s' - s_{nm}) \Big]^2,\end{aligned} \tag{4.3}$$

where $\mathcal{K}(\cdot)$ is a smoothing kernel function and $h_1, h_2, h_3, h_4 > 0$ are bandwidths. We refer to this procedure as *4D smoothing*. However, there are two issues with 4D smoothing. Firstly, the curse of dimensionality results in an estimator of poor quality, unless surfaces are observed relatively densely and many replications are available. And secondly, especially when the latter is true, the computational costs of smoothing in a higher dimension can be excessive. We make the assumption of separability mainly to cope with these two issues, which is often the case in the literature already when working with fully observed data (Gneiting et al., 2006; Genton, 2007; Pigoli et al., 2018). We do not see separability as a critical modeling assumption, but rather as a regularization, which possibly introduces a bias. Separability reduces both the statistical and the computational complexity of the covariance estimation task. This is always important when working with random surfaces, whatever their mode of observation, but becomes particularly crucial when working with sparsely observed surfaces.

In the following, we provide a heuristic on how separability can be used to our advantage in the sparse observation regime. Assuming zero mean for now, we have

$$\mathbb{E}G_{nmm'} = \mathbb{E}Y_{nm}Y_{nm'} = a(t_{nm}, t_{nm'})b(s_{nm}, s_{nm'}) + \sigma^2 \mathbb{1}_{[m=m']}.$$

Imagine for a moment that the spatial kernel $b = b(s, s')$ is known, and consider the set of values

$$\left\{ \frac{Y_{nm}Y_{nm'}}{b(s_{nm}, s_{nm'})} \;\middle|\; m, m' = 1, \ldots, M_n, m \neq m', n = 1, \ldots, N \right\}. \tag{4.4}$$

The expectation of every point in this set is $a(t_{nm}, t_{nm'})$, so we can chart these points in a scatter plot as

$$\left(t_{nm}, t_{nm'}, \frac{Y_{nm}Y_{nm'}}{b(s_{nm}, s_{nm'})} \right),$$

and use a two-dimensional surface smoother to obtain an estimator of $a = a(t, t')$.

Correspondingly, if one knew the temporal kernel $a = a(t,t')$ instead, the set of values (4.4) (with a in the denominator instead of b) could be arranged against s_{nm} and $s_{nm'}$ to obtain an estimator of $b = b(s,s')$. When neither the temporal kernel a nor the spatial kernel b is known, one can start with a fixed b and iterate between updates of a and b, smoothing a scatterplot once per every single update.

However, there are two issues with such an approach. Firstly, for small denominators, the corresponding points on the scatterplot are not reliable, and using them as they are can have a severe negative impact on estimation quality. Secondly, unless very few observations per surface are available, the procedure above can still be extremely demanding to compute, see Section 4.8.

To cope with these issues, we will use weights for the surface smoother and utilize gridding, i.e. split the domain into disjoint intervals and work on a grid. While, gridding can significantly reduce computations already on a univariate domain (Yao et al., 2005a), the gains are much bigger in higher dimensions. In the following section, we introduce our methodology in full from the theoretical perspective, while computational aspects are deferred to Section 4.5.

4.3 Estimation of the Model Components

We use local linear regression surface smoothers (Fan and Gijbels, 1996) to formalise the heuristic described in the previous section and estimate the components of the model from Section 4.1, i.e. the mean $\mu = \mu(t,s)$, the temporal kernel $a = a(t,t')$, the spatial kernel $b = b(s,s')$, and the noise level σ^2.

By applying a surface smoother to the set $\{(x_k, y_k, z_k) \mid k = 1,\dots,M\} \subset \mathbb{R}^3$ with given weights $\{w_k \mid k = 1,\dots,M\}$, we understand calculating $\widehat{\gamma}_0 = \widehat{\gamma}_0(x,y)$ as the minimizer of the weighted sum of squares

$$(\widehat{\gamma_0}, \widehat{\gamma_1}, \widehat{\gamma_2}) = \underset{\gamma_0,\gamma_1,\gamma_2}{\arg\min} \sum_{k=1}^{M} \mathcal{K}\left(\frac{x - x_k}{h_1}\right) \mathcal{K}\left(\frac{y - y_k}{h_2}\right) w_k \Big[z_k - \gamma_0 - \gamma_1(x - x_k) - \gamma_2(y - y_k) \Big]^2 \tag{4.5}$$

for every fixed $(x,y) \in [0,1]^2$, where $\mathcal{K}(\cdot)$ is a smoothing kernel function and $h_1, h_2 > 0$ are bandwidths. Throughout this chapter, we use the Epanechnikov kernel, utilize cross-validation to select the bandwidths, and mention weights only when they are not all equal.

First, we estimate the mean by applying the surface smoother to the set

$$\{(t_{nm}, s_{nm}, Y_{nm}) \mid m = 1,\dots,M_n,\ n = 1,\dots,N\}. \tag{4.6}$$

Denote the resulting estimator by $\widehat{\mu} = \widehat{\mu}(t,s)$.

Next, consider the *raw covariances* $G_{nmm'} = \Big[Y_{nm} - \widehat{\mu}(t_{nm}, s_{nm})\Big]\Big[Y_{nm'} - \widehat{\mu}(t_{nm'}, s_{nm'})\Big]$. We begin by applying the surface smoother to the set

$$\{(t_{nm}, t_{nm}, G_{nmm'}) \mid m, m' = 1, \dots, M_n,\ m \neq m',\ n = 1 \dots, N\}$$

to obtain a preliminary estimator of $a = a(t, t')$, denoted by $\widehat{a}_0 = \widehat{a}_0(t, t')$. Then we use this preliminary estimator to calculate a proxy of $b = b(s, s')$. Namely, we apply the surface smoother to the set

$$\left\{ \left(s_{nm}, s_{nm'}, \frac{G_{nmm'}}{\widehat{a}_0(t_{nm}, t_{nm'})} \right) \,\middle|\, m, m' = 1, \dots, M_n,\ m \neq m',\ n = 1 \dots, N \right\} \tag{4.7}$$

using weights $\{\widehat{a}_0^2(t_{nm}, t_{nm'})\}$ to obtain $\widehat{b}_0 = \widehat{b}_0(s, s')$. If the denominator in the set above is ever zero, we remove the corresponding point from the set. Note that since the weights are equal exactly to the denominators squared, it makes sense even formally that such a point is never considered for the surface smoother. As the next step, we refine our estimator of a by applying the surface smoother to the set

$$\left\{ \left(t_{nm}, t_{nm'}, \frac{G_{nmm'}}{\widehat{b}_0(s_{nm}, s_{nm'})} \right) \,\middle|\, m, m' = 1, \dots, M_n,\ m \neq m',\ n = 1 \dots, N \right\} \tag{4.8}$$

using weights $\{\widehat{b}_0^2(s_{nm}, s_{nm'})\}$ from which we obtain $\widehat{a} = \widehat{a}(t, t')$. Finally, we refine the estimator of b by applying the surface smoother to set (4.7) with $\widehat{a}_0$ replaced by $\widehat{a}$ and the weights adopted accordingly, resulting in the estimator $\widehat{b} = \widehat{b}(s, s')$.

Finally, once both the mean and the separable covariance have been estimated, it remains to estimate the noise level σ^2, which is of interest e.g. for the purposes of prediction. We begin by applying the surface smoother to the set

$$\{(t_{nm}, s_{nm}, G_{nmm}) \mid m = 1, \dots, M_n,\ n = 1 \dots, N\}$$

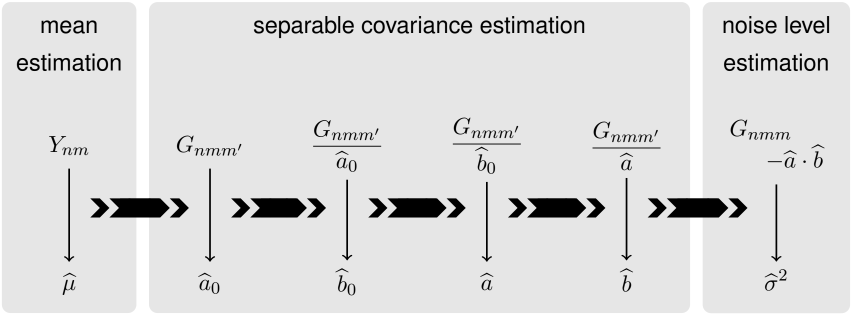
to obtain $\widehat{V} = \widehat{V}(t, s)$. Note that since $\mathbb{E} G_{nmm} \approx a(t_{nm}, t_{nm}) b(s_{nm}, s_{nm}) + \sigma^2$, we can estimate σ^2 by

$$\widehat{\sigma}^2 = 4 \int_{1/4}^{3/4} \int_{1/4}^{3/4} \Big[\widehat{V}(t, s) - \widehat{a}(t, t)\widehat{b}(s, s)\Big] \, \mathrm{d}t \, \mathrm{d}s,$$

where (similarly to Yao et al., 2005a) we integrate only along the middle part of the domain to mitigate boundary issues.

The workflow of the estimation scheme described above is visualised in Figure 4.1. The main novelty of our approach lies in the part where the separable covariance is estimated. Separability allows us to reduce dimensionality of the problem. Hence only two-dimensional smoothing is required, while a straightforward multi-dimensional generalization of e.g. the PACE approach (Yao et al., 2005a) not utilizing separability

Figure 4.1: Workflow of the proposed estimation procedure. We estimate firstly the mean from the data $\{Y_{nm}\}$, then the separable covariance (in several steps) from the raw covariances $\{G_{nmm'}\}$, and finally the noise level. A surface smoother over a 2D domain is utilized in every step (once per a single thin arrow).


mean estimation
separable covariance estimation
noise level estimation
Y_{nm}
$\widehat{\mu}$
$G_{nmm'}$
$\widehat{a}_0$
$\frac{G_{nmm'}}{\widehat{a}_0}$
$\widehat{b}_0$
$\frac{G_{nmm'}}{\widehat{b}_0}$
$\widehat{a}$
$\frac{G_{nmm'}}{\widehat{a}}$
$\widehat{b}$
$G_{nmm} - \widehat{a} \cdot \widehat{b}$
$\widehat{\sigma}^2$

would require four-dimensional smoothing to estimate the covariance.

The estimation of the separable terms can be viewed as an iterative procedure, where either a or b is kept fixed while the other term is being updated. The initial proxy $\widehat{a}_0$ is also obtained in this way, starting from $\widehat{b}_0 \equiv 1$. A natural question is whether one should iterate this process until convergence, or simply stop after a single step, and use e.g. $\widehat{a}_0$ as the estimator of a. The approach we advocate for uses exactly two steps (for both a and b). The reason is the following. As will be shown in Section 4.7, the asymptotic distribution of $\widehat{b}$ does not depend on $\widehat{a}_0$, and the same is true for $\widehat{a}$. This fact follows from separability. However, one can anticipate the finite sample performance of $\widehat{a}$ to be better than that of $\widehat{a}_0$, which is verified in our simulation study. One can think of the first step as estimating the optimal weights consistently, and the second step as using those consistently estimated weights to produce the estimators, which are expected to outperform the one-step proxies.

4.3.1 Alternative Number of Steps

While we propose to estimate the separable covariance as summarized in Figure 4.1, one can easily envision a multi-step procedure, instead of the proposed two-step variant. Due to the kernel smoothing step, we are unable to show convergence of a fully iterated procedure. On one hand, this is not a big issue, because low number of steps is usually sufficient, as shown in Section 4.8. On the other hand, especially when the covariance c satisfies

$$\int_{[0,1]^4} c(t,s,t',s')dtdsdt'ds' \approx 0,$$

more than two steps can sometimes lead to a better performance. In this section, we introduce a cross-validation (CV) scheme to choose the number of steps in a data-driven way.

Assume we are working on a grid of size $d \times d$, and let the n-th surface X_n be observed at $\Omega_n \subset \{1, \ldots, d\} \times \{1, \ldots, d\}$. Then, the observations are stored as matrices $\mathbf{X}_n$, $n = 1, \ldots, N$, where $\mathbf{X}[i, j]$ is available if and only if $(i, j) \in \Omega_n$. Let l denote the number of steps and $\widehat{\mathbf{C}}^{(l)}$ the estimator obtained using l steps. Let L denote the maximum candidate value for l, in practice one can choose e.g. $L = 5$. For a candidate estimator $\widehat{\mathbf{C}}^{(l)}$, its "goodness of fit" can be measured by $\mathbb{E}\|\widehat{\mathbf{C}}^{(l)} - \mathbf{C}\|_2^2$. Of course, we cannot evaluate this objective since we do not know $\mathbf{C}$. Our strategy will be to find a proxy, which can be evaluated based on the available data, and then choose the number of iterations l that minimizes such a proxy.

Firstly, since

$$\mathbb{E}\|\widehat{\mathbf{C}}^{(l)} - \mathbf{C}\|_2^2 = \mathbb{E}\|\widehat{\mathbf{C}}^{(l)}\|_2^2 - 2\mathbb{E}\langle \widehat{\mathbf{C}}^{(l)}, \mathbf{C}\rangle + \mathbb{E}\|\mathbf{C}\|_2^2,$$

where the last term does not depend on k, we only need to estimate the expected inner product in the previous equation. Let $\widehat{\mathbf{C}}_{-n}^{(l)}$ denote the estimator obtained without the n-th surface. Then we have

$$\mathbb{E}\langle \widehat{\mathbf{C}}^{(l)}, \mathbf{C}\rangle \approx \mathbb{E}\langle \widehat{\mathbf{C}}_{-n}^{(l)}, \mathbf{C}\rangle = \langle \mathbb{E}\widehat{\mathbf{C}}_{-n}^{(l)}, \mathbb{E}\mathbf{X}_n \otimes \mathbf{X}_n\rangle = \mathbb{E}\langle \widehat{\mathbf{C}}_{-n}^{(l)}, \mathbf{X}_n \otimes \mathbf{X}_n\rangle = \mathbb{E}\langle \widehat{\mathbf{C}}_{-n}^{(l)}, \mathbf{X}_n \otimes \mathbf{X}_n\rangle_\star, \tag{4.9}$$

where the middle equation is due to independence between the samples, $\langle \cdot, \cdot \rangle_\star$ is the sparse version of the inner product given by

$$\langle \widehat{\mathbf{C}}_{-n}^{(l)}, \mathbf{X}_n \otimes \mathbf{X}_n\rangle_\star = \\ = \frac{\|\widehat{\mathbf{C}}_{-n}^{(l)}\|_2^2}{\sum_{(i,j),(i',j') \in \Omega_n} (\widehat{\mathbf{C}}_{-n}^{(l)}[i,j,i',j'])^2} \sum_{\substack{(i,j),(i',j') \in \Omega_n \\ (i,j) \neq (i',j')}} \widehat{\mathbf{C}}_{-n}^{(l)}[i,j,i',j'] \mathbf{X}_n[i,j] \mathbf{X}_n[i',j']$$

and the last expectation in (4.9) corresponds to averaging over the sampling patter as well. Note that $\langle \cdot, \cdot \rangle_\star$ is defined such that it is equal to $\langle \cdot, \cdot \rangle$ in the case of a fully observed and noiseless surface $\mathbf{X}_n$. Now, the expected inner product in (4.9) can be naturally estimated using the available sample as the following weighted average:

$$\frac{\sum_{(i,j),(i',j') \in \Omega_n} (\widehat{\mathbf{C}}_{-n}^{(l)}[i,j,i',j'])^2}{\sum_{n=1}^N \sum_{(i,j),(i',j') \in \Omega_n} (\widehat{\mathbf{C}}_{-n}^{(l)}[i,j,i',j'])^2} \sum_{n=1}^N \langle \widehat{\mathbf{C}}_{-n}^{(l)}, \mathbf{X}_n \otimes \mathbf{X}_n\rangle_\star = \\ = \frac{\|\widehat{\mathbf{C}}_{-n}^{(l)}\|_2^2}{\sum_{n=1}^N \sum_{(i,j),(i',j') \in \Omega_n} (\widehat{\mathbf{C}}_{-n}^{(l)}[i,j,i',j'])^2} \sum_{\substack{(i,j),(i',j') \in \Omega_n \\ (i,j) \neq (i',j')}} \widehat{\mathbf{C}}_{-n}^{(l)}[i,j,i',j'] \mathbf{X}_n[i,j] \mathbf{X}_n[i',j'].$$

Finally, we have a fully calculable objective, which can be minimized over different values of l to suggest the optimal number of steps. The procedure above corresponds to the leave-one-out CV scheme. In practice, we use 10-fold CV instead. We show in Section 4.8 that the strategy presented here leads to a reasonable choice of the the number of steps l.

4.4 Weighting Scheme

One of the distinctive features of our methodology is the use of this explicit weighing scheme, where the weights for each of the two covariance kernels depend on the other covariance kernel. Here we explain the specific (quadratic) form of weight.

Firstly, we discuss how the weighting scheme using a fixed part of the separable covariance can be understood as an alternative to local weighting via the smoothing kernel. Assume we observe zero-mean surfaces sparsely, and one of these surfaces, say X_1, is observed at four locations only, as depicted in Figure 4.2. First, let us describe the 4D smoothing estimator at a fixed location (t, s, t', s'), when the bandwidths are also fixed. X_1 contributes to $\widehat{c}(t, s, t', s')$ only if there is a pair of two locations, where X_1 is observed, such that one of the locations is close to (t, s), and the other is close to (t', s'). In this case, closeness in time, resp. space, is measured by h_t, resp. h_s, which control the bandwidth of the smoothing kernel. In Figure 4.2 (left), only a single pair of locations contributes to estimation at (t, s, t', s').

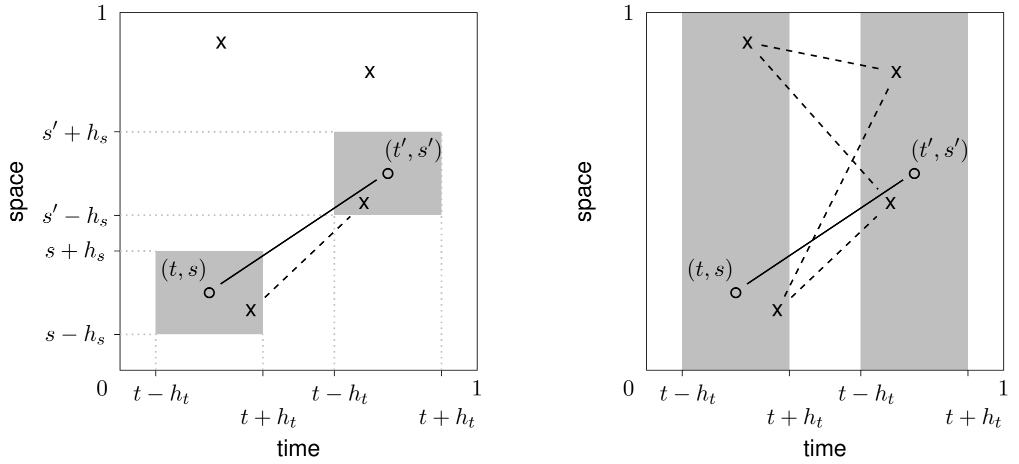
Now, let us contrast this to a single step in the proposed estimating procedure. Assume that $b = b(s, s')$ is fixed in the current step, and we are estimating $a = a(t, t')$. In this step, the spatial dimensions are not explicitly considered, we are performing smoothing only in the temporal dimensions. Hence the product of any two locations, where X_1 is observed, contributes to $\widehat{a}$, as long as the locations are close to (t, s) and (t', s') in the temporal domain. In the situation displayed in Figure 4.2 (right), this leads to four contributing raw covariance pairs. In other words, when estimating the temporal part of the covariance at (t, s, t', s'), we can consider even raw covariances, *which are spatially far* from (t, s, t', s'). This is meaningful due to separability. The adopted weighting scheme then ensures that raw covariances arising from points which are spatially distant are appropriately weighted.

To sum up, 4D smoothing can be understood as averaging over information about $c(t, s, t', s')$ captured in raw covariances, whose locations are close to (t, s, t', s'). Under separability, however, the proposed methodology borrows information in a different manner, always allowing for more freedom in one dimension or the other, depending on which dimension is currently held fixed.

Secondly, we provide a heuristic justification for the quadratic choice of weights in the smoothers for the estimation of the covariance kernels $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$, such as (4.7) or (4.8). The quadratic weights can be motivated by the connection to *weighted least squares*. We recall that weighted least squares are used for linear regression models where the model errors are not necessarily i.i.d. Their covariance matrix is assumed to be a diagonal matrix known up to a multiplicative constant:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \qquad \mathbb{E}[\varepsilon|\mathbf{X}] = 0, \qquad \mathrm{Var}(\varepsilon|\mathbf{X}) = \sigma_\varepsilon^2 \operatorname{diag}(\mathbf{v}), \qquad \sigma_\varepsilon^2 > 0, \tag{4.10}$$

Figure 4.2: A single random surface is observed at four locations (depicted by “x”), and these observations contribute differently to estimation of the covariance at a fixed location (t, s, t', s') in the case of 4D smoothing (**left**) and one step of the proposed approach leading to the estimator of $a(t, t')$ (**right**). The gray areas depict active neighborhoods and the dashed lines depict the contributing raw covariances (products of the values at the connected locations).



where $\mathbf{y} = (y_1, \dots, y_I)^\top$ is the response, $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_I^\top)^\top$ is the model matrix, and $\operatorname{diag}(\mathbf{v})$ denotes the diagonal matrix with the known vector $\mathbf{v} = (v_1, \dots, v_I)^\top$ on its diagonal. The regression coefficients β in the model (4.10) are estimated by the weighted least squares:

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{I} w_i \left(y_i - \mathbf{x}_i \beta \right)^2, \qquad \text{where} \quad w_i = \frac{1}{v_i}, \ i = 1, \dots, I. \tag{4.11}$$

Consider the surface smoother of

$$\left\{ \left(s_{nm}, s_{nm'}, \frac{G_{nmm'}}{\alpha(t_{nm}, t_{nm'})} \right) \middle| \, m, m' = 1, \dots, M_n, \, m \neq m', \ n = 1 \dots, N \right\}$$

where $\alpha(t, t')$, $t, t' \in [0, 1]$, is a fixed deterministic kernel. The kernel smoothing technique we deploy is based on fitting a linear regression locally. In view of model (4.10) we want to assess the variance of the response $G_{nmm'}/\alpha(t_{nm}, t_{nm'})$ to improve the estimation procedure:

$$\operatorname{Var}\left(\frac{G_{nmm'}}{\alpha(t_{nm}, t_{nm'})} \right) = \frac{1}{\alpha^2(t_{nm}, t_{nm'})} \operatorname{Var}\left(G_{nmm'} \right). \tag{4.12}$$

The variance of $G_{nmm'}$ is unknown and therefore cannot be used to improve the estimation. Still, we observe in equation (4.12) that the variance is multiplied by the reciprocal of $\alpha^2(t_{nm}, t_{nm'})$. Therefore, we would define the weights for the weighted least squares (4.11) as $w_i = \alpha^2(t_{nm}, t_{nm'})$, to utilize the knowledge we actually have.

Finally, we provide a more precise justification, showing that the quadratic choice of weights corresponds to the optimal choice, when data are observed densely. With fully observed surfaces, the separable model can be estimated via the generalized power iteration method, where a single step is given by the partial inner product between the empirical covariance and the previous step, as in Chapter 3. For example, when $b = b(s, s')$ is fixed, one step of the power iteration method is due to Proposition 10 given by

$$\widehat{a}(t, t') = \int_0^1 \int_0^1 b(s, s') \widehat{c}_N(t, s, t', s') \,\mathrm{d}s \,\mathrm{d}s' \Big/ \int_0^1 \int_0^1 b^2(s, s') \,\mathrm{d}s \,\mathrm{d}s', \tag{4.13}$$

where $\widehat{c}_N$ is the empirical covariance estimator. In this section, we demonstrate that, with fully observed data and with no smoothing conducted, the estimation methodology of Section 4.3 corresponds to the power iteration step (4.13).

Firstly, assume that b is fixed, and we are using the surface smoother on the set of points (4.8) to obtain $\widehat{a}$. Assume that the n-th surface is observed twice at the temporal location t, i.e. at two locations (t, s_1) and (t, s_2), and once more in a general location (t', s'). Let us denote the raw covariance corresponding to the n-th surface and locations (t, s) and (t', s') explicitly by $G_n(t, s, t', s')$. Then, two values are available for the location (t, t') in set (4.8):

$$\frac{G_n(t, s_1, t', s')}{b(s_1, s')} \quad \& \quad \frac{G_n(t, s_2, t', s')}{b(s_2, s')}.$$

The corresponding weights are $b^2(s_1, s')$ and $b^2(s_2, s')$, respectively. If the bandwidth is small enough, and no other observations are available for this location, $\widehat{a}(t, t')$ is calculated as a weighted average:

$$\left[b^2(s_1, s') \frac{G_n(t, s_1, t', s')}{b(s_1, s')} + b^2(s_2, s') \frac{G_n(t, s_2, t', s')}{b(s_2, s')} \right] \Big/ \left[b^2(s_1, s') + b^2(s_2, s') \right]$$

Also, for the purposes of the surface smoother, using the two points separately with their separate quadratic weights is equivalent to using the weighted average with the weight $b^2(s_1, s') + b^2(s_2, s')$.

When the temporal slice t of the n-th surface is observed fully, the weighted averaging can be done continuously:

$$\widehat{a}(t, t') = \int_0^1 b(s, s') G_n(t, s, t', s') ds \Big/ \int_0^1 b^2(s, s') \,\mathrm{d}s.$$

When the temporal slice t' of the n-th surface is also observed fully, the weighted average becomes

$$\widehat{a}(t, t') = \int_0^1 \int_0^1 b(s, s') G_n(t, s, t', s') \,\mathrm{d}s \,\mathrm{d}s' \Big/ \int_0^1 \int_0^1 b^2(s, s') \,\mathrm{d}s \,\mathrm{d}s'.$$

When this is true for all N surfaces, the result is averaged over all the independent realizations, and we arrive directly to (4.13), since $\widehat{c}_N(t,s,t',s') = \frac{1}{N}\sum_{n=1}^{N} G_n(t,s,t',s')$.

Altogether, our estimation procedure can be thought of (due to the specific weighting scheme used) as a sparse version of the partial inner product introduced in Chapter 3. This link has important computational implications, which are discussed in the following section.

4.5 Implementation Details

In our implementation, we assume that data arrive as sparse matrices (i.e. matrices with a substantial number of missing entries). A single step of our estimation procedure for the separable model can be understood as the sparsified version of the partial inner product. The partial inner product is in turn a marginalization operator. In the case of sparse data, marginalization process corresponds to preparation of the raw covariances for the 2D smoothing step, i.e. charting the raw covariances either in time or in space and weighting them as in formulas (4.8) and (4.7). Similarly to formulas (3.14) in the previous chapter, the marginalization step can be performed effectively on the level of data. More importantly, the scatter points can be pooled together during the marginalization process, resulting in substantial computational savings during the subsequent smoothing step.

Assume we observe matrices $\mathbf{Y}_1,\ldots,\mathbf{Y}_N \in \mathbb{R}^{K_1\times K_2}$ with only some of their entries known, i.e. most of the entries are missing. The marginal covariance kernels $a = a(t,t')$ and $b = b(s,s')$ are replaced by matrices $\mathbf{A} \in \mathbb{R}^{K_1\times K_1}$ and $\mathbf{B} \in \mathbb{R}^{K_2\times K_2}$, respectively. We assume again for simplicity that the mean $\mu = \mu(t,s)$ is zero. The raw covariances, stemming from a single latent surface, then form a tensor $\mathbf{G}_n = \mathbf{Y}_n \otimes \mathbf{Y}_n \in \mathbb{R}^{K_1\times K_2\times K_1\times K_2}$ with entries $\mathbf{G}_n[i,j,i',j'] = \mathbf{Y}_n[i,j]\mathbf{Y}_n[i',j']$ of which many are missing again.

Again, assume that $\mathbf{B}$ is fixed, and we are using the surface smoother on the discrete equivalent to set (4.8), i.e.

$$\left\{ \left(i, i', \frac{\mathbf{G}_n[i,j,i',j']}{\mathbf{B}[j,j']} \right) \middle| \; \mathbf{Y}_n \text{ observed at } (i,j) \text{ and } (i',j'),\ (i,j) \neq (i',j'),\ n = 1,\ldots,N \right\} \tag{4.14}$$

to obtain $\widehat{\mathbf{A}}$. Like in the previous section, assume $\mathbf{Y}_n$ was observed at locations at $[i,j_1]$, $[i,j_2]$ and $[i',j']$, where no two locations are the same. As explained in the previous section, it is equivalent for the surface smoother to replace the corresponding two values from (4.14), i.e.

$$\frac{\mathbf{G}_n[i,j_1,i',j']}{\mathbf{B}[j_1,j']} \quad \& \quad \frac{\mathbf{G}_n[i,j_2,i',j']}{\mathbf{B}[j_2,j']}$$

with weights $(\mathbf{B}[j_1, j'])^2$ and $(\mathbf{B}[j_2, j'])^2$, by a single value

$$\Big(\mathbf{B}[j_1, j']\mathbf{G}_n[i, j_1, i', j'] + \mathbf{B}[j_2, j']\mathbf{G}_n[i, j_2, i', j']\Big)\Big/\Big(\mathbf{B}^2[j_1, j'] + \mathbf{B}^2[j_2, j']\Big) \tag{4.15}$$

with the aggregated weight $(\mathbf{B}[j_1, j'])^2 + (\mathbf{B}[j_2, j'])^2$.

Let $\mathbf{y}_{n,i}$ (resp. $\mathbf{y}_{n,i'}$) denote the i-th (resp. i'-th) column of $\mathbf{Y}_n$. Let $\mathbf{q}_i$ denote the identifier of whether the entries of $\mathbf{y}_{n,i}$ were observed (and similarly $\mathbf{q}_{i'}$), i.e.

$$\mathbf{q}_i[l] = \begin{cases} 1, & l \in \{j_1, j_2\} \\ 0, & \text{otherwise} \end{cases} \qquad \& \quad \mathbf{q}_{i'}[l] = \begin{cases} 1, & l = j' \\ 0, & \text{otherwise.} \end{cases}$$

Then, value (4.15) can be calculated as $\mathbf{y}_{n,i}^\top \mathbf{B} \mathbf{y}_{n,i} / \mathbf{q}_i^\top \mathbf{B}_2 \mathbf{q}_{i'}$ with the aggregated weight given by $\mathbf{q}_i^\top \mathbf{B}_2 \mathbf{q}_{i'}$, where $\mathbf{B}_2$ is the entry-wise square of $\mathbf{B}$. Naturally, this can be generalized to the case when arbitrary number of entries in the i-th and i'-th columns of $\mathbf{Y}_n$ are observed. But more importantly, it can be also generalized to account for different pairs of columns of $\mathbf{Y}_n$ at the same time.

Let $\mathbf{Q}_n$ be formed by vectors $\mathbf{q}_i$, i.e. $\mathbf{Q}_n[i,j] = \mathbb{1}_{[\mathbf{Y}[i,j] \text{ is observed}]}$. Then the contribution of the n-th surface $\mathbf{Y}_n$ into set (4.14) can be calculated at once as

$$\mathbf{Y}_n^\top \widetilde{\mathbf{B}} \mathbf{Y}_n / \mathbf{Q}^\top \widetilde{\mathbf{B}}_2 \mathbf{Q} \tag{4.16}$$

where $\widetilde{\mathbf{B}}$ is $\mathbf{B}$ with the diagonal values replaces by zeros and $\widetilde{\mathbf{B}}_2$ is the entry-wise square of $\widetilde{\mathbf{B}}$. The diagonal values of $\mathbf{B}$ are replaced by zeros as described in order to discard products of the type $(\mathbf{Y}[i,j])^2$, which are burdened by noise.

The situation is analogous in the other step, when $\mathbf{A}$ is fixed and $\mathbf{B}$ is calculated. The whole procedure of estimating the separable covariance based on gridded sparse measurements is outlined in Algorithm 4.2.

Separability offers reductions in both time and memory complexities already when data are observed fully. Now, we argue that computational gains of separability are even greater, when data are observed sparsely and kernel smoothing is used.

Kernel smoothers are known to be computationally demanding. To directly evaluate a kernel smoother in d_1 locations using d_2 observations takes $\mathcal{O}(d_1 d_2)$ operations. Table 4.3 shows these quadratic complexities in our situation, explained below. Assume that N surfaces were observed on a grid of size $K \times K$ relatively densely (i.e. a fixed percentage of the grid was observed – this is not unrealistic since one often chooses the grid size in such a way), and an unbounded kernel was used. The quadratic complexity of kernel smoother translates into estimating a general covariance by a surface smoother in $\mathcal{O}(NK^8)$ operations, because all $\mathcal{O}(NK^4)$ raw covariances have to be accessed at every single one of $\mathcal{O}(K^4)$ grid points. When we consider N fixed, the resulting complexity in K, i.e. $\mathcal{O}(K^8)$, is huge. Under separability, not using the marginalization procedure, the complexity is $\mathcal{O}(K^6)$, because $\mathcal{O}(K^4)$ raw covariances have to be accessed at $\mathcal{O}(K^2)$ grid points. With marginalization, i.e. using formula (4.16), the time complexity drops down to $\mathcal{O}(K^4)$, because the number of raw covariances that has to be accessed at every grid point decreases to $\mathcal{O}(K^2)$.

Table 4.2 Algorithm for estimation of the separable model from sparsely observed (w.l.o.g. zero-mean) surfaces.

```
Input $\mathbf{Y}_1, \dots, \mathbf{Y}_N \in (\mathbb{R} \cup \{\diamond\})^{K_1 \times K_2}$, where $\diamond$ represents a missing value
  $\mathbf{Q}_n := \mathbb{1}_{[\mathbf{Y}_n \neq \diamond]} \in \{0,1\}^{K_1 \times K_2}$, for $n = 1, \dots, N$
  replace all $\diamond$ entries in $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ by zeros
  $\mathbf{A} := \big(1\big)_{i,j=1}^{K_1 \times K_1}$
repeat
    for $n = 1, \dots, N$
        $\widetilde{\mathbf{B}} := \mathbf{B}$ with diagonal entries replaced by zeros
        $\widetilde{\mathbf{B}}_2 :=$ entry-wise square of $\widetilde{\mathbf{B}}$
        $\mathbf{W}_n := \mathbf{Q}_n \widetilde{\mathbf{B}}_2 \mathbf{Q}_n^\top$
        $\mathbf{Z}_n := \mathbf{Y}_n \widetilde{\mathbf{B}} \mathbf{Y}_n^\top$ entry-wise divided by $\mathbf{W}_n$
    end for
      $\mathbf{A} :=$ surface smoother of $\{\mathbf{Z}_1, \dots, \mathbf{Z}_N\}$ with $\{\mathbf{W}_1, \dots, \mathbf{W}_N\}$ as the smoothing weights
    for $n = 1, \dots, N$
        $\widetilde{\mathbf{A}} := \mathbf{A}$ with diagonal entries replaced by zeros
        $\widetilde{\mathbf{A}}_2 :=$ entry-wise square of $\widetilde{\mathbf{A}}$
        $\mathbf{W}_n := \mathbf{Q}_n^\top \widetilde{\mathbf{A}}_2 \mathbf{Q}_n$
        $\mathbf{Z}_n := \mathbf{Y}_n^\top \widetilde{\mathbf{A}} \mathbf{Y}_n$ entry-wise divided by $\mathbf{W}_n$
    end for
      $\mathbf{B} :=$ surface smoother of $\{\mathbf{Z}_1, \dots, \mathbf{Z}_N\}$ with $\{\mathbf{W}_1, \dots, \mathbf{W}_N\}$ as the smoothing weights
until convergence (or only twice)
Output $\mathbf{A}, \mathbf{B}$
```

Table 4.3: Complexities for covariance estimation of a random surface observed on a $K \times K$ grid.

Complexity	Separability	Separability w/o Marginalization	4D smoothing
Time	$\mathcal{O}(K^4)$	$\mathcal{O}(K^6)$	$\mathcal{O}(K^8)$
Memory	$\mathcal{O}(K^2)$	$\mathcal{O}(K^4)$	$\mathcal{O}(K^4)$

In practice, the quadratic complexity of plain and simple kernel smoothers becomes intractable and there exist many computational approaches to reduce the burden. Most notably, the fast Fourier transform can be used on equispaced domains to reduce quadratic complexity to log-linear (Silverman, 1982), effectively cutting down the powers of K in the first row of Table 4.3. Many other accelerating approaches exist, see e.g. Raykar et al. (2010) or Langrené and Warin (2019), and references therein. However, software availability utilizing these computationally efficient approaches is rather limited, and this is particularly true for multi-dimensional problems.

We do not provide our own implementation of kernel smoothing. For the proposed approach, we implement Algorithm 1, which uses a "surface smoother". To this end, we utilize local linear smoothers provided in the `fdapace` package (Yao et al., 2005a; Chen et al., 2020b). We also utilize internal functions from `fdapace` to perform cross-validation for the choice of bandwidths.

For 4D smoothing, which we consider only for comparison, we use the `np` package (Hayfield and Racine, 2008), which is to the best of our knowledge the only `R` (R Core Team, 2020) package able to perform local linear polynomial regression surface smoothing in more than two dimensions. The 4D smoothing estimator requires smoothing in four dimensions. Even though the `np` package implements cross-validation to choose the bandwidths, we found the computational burden to be huge and the performance rather poor in our simulation study. Hence, whenever we use 4D smoothing, we fix the unknown bandwidths as chosen by cross-validation for the proposed separable model. While this intuitively leads to smaller than optimal bandwidths, we found out in our simulation study that bandwidths are governed mainly by smoothness of the underlying covariance rather than the number of points per surface available. Since the smoothness of a four-dimensional covariance and its separable proxy is similar, optimal bandwidths chosen for the proposed (separable) approach seem to be reasonable for 4D smoothing as well, and this was verified in our experiments. Regardless, we can hardly afford other strategy for choosing the four bandwidths for 4D smoothing. Even the sophisticated combination of cross-validation and optimization provided in the `np` package leads huge runtimes in our setups, see Section 4.8.

4.6 Prediction

Another objective of our methodology is the recovery of the latent surfaces based on the sparse and noisy observations thereon. The prediction method we are going to present in this section follows the principle of *borrowing strength* across the entire data set, an expression framed by Yao et al. (2005a). Specifically, consider the training data set $\{Y_{nm} : m = 1, \dots, M_n,\, n = 1, \dots, N\}$ composed of the observations made on random surfaces $X_1, \dots, X_N$ via (4.2), and a new random surface $X^\star$ observed under the same sampling protocol:

$$Y_m^\star = X^\star(t_m^\star, s_m^\star) + \varepsilon_m^\star, \qquad m = 1, \dots, M^\star. \tag{4.17}$$

We assume that $X^\star$ comes from the same population as $X_1, \dots, X_N$. In fact, it may be set (together with its sparse measurements) to one of the training surfaces $X_1, \dots, X_N$ if the task is to predict one of those.

Our prediction method is calibrated on all the observations, and this information is used for the prediction of $X^\star$. This contrasts to the pre-smoothing step often used in the functional data literature (Ramsay and Silverman, 2005, 2007), typically in the dense regime, where the prediction (usually some kind of a smoother) of $X^\star$ is based only on $\{Y_m^\star : m = 1, \dots, M^\star\}$.

Let $X^\star$ be a random surface with the mean $\mu(t,s)$, $t, s \in [0,1]$, and the covariance kernel $a(t,t')b(s,s')$, $t, t', s, s' \in [0,1]$, observed through sparse measurements (4.17). Then the best linear unbiased predictor of the latent surface $X^\star$ given the sparsely observed data $\mathbb{Y}^\star = (Y_1^\star, \dots, Y_{M^\star}^\star)$, denoted as $\Pi(X^\star|\mathbb{Y}^\star)$, is given by (c.f. Henderson, 1975):

$$\Pi(X^\star(t,s)|\mathbb{Y}^\star) = \mu(t,s) + \mathrm{Cov}(X^\star(t,s), \mathbb{Y}^\star)\, [\mathrm{Var}(\mathbb{Y}^\star)]^{-1}\, (\mathbb{Y}^\star - \mathbb{E}\mathbb{Y}^\star), \qquad t, s \in [0,1], \tag{4.18}$$

where

$$\mathrm{Cov}(X^\star(t,s), \mathbb{Y}^\star) = \Big(a(t, t_m^\star) b(s, s_m^\star) \Big)_{m=1}^{M^\star} \in \mathbb{R}^{M^\star}, \qquad t, s \in [0,1], \tag{4.19}$$

$$\mathrm{Var}(\mathbb{Y}^\star) = \Big(a(t_m^\star, t_{m'}^\star) b(s_m^\star, s_{m'}^\star) + \sigma^2 \mathbb{1}_{[m=m']} \Big)_{m,m'=1}^{M^\star} \in \mathbb{R}^{M^\star \times M^\star}, \tag{4.20}$$

$$\mathbb{E}\mathbb{Y}^\star = \Big(\mu(t_m^\star, s_m^\star) \Big)_{m=1}^{M^\star} \in \mathbb{R}^{M^\star}.$$

Formula (4.18) contains the unknown mean surface $\mu(\cdot,\cdot)$, covariance kernels $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ as well as the measurement error variance σ^2. In reality, we estimate these quantities from the training data set, say sparsely observed measurements on $X_1, \dots, X_N$, and plug-in the estimates $\widehat{\mu}, \widehat{a}, \widehat{b}$, and $\widehat{\sigma}^2$ into (4.18). We shall denote such predictor as $\widehat{\Pi}(X^\star|\mathbb{Y}^\star)$. We show in the following section (Theorem 11) that the predictor $\widehat{\Pi}(X^\star|\mathbb{Y}^\star)$

converges to its theoretical counterpart $\Pi(X^\star|\mathbb{Y}^\star)$ as the number of training samples N grows to infinity.

In the rest of this section, we turn our attention to the construction of confidence bands under a Gaussian assumption.

(A2) The random surface $X^\star$ is a Gaussian random element in $\mathcal{L}^2([0,1]^2)$ and the measurement error ensemble $\{\varepsilon_m^\star\}_{m=1}^{M^\star}$ is a Gaussian random vector.

First note that under the assumption (A2), the best linear unbiased predictor (4.18) actually corresponds to the conditional expectation $\mathbb{E}[X^\star(t,s)|\mathbb{Y}^\star]$. Furthermore, the conditional covariance structure is given for $t, t', s, s' \in [0,1]$ by

$$\begin{aligned}&\operatorname{Cov}\left(X^\star(t,s), X^\star(t',s')|\mathbb{Y}^\star\right)\\ &\qquad = a(t,t')b(s,s') - \operatorname{Cov}(X^\star(t,s),\mathbb{Y}^\star)\left[\operatorname{Var}(\mathbb{Y}^\star)\right]^{-1}\left[\operatorname{Cov}(X^\star(t',s'),\mathbb{Y}^\star)\right]^\top. \end{aligned}\tag{4.21}$$

Moreover, we denote $\widehat{\operatorname{Cov}}\left(X^\star(t,s), X^\star(t',s')|\mathbb{Y}^\star\right)$ the empirical counterpart to (4.21), where the unknown quantities a, b and σ^2 are replaced by their estimators.

For $(t,s) \in [0,1]^2$ and $\alpha \in (0,1)$, the $(1-\alpha)$-confidence interval for $X^\star(t,s)$ is given by

$$\widehat{\Pi}(X^\star(t,s)|\mathbb{Y}^\star) \pm u_{1-\alpha/2}\sqrt{\widehat{\operatorname{Cov}}\left(X^\star(t,s), X^\star(t,s)|\mathbb{Y}^\star\right)} \tag{4.22}$$

where $u_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard Gaussian law. The point-wise confidence band is then constructed by connecting the intervals (4.22) for all $(t,s) \in [0,1]^2$.

The construction of the simultaneous confidence band is more involved, and we shall use the technique proposed by Degras (2011). Define the conditional correlation kernel

$$\widehat{\operatorname{Corr}}\left(X^*(t,s), X^*(t',s')|\mathbb{Y}^*\right) = \frac{\widehat{\operatorname{Cov}}\left(X^*(t,s), X^*(t',s')|\mathbb{Y}^*\right)}{\sqrt{\widehat{\operatorname{Var}}\left(X^*(t,s)|\mathbb{Y}^*\right)\widehat{\operatorname{Var}}\left(X^*(t',s')|\mathbb{Y}^*\right)}} \tag{4.23}$$

if the division on the right-hand side makes sense and zero otherwise, and where we define $\widehat{\operatorname{Var}}\left(X^*(t,s)|\mathbb{Y}^*\right) = \widehat{\operatorname{Cov}}\left(X^*(t,s), X^*(t,s)|\mathbb{Y}^*\right)$.

Then, construct the simultaneous confidence band by connecting the intervals

$$\widehat{\Pi}(X^\star(t,s)|\mathbb{Y}^\star) \pm \widehat{u}_{1-\alpha}\sqrt{\widehat{\operatorname{Cov}}\left(X^\star(t,s), X^\star(t,s)|\mathbb{Y}^\star\right)}, \tag{4.24}$$

where $\widehat{u}_{1-\alpha}$ is the $(1-\alpha)$-quantile of the law of

$$\widehat{W} = \sup_{t,s\in[0,1]^2}\left|\widehat{Z}(t,s)\right| \tag{4.25}$$

with $\widehat{Z}$ being a Gaussian element with $\mathrm{Cov}(\widehat{Z}(t,s),\widehat{Z}(t',s')) = \widehat{\mathrm{Corr}}\,(X^\star(t,s), X^\star(t',s')|\mathbb{Y}^\star)$ for $t,t',s,s' \in [0,1]$. Therefore, by the definition, $\mathbb{P}(\sup_{(t,s)\in[0,1]^2} |\widehat{Z}(t,s)| \leq \widehat{u}_{1-\alpha}) = 1-\alpha$. Numerical calculation of this quantile is explained by Degras (2011), who also concludes that $\widehat{u}_{1-\alpha} < u_{1-\alpha/2}$. Therefore the point-wise confidence band is always enveloped by the simultaneous confidence band, as expected.

The asymptotic coverage of the point-wise band (4.22) and the simultaneous band (4.24) is verified in Theorem 11 in the following section.

4.7 Asymptotic Properties

In this section, we establish consistency and convergence rates of the estimators $\widehat{\mu} = \widehat{\mu}(t,s)$, $\widehat{a} = \widehat{a}(t,t')$, and $\widehat{b} = \widehat{b}(s,s')$, as well as the measurement error variance σ^2.

The following assumptions refine the sparse observation scheme introduced in Section 4.1.

(B1) The counts of measurements per surface M_n are independent identically distributed random variables with the law $M_n \sim \mathcal{M} > 0$ such that $\mathbb{P}(\mathcal{M} > 1) > 0$ and $\mathcal{M} \leq M^{max}$ where $M^{max} \in \mathbb{N}$ is a constant.

(B2) The measurement locations (t_{nm}, s_{nm}), $m = 1,\dots,M_n$, $n = 1,\dots,N$, are independent identically distributed random variables generated from the density $f_{(t,s)}(\cdot,\cdot)$ on $[0,1]^2$. The density $f_{(t,s)}(\cdot,\cdot)$ is assumed to be twice continuously differentiable and positive on $[0,1]^2$.

(B3) The counts (M_n), the locations (t_{nm}, s_{nm}), and the latent surfaces (X_n) are independent.

The following two assumptions are required for consistent estimation of the mean surface.

(B4) The mean surface $\mu(\cdot,\cdot)$ is twice continuously differentiable on $[0,1]^2$.

(B5) There exists $\rho > 2$ such that the random surface X_1 and the measurement error ε_{11} satisfy

$$\sup_{(t,s)\in[0,1]^2} \mathbb{E}\left[|X_1(t,s)|^\rho\right] < \infty, \qquad \mathbb{E}\left[|\varepsilon_{11}|^\rho\right] < \infty.$$

(B6) The bandwidths $h_{\mu,1}, h_{\mu,2}$ for the mean estimator satisfy $(\log N)/(N h_{\mu,1} h_{\mu,2}) = o(1)$ and furthermore we assume that they decay with the same rate: $h_{\mu,1} \asymp h$ and $h_{\mu,2} \asymp h$ as $N \to \infty$. The statement $x_n \asymp x'_n$ as $n \to \infty$ for two sequences $\{x_n\}$ and $\{x'_n\}$ is understood as $\lim_{n\to\infty} x_n/x'_n \in (0,\infty)$, i.e. x_n and x'_n differ asymptotically up to a multiplicative constant

The common decay rate assumption, which is included in (B6), is not really required for our asymptotic theory; it is imposed to simplify the statements on the convergence rates. The same argument applies to the other bandwidths below. Because all smoothing in our methodology is restricted to two dimensions, the bandwidths are indeed expected to decay with the same rate, and we may assume that they differ asymptotically up to a multiplicative constant.

In order to estimate the covariance kernels $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ we need the following assumptions:

(B7) The covariance kernels $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are twice continuously differentiable on $[0,1]^2$.

(B8) There exists $\rho' > 2$ such that the random surface X_1 and the measurement error ε_{11} satisfy

$$\sup_{(t,t',s,s')\in[0,1]^4} \mathbb{E}\left[|X_1(t,s)X_1(t',s')|^{\rho'}\right] < \infty, \qquad \mathbb{E}\left[|\varepsilon_{11}|^{2\rho'}\right] < \infty.$$

(B9) The bandwidths h_a and h_b used for smoothing the covariance kernel $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ satisfy $(\log N)/(Nh_a^2) = o(1)$ and $(\log N)/(Nh_b^2) = o(1)$ as $N \to \infty$ and, for simplicity of the convergence rates statements, we assume $h_a \asymp h$ and $h_b \asymp h$ as $N \to \infty$ where h is from assumption (B6).

(B10) The true value of the covariance kernel $b(\cdot,\cdot)$, of the separable model (4.1) satisfies

$$\Theta \overset{\text{def}}{=} \int_0^1 \int_0^1 b(s,s') f_s(s) f_s(s') \,\mathrm{d}s\,\mathrm{d}s' \neq 0 \tag{4.26}$$

where $f_s(s) = \int_0^1 f_{(t,s)}(t,s)\,\mathrm{d}t$ is the marginal density of the random location s_{11}.

While the other assumptions are standard in the smoothing literature, assumption (B10) might be surprising, especially considering it is not symmetric between $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$. The reason behind this asymmetry is that our estimation methodology starts by smoothing the raw covariances $G_{nmm'}$ against $(t_{nm}, t_{nm'})$ in order to produce the preliminary estimator $\widehat{a}_0(\cdot,\cdot)$. The condition (4.26) ensures that the estimator $\widehat{a}_0(\cdot,\cdot)$ converges to a nonzero quantity, see the constant Θ in Theorem 10. By contrast, this issue is not present in the follow-up steps. Due to positive semi-definitness of $b(\cdot,\cdot)$, the constant Θ can only be zero if all eigenfunctions of $b(\cdot,\cdot)$ are orthogonal to f_s. This cannot happen e.g. unless B is exactly low-rank, and with all the eigenfunctions changing signs. From the practical perspective, the condition is merely a technicality.

The estimation of the noise level σ^2 furthermore requires the following assumption.

(B11) The bandwidths $h_{V,1}, h_{V,2}$ for the smoother $\widehat{V}(\cdot,\cdot)$ satisfy $(\log N)/(N h_{V,1} h_{V,2}) = o(1)$ and, for simplicity of the convergence rates statements, we assume $h_{V,1} \asymp h$ and $h_{V,2} \asymp h$ as $N \to \infty$ where h is from assumption (B6).

The mean surface asymptotic theory is presented as the following proposition. Note that the separability assumption (A1) is not required here.

Proposition 12. *Under assumptions (B1) – (B6):*

$$\sup_{(t,s)\in[0,1]^2} |\widehat{\mu}(t,s) - \mu(t,s)| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N h^2}} + h^2\right) \qquad as \quad N \to \infty.$$

Our main asymptotic result, the consistency and the convergence rates for the separable model components (4.1), is presented in the following theorem.

Theorem 10. *Under assumptions (A1), (B1) – (B10):*

$$\sup_{(t,t')\in[0,1]^2} |\widehat{a}(t,t') - \Theta a(t,t')| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N h^2}} + h^2\right), \tag{4.27}$$

$$\sup_{(s,s')\in[0,1]^2} \left|\widehat{b}(s,s') - \frac{1}{\Theta} b(s,s')\right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N h^2}} + h^2\right), \tag{4.28}$$

as $N \to \infty$, where Θ is defined in (4.26)*.*

The separable decomposition (4.1) is not identifiable, because a constant can multiply one component while dividing the other, i.e. $a(t,t')b(s,s') = [\lambda a(t,t')]\,[(1/\lambda) b(s,s')]\,, t,t',s,s' \in [0,1]$, for any $\lambda \in (0,\infty)$. Therefore we can only aim to recover the covariance kernels $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ up to a multiplicative constant and its reciprocal, respectively. The number Θ in statements (4.27) and (4.28) plays the role of such a constant and depends on the initialization of the algorithm, in our case on the fact that the first estimator $\widehat{a}_0$ smooths the raw covariances $G_{nmm'}$ without any weighting. Still, the product $\widehat{a}(t,t')\widehat{b}(s,s')$, $t,t',s,s' \in [0,1]$, estimates consistently the covariance structure $c(t,s,t',s') = a(t,t')b(s,s')$, $t,t',s,s' \in [0,1]$, which is summarised in the following corollary.

Corollary 4. *Under assumptions (A1), (B1) – (B10):*

$$\sup_{(t,s,t',s')\in[0,1]^4} \left|\widehat{a}(t,t')\widehat{b}(s,s') - a(t,t')b(s,s')\right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N h^2}} + h^2\right)$$

as $N \to \infty$.

Finally, the asymptotic behaviour of the noise level σ^2 is given as the following proposition.

Proposition 13. *Under assumptions (A1), (B1) – (B11):*

$$\widehat{\sigma}^2 = \sigma^2 + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right) \qquad as \quad N \to \infty.$$

This completes the asymptotic theory for our estimsators, and we now turn to prediction. The following theorem shows that the predictor $\widehat{\Pi}(X^{new}|\mathbb{Y}^{new})$ defined in Section 4.6 converges – as the sample size grows to infinity – to its oracle counterpart (4.18), which assumes the knowledge of the true distribution of the data. Moreover, the theorem also proves the asymptotic coverage of the point-wise and simultaneous confidence bands (4.22) and (4.24).

Theorem 11. *Under assumptions (A1), (B1) – (B11):*

$$\sup_{(t,s)\in[0,1]^2} \left|\widehat{\Pi}(X^{new}(t,s)|\mathbb{Y}^{new}) - \Pi(X^{new}(t,s)|\mathbb{Y}^{new})\right| = o_{\mathbb{P}}(1), \qquad as \quad N \to \infty, \tag{4.29}$$

conditionally on $\mathbb{Y}^{new}$.

Assuming further (A2) and fixing $\alpha \in (0,1)$:

$$\forall (t,s) \in [0,1]^2 \lim_{N\to\infty} \mathbb{P}\left(\left|\widehat{\Pi}(X^{new}(t,s)|\mathbb{Y}^{new}) - X^{new}(t,s)\right| \leq u_{1-\alpha}\sqrt{\widehat{v}} \,\middle|\, \mathbb{Y}^{new}\right) = 1-\alpha,$$

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{(t,s)\in[0,1]^2} \widehat{v}^{-1/2}\left|\widehat{\Pi}(X^{new}(t,s)|\mathbb{Y}^{new}) - X^{new}(t,s)\right| \leq \widehat{z}_{1-\alpha} \,\middle|\, \mathbb{Y}^{new}\right) = 1-\alpha,$$

where $\widehat{v} = \widehat{\mathrm{Var}}\,(X^{new}(t,s)|\mathbb{Y}^{new})$.

The rates established in this section manifest the statistical consequences of separability. Corollary 4 shows that the complete covariance structure is estimated with the rate $O_{\mathbb{P}}(\sqrt{(\log N)/(Nh^2)} + h^2)$, which is the known optimal minimax convergence rate (Fan and Gijbels, 1996) for two dimensional non-parametric regression. By steps similar to our proofs (postponed to the appendix), it could be shown that the empirical covariance smoother yields the convergence rate $O_{\mathbb{P}}(\sqrt{(\log N)/(Nh^4)}+h^2)$. The empirical covariance smoother's convergence rate is thus slower than the one found in Corollary 4, achieved via the separable model.

4.8 Simulation Study

We explore the finite sample performance of the proposed methodology by means of a moderate simulation study (total runtime of about one thousand CPU hours). Computational efficiency (relatively small runtimes) is achieved by working on a 20 × 20 grid, like described in Section 4.5. Every surface $\mathbf{X}_1, \ldots, \mathbf{X}_{100}$ is first sampled fully on this grid as a zero-mean matrix-variate Gaussian with covariance $\mathbf{C}$ (to be specified), superposed with noise (zero-mean i.i.d. Gaussian entries with variance σ^2), and then sub-sampled in a way that only a fraction of the entries, selected at random, is retained. The covariance $\mathbf{C}$ is always standardized to have trace one, and σ^2 is chosen such that the gridded white noise process is also trace one.

Methods compared. We compare the proposed separable estimator $\widehat{c} = \widehat{a} \cdot \widehat{b}$ against the non-separable empirical estimator obtained by local linear smoothing in four dimensions (4D smoothing), and also against the nearest Kronecker product (NKP Van Loan and Pitsianis, 1993) approximation to $\widehat{\mathbf{C}}_N$ obtained from the fully observed and noise-free surfaces. We also compare the proposed estimator against its *one-step* version ($\widehat{c} = \widehat{a}_0 \cdot \widehat{b}_0$, cf. Figure 4.1).

Covariance choices. We consider four specific choices for the covariance.

(a) The Fourier scenario, where $a = a(t,t')$ and $b = b(s,s')$ are chosen to be the same, such that they have the trigonometric basis as their eigenfunctions and power decay of their eigenvalues, resulting in a rather wiggly univariate covariance displayed in Figure 4.3 (left). The covariance is then set as $c(t,s,t',s') = a(t,t')\, b(s,s')$, resulting in a separable covariance.

(b) The Brownian scenario, where $a = a(t,t')$ and $b = b(s,s')$ are both chosen as the covariance of the Wiener process, i.e. $a(t,t') = \min(t,t')$ and $b(s,s') = \min(s,s')$, resulting in a rather flat covariance displayed in Figure 4.3 (center). The covariance is then set as $c(t,s,t',s') = a(t,t')\, b(s,s')$, i.e. it is separable again.

(c) The Gneiting scenario, where the covariance has the following parametric form:

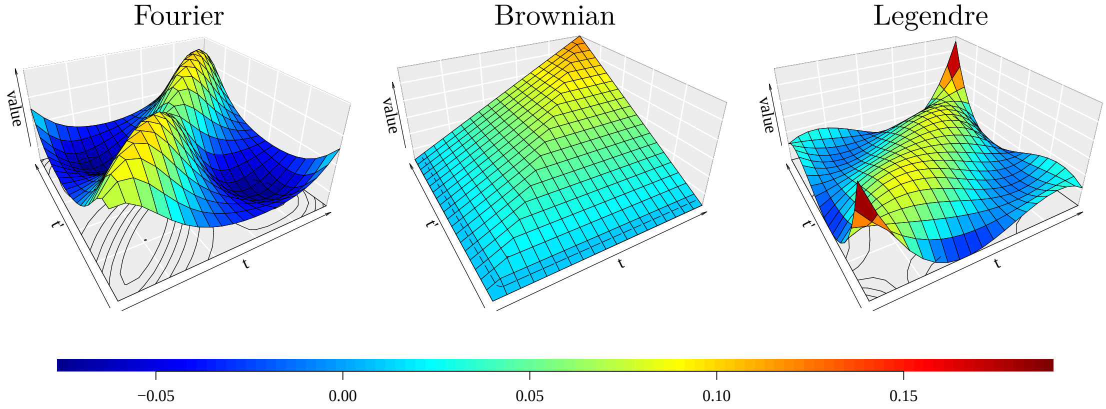
$$c(t,s,t',s') = \frac{\sigma^2}{(a^2|t-t'|^{2\alpha}+1)^{\tau}} \exp\left(\frac{b^2|s-s'|^{2\gamma}}{(a^2|t-t'|^{2\alpha}+1)^{\beta\gamma}} \right), \tag{4.30}$$

where $a = b = \tau = \alpha = \gamma = \sigma^2 = 1$ and $\beta = 0.7$. This covariance is non-separable (Gneiting, 2002), but it is rather flat.

(d) The Fourier-Legendre scenario, where we choose $a_1 = a_1(t,t')$ and $b_1 = b_1(s,s')$ as the Fourier univariate covariances specified above. Furthermore, $a_2 = a_2(t,t')$ and $b_2 = b_2(s,s')$ are both chosen as rank-4 covariances with shifted Legendre basis as their eigenfunctions, resulting in rather wiggly univariate covariances (see Figure 4.3, right). The covariance is then chosen as $c(t,s,t',s') = a_1(t,t')\, b_1(s,s') + a_2(t,t')\, b_2(s,s')$, resulting in a non-separable covariance.

To understand the simulation results, it is only important to point out the following. Firstly, while the Fourier setup (a) and the Brownian setup (b) are separable, the Gneiting setup (c) and the Fourier-Legendre setup (d) are non-separable. Secondly, while the

Figure 4.3: Covariances used as building blocks in the simulation study.



Brownian setup (b) and Gneiting setup (c) lead to rather flat covariances, the Fourier setup (a) and the Fourier-Legendre setup (d) lead to quite wiggly (though infinitely smooth) covariances.

Sparsity. In all simulation setups, we consider different percentages of the entries observed $p \in \{2, 5, 10, 20, 40, 70\}$. Since the grid size is 20×20, this means for example that for $p = 2$ we have $2/100 \cdot 20^2 = 8$ observations per surface. For all the setups and percentages, we report the relative estimation errors $\|\widehat{\mathbf{C}} - \mathbf{C}\|_2/\|\mathbf{C}\|_2$, where $\widehat{\mathbf{C}}$ is an estimator obtained by one of the four methods above (4D smoothing, one-step, proposed, or NKP). The results are shown in Figure 4.4. Every reported error was calculated as an average of 100 Monte Carlo runs.

Error components. The reported estimation errors can be thought of having four components:

(i) asymptotic bias, which is zero if the true covariance is separable, i.e. in cases (a) and (b);

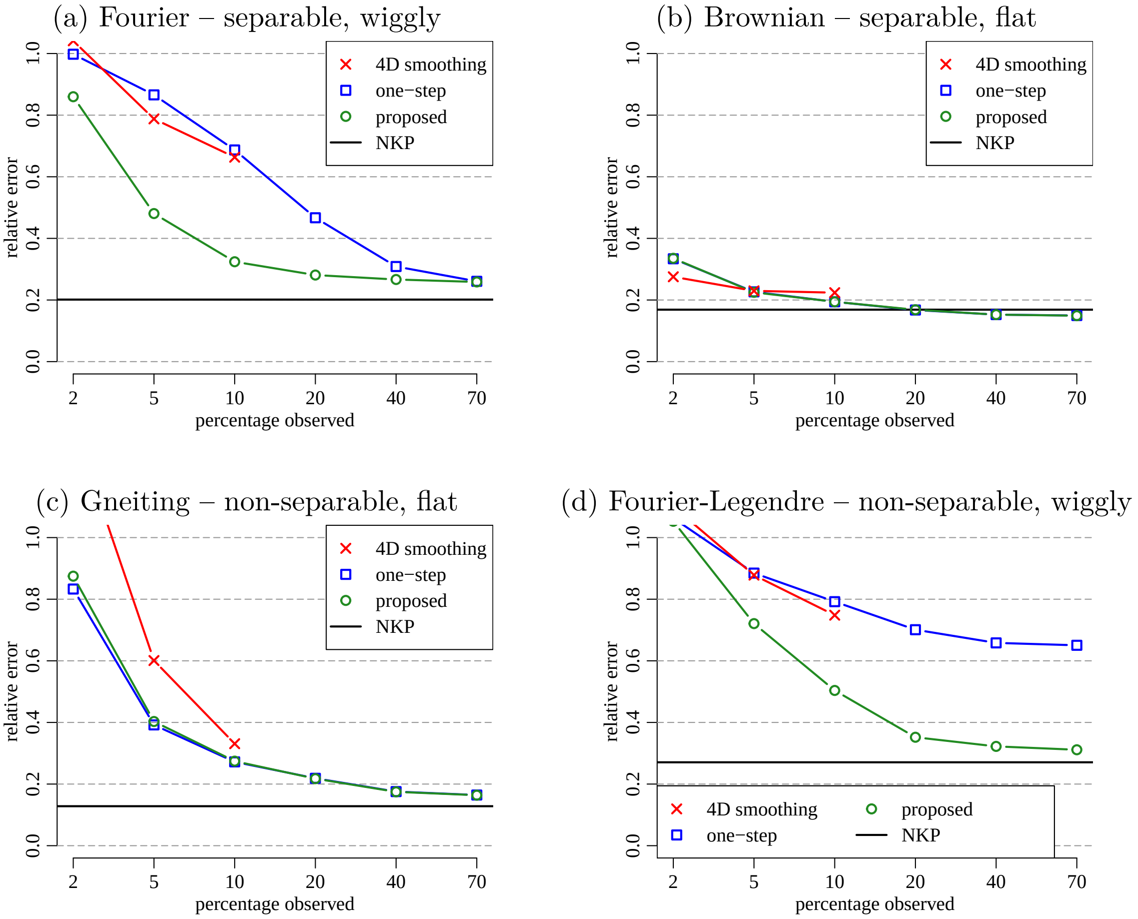
(ii) error due to finite number of samples ($N = 100$);

(iii) error due to sparse observations (i.e. not observing the full surfaces); and

(iv) noise contribution. NKP errors are always free of the latter two, providing a baseline.

The effect of not observing the full surfaces is displayed for different values of p. Finally, the noise contamination prevents smoothing approaches to reach the performance of NKP even for p large. Although our methodology explicitly handles noise, the finite sample performance is better with noise-free data, which only NKP has access to.

Figure 4.4: Relative estimation errors depending on percentages of the surfaces observed p for 4 ground truth covariance choices (a)-(d) and 4 methods compared. NKP provides a baseline, having access to full surfaces and hence not depending on p. For 4D smoothing, only results for small p are reported.



Results. There are several comments to be made about the results in Figure 4.4:

1. In the setups where the covariance is flat, i.e. (b) and (c), the one-step and the proposed approaches work the same, and 4D smoothing also works relatively well. These two setups are simple in a sense, because information can be borrowed quite efficiently via smoothing, regardless of whether the truth is separable or not. Still, the proposed approach utilizing separability does not perform worse than 4D smoothing even in the non-separable case (c), having the advantage of being much faster to obtain. For $p = 10$ the proposed estimator takes only a couple of seconds while 4D smoothing takes about 40 minutes even at this relatively small size of data.

2. When the true covariance is wiggly, the proposed methodology clearly outperforms 4D smoothing, both for the separable truth (a) and the non-separable truth (d).

The reason is that smoothing procedures are not very efficient in this case, and borrowing strength via separability is imperative.

3. The reason why error curves for 4D smoothing are only calculated up to $p = 10$ is the computational cost of smoothing in higher dimensions. In fact, performing 4D smoothing when $p = 10$ took more time than calculation of all the remaining results combined. This even takes into account the fact that cross-validation is not performed for 4D smoothing, cf. Section 4.5. Should we opt for cross-validation in the four-dimensional days, the total runtime of this simulation study would increase from the current one thousand CPU hours to over one thousand CPU days.

Now, we show runtimes for the Fourier scenario from our simulation study in Figure 4.5 (left). The runtimes look similarly for any of the remaining scenarios (not reported). To demonstrate effectiveness of the marginalization procedure described in detail in Section 4.5, we also show runtimes for the non-pooled procedure, considering all raw covariances in sets (4.8) or (4.7) as separate points for the purposes of smoothing. It leads to the same results as the proposed approach but, as more and more points per surface are observed, the number of scatter points supplied to the smoothing procedure increases rapidly, which increases the runtimes. But more importantly, at the edge of computational feasibility for the 4D smoothing approach (i.e. with the percentage $p = 10$) the proposed separable estimator is calculated about 200 times faster than the 4D smoothing estimator.

Next, it was observed above that the proposed approach outperforms the one-step version of our estimator $\widehat{a}_0 \cdot \widehat{b}_0$, cf. Figure 4.1. While the proposed approach can be seen as a two-step version, a natural question arises whether a multi-step version of the estimator could not be much better. Figure 4.5 compares the estimation errors achieved by the proposed approach and by a three-step approach in all four scenarios considered in Section 4.8. The third step offers a significant improvement in only one of the scenarios, and even then the improvement is relatively small compared to the improvements achieved by using the proposed two-step methodology as opposed to 4D smoothing. Figure 4.5 (right) also shows the errors achieved by the cross-validated choice of the number of steps as described in Section 4.3.1. The considered candidate values for the number of steps were $l = 1, \ldots, 5$. While the CV scheme clearly performs well, occasionally outperforming both the two-step and three-step approaches (this is only for small percentages of entries observed and due to the associated variability across different Monte Carlo runs). Still, the improvements are very small, only supporting the straightforward approach of fixing the number of iterations at $l = 2$.

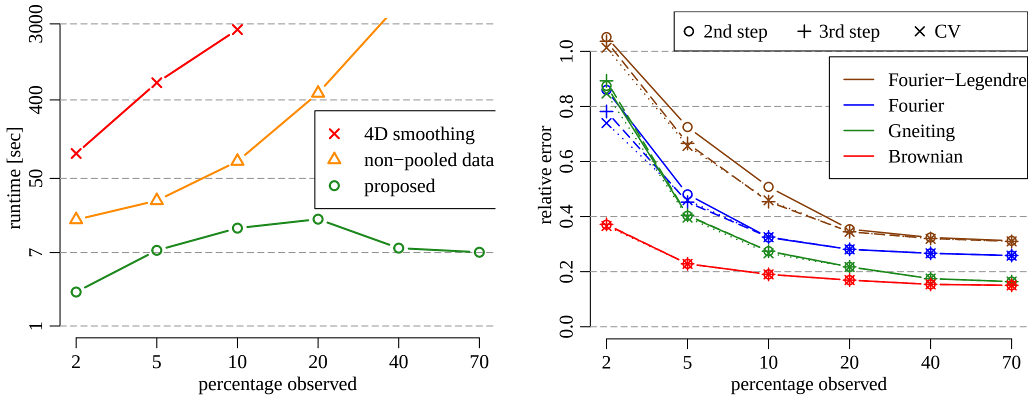
Finally, Table 4.4 shows relative estimation errors of the 4D smoothing approach in all four scenarios used in Section 4.8, but only with the smallest considered percentage $p = 2$. The cross-validated choice of bandwidths is compared against the choice of bandwidth suggested by the separable model (which is used in Section 4.8). It is clear

Table 4.4: Relative estimation errors of 4D smoothing with cross-validated bandwidths and bandwidths suggested by the separable model for the four scenarios considered in Section 4.8 with $p = 2$ percentages of the surfaces observed.

	(a) Fourier	(b) Brownian	(c) Gneiting	(d) Fourier-Legendre
cross-validation	1.07	1.06	1.04	1.39
separable choice	0.04	0.26	1.34	1.1

that cross-validation fails here, because even in the simplest Brownian scenario, cross-validated relative error is larger that one. The reason for that is likely the following. Cross-validation for 4D smoothing, as implemented in the `np` package, does not evaluate its objective function on a grid. In order to reduce the computational burden, the cross-validation objective is optimized is a step-wise manner, until a stopping criterion is met. The stopping criterion is set as a *tolerance* (defaults to 10^{-8}), and the iterative optimization is stopped once both the change in objective value and change in the bandwidths is smaller than the tolerance for two consecutive iterations. While this saves computation time compared to creating a grid over potential bandwidth values and evaluating the objective function in all the grid points, the optimization can get stuck in a local minimum. This is the reason why the cross-validated errors in Table 4.4 are so high. The sampling pattern is very sparse with $p = 2$, and there likely are many local minima. However, the cross-validation still requires fitting the covariance for different values of the bandwidths. We tried to obtain results for larger values of p as well, however we ran out of time (with a single task) at 70 hours with $p = 5$ even with the tolerance decreased to 10^{-2}. Hence, we have no other choice but to use the bandwidths suggested by the separable model also for 4D smoothing.

Figure 4.5: Left: Runtimes for Fourier simulations. **Right:** The proposed approach (two-step) compared to the proposed approach with included third step and with cross-validated number of steps (CV) for all the four scenarios considered in Section 4.8.



4.9 Data Analysis: Implied Volatility Surfaces

A *European call option* is a contract granting its holder the right, but not the obligation, to buy an underlying asset (for example a stock) for an agreed-upon strike price at a defined expiration time. Finding a model and deriving a pricing formula for the fair price of a European call option was a milestone problem in quantitative finance and stochastic calculus. Black and Scholes (1973) and Merton (1973) solved this problem and under the so-called *Black-Scholes-Merton* model they shown that the fair price of the European call option on a non-divident paying asset is given by the *Black-Scholes formula* (Hull, 2006):

$$C_t^{BS}(m, \tau, \sigma_S) = S_t F_{N(0,1)}(d_1) - \kappa e^{-rt} F_{N(0,1)}(d_2), \tag{4.31}$$

$$d_1 = \frac{-\log m + \tau(r + \sigma_S^2/2)}{\sigma_S\sqrt{t}}, \qquad d_2 = \frac{-\log m + \tau(r - \sigma_S^2/2)}{\sigma_S\sqrt{t}},$$

where $m = \kappa/S_t$ is the *moneyness* defined as the ratio of the strike κ and the current underlying asset price S_t at the current time t, $\tau = T - t$ denotes the time to expiration, σ_S is the volatility parameter in the Black-Scholes-Merton model, r is the risk-free interest rate, and $F_{N(0,1)}(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The only unknown quantity among the inputs in (4.31) is the volatility σ_S.

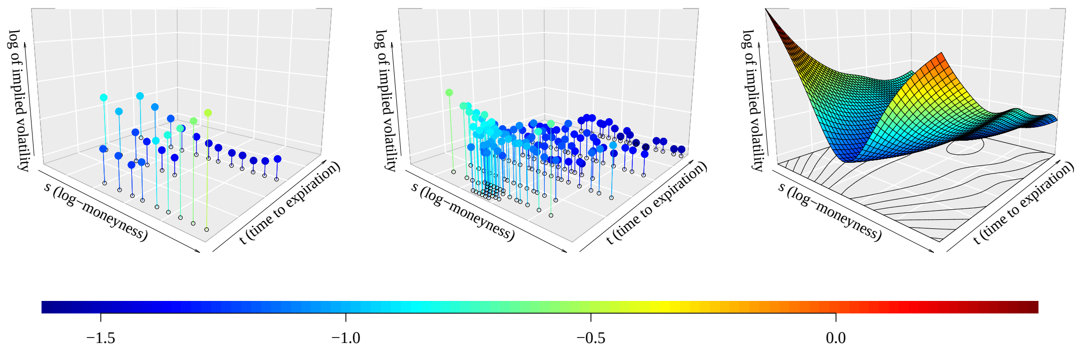
Besides calculating the fair price of an option given the (estimated/realised) volatility, the Black-Scholes formula (4.31) can be used in reverse: having observed the market price of the option, denoted as $C_t^*(m, \tau)$, find the value of $\sigma_t^{IV}(m, \tau)$ that solves the equation

$$C_t^{BS}(m, \tau, \sigma_t^{IV}(m, \tau)) = C_t^*(m, \tau).$$

It can be shown that such value $\sigma_t^{IV}(m, \tau) > 0$, called the *implied volatility*, exists uniquely for each triplet of $m > 0$, $\tau > 0$, and $C_t^*(m, \tau) > 0$. Now, if the market indeed followed the Black-Scholes-Merton model and the investors were rational, the implied volatility $\sigma_t^{IV}(m, \tau)$ for various m and τ would be constant. However, this is not true for real market data pointing to the shortcomings of the Black-Scholes-Merton model. Despite these shortcomings, the Black-Scholes formula (4.31) is widely used for *transforming* the observed option prices into an ensemble of implied volatilities in a bijective manner. The advantage of considering the implied volatilities as opposed to the market prices of the options is that the implied volatility surfaces tend to be smoother and comparable across assets. Thus we can take advantage of the functional data analysis framework.

In contrast to the European options, an *American call option* grants the right to buy the underlying asset anytime until the expiration time T. The pricing of American options on possibly dividend paying stocks is more complicated because the pricing involves the optimal stopping problem. In general, no closed form solution exists and numerical algorithms are required (Cox et al., 1979). Likewise, the observed market option prices can be transformed into implied volatilities.

Figure 4.6: Two sample snapshots of the considered log implied volatility surfaces corresponding to the call options on the stocks of Dell Technologies Inc on 01/19/2006 **(a)** and Qualcomm Inc on 02/07/2018 **(b)**, and the mean surface of the implied volatility gained from pooling all the data together **(c)**.


(a) Dell on 01/19/2006 (b) QCOM on 02/07/2018 (c) mean surface

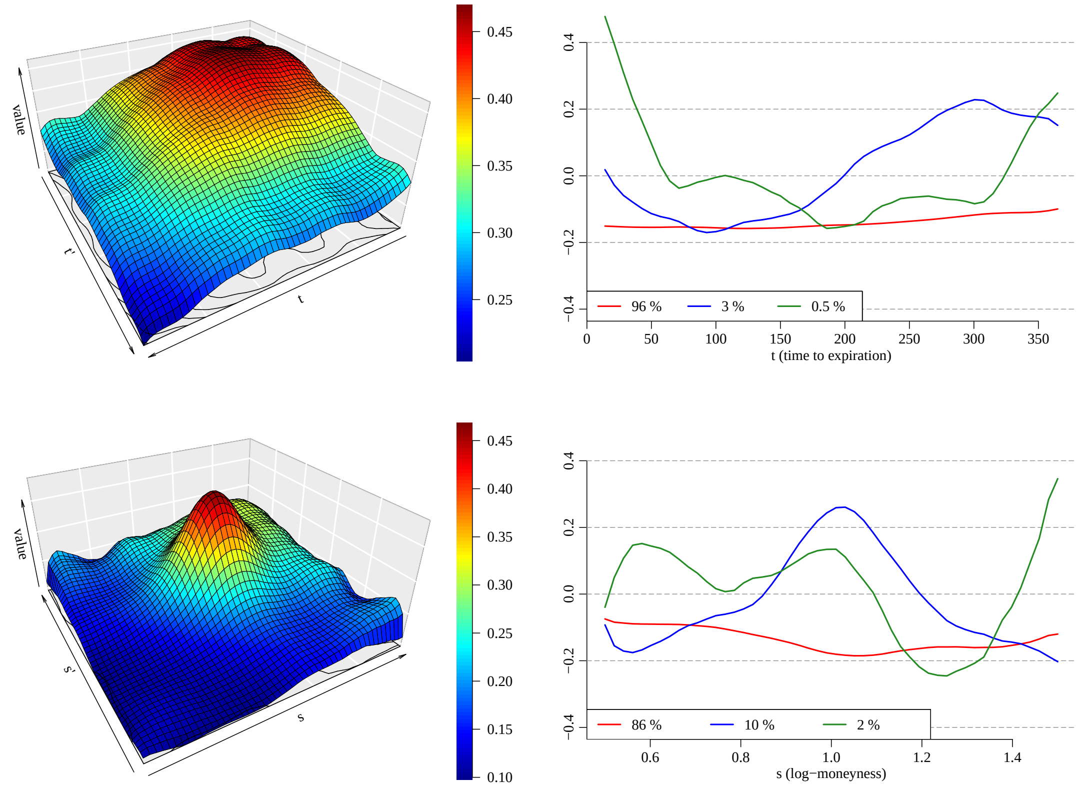
In this section we consider the options data offered by DeltaNeutral[1]. This free data set contains the end-of-day prices as well as the calculated implied volatilities for options on U.S. Equities markets. The data covers the period from January 2003 until April 2019 but limits each month to contain the daily options data on only one symbol (a stock or an index). The currently included symbol changes every month and the options on some of the symbols are American while some are European. For each month we pick randomly only one trading day with the data on the currently available symbol and discard the other trading days. Therefore the sample we analyze contains 196 snapshots with option prices and implied volatilities. We discard the non-liquid options and consider the contracts with the log-moneyness $\log m = \log(K/S_t) \in [-0.5, 0.5]$ and the time to expiration $\tau = T - t \in [14, 365]$ (in days). Moreover, we take the logarithm of the implied volatilities to transform them from the domain $(0, \infty)$ onto the real line. To reduce computational costs we round the log-moneyness and the time to expiration to fall on a common 50×50 grid, cf. Section 4.5.

Figure 4.6 shows two observations in our samples. The snapshot of Qualcomm Inc (QCOM) features a cummulation of observation at the short expiration. Here, it happens five times that two raw observations fall in the same pixel on the common 50×50 grid. In these few cases we calculate the average of the two observations in each pair. Due to smoothness, the option prices (and hence the implied volatilities) attain very similar values and thus this rounding and averaging does not change the conclusions of our analysis. We have observed this fact by fitting the model on finer grids while the estimates remained similar.

---

[1] Available at https://www.historicaloptiondata.com/content/free-data, retrieved on 2020-07-15.

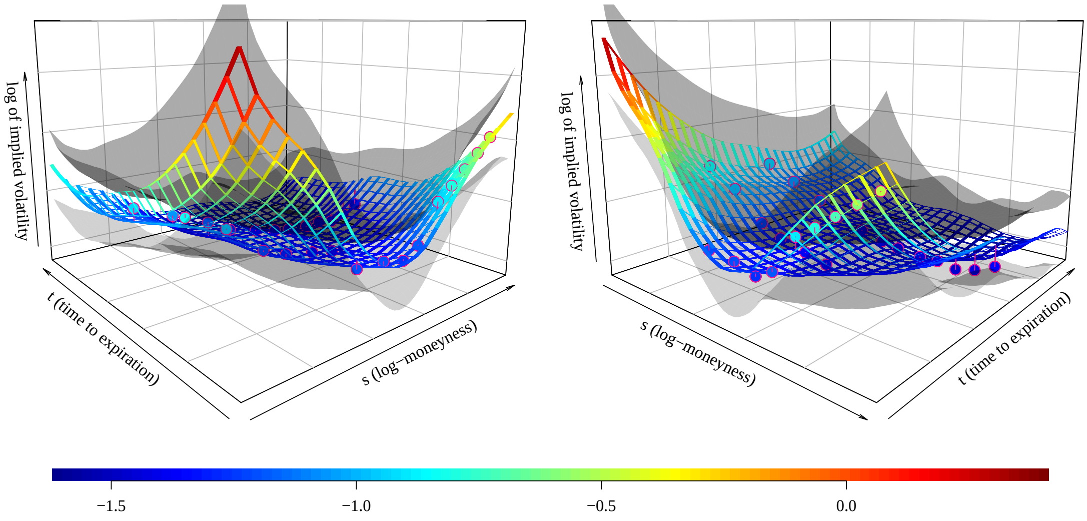
Figure 4.7: Top-left: The estimated covariance kernel $\widehat{a} = \widehat{a}(t, t')$ corresponding to the time to expiration variable. **Top-right:** The three leading eigenfunctions of the spectral decomposition of the covariance kernel $\widehat{a} = \widehat{a}(t, t')$. **Bottom-left and botton-right:** The same as above but for the estimated covariance kernel $\widehat{b} = \widehat{b}(s, s')$ corresponding to the log-moneyness variable.



The estimated mean surface is displayed on the right-hand side of Figure 4.6. The mean surface captures the typical feature of the implied volatility surfaces: the volatility smile (Hull, 2006). The implied volatility is typically greater for the options with moneyness away from 1, while this aspect is more significant for shorter times to expiration.

Figure 4.7 displays the estimates of the separable covariance components by our methodology presented in Section 4.3. The moneyness component demonstrates the highest marginal variability at the center of the covariance surface, meaning that the log implied volatility oscillates the most for the options with the log-moneyness around 0 (i.e. moneyness 1). The marginal variance is lower as the log-moneyness departs from 0. The eigendecomposition plot of the moneyness covariance kernel shows that the most of variability is explained by a nearly constant function with a small bump at the log-moneyness 0. The second leading eigenfunction adjusts the peak at the log-moneyness around 0 to a greater extent than the first eigenfunction. The covariance kernel corresponding to the time to expiration variable is smoother and demonstrates slightly higher marginal variability at shorter expiration. This phenomenon is well known for implied volatility (Hull, 2006). The eigendecomposition of this covariance kernel indicates that the log

Figure 4.8: Two views on prediction based on the call options written on the stock of Dell Technologies Inc on 01/19/2006. The circles depict the available sparse observations, the ribbons depict the predicted latent surface by the method of Section 4.6, where the covariance structure was assumed separable, and finally the transparent gray surfaces depict the 95 % simultaneous confidence band for the latent log implied volatility surface.



implied volatility variation is mostly driven by the constant function while the second leading eigenfunction adjusts the slope of the surface for varying time to expiration.

Figure 4.8 demonstrates our prediction techniques presented in Section 4.6 together with the 95 % simultaneous confidence band. We recall that the confidence band aims to capture the latent smooth random surface itself, while our raw observations are modelled by adding an error term. Therefore, the raw data are not guaranteed to be covered in the confidence band.

4.9.1 Quantitative Comparison

The prediction method outline in Section 4.6 requires as an input the pairwise covariances regardless whether they have been estimated by the separable estimator $\widehat{a}(t,t')\widehat{b}(s,s')$ or the 4D smoother $\widehat{c}(t,s,t',s')$. As the benchmark for our comparison, we choose the locally linear kernel smoother (Fan and Gijbels, 1996) applied individually for each surface as such smoothers constitute a usual pre-processing step. We will refer to this predictor as *pre-smoothing*. In this section, we demonstrate that the predictive performance is comparable for both covariance estimator strategies (the separable and the 4D smoother), and that both of these approaches are superior to pre-smoothing. Moreover, the separable smoother is substantially faster than the 4D smoother.

We compare the prediction error by performing a 10-fold cross-validation, where the covariance structure is fitted always on varying 90 % of the surfaces, with the remaining

10 % used for out-of-sample prediction. In the set that is held out for prediction, we select some of the sparse observations and predict them based on the remaining observations on that surface. We use the following hold-out patterns:

(a) *Leave one chain out.* Since the options are quotes always for a range of strikes, they constitute features known as option chains (cf. Figure 4.6) where multiple option prices (or equivalently implied volatilities) are available for a fixed time to expiration. For those surfaces that include at least two such chains, we remove gradually each chain and predict it based on the other chains. Therefore the number of prediction tasks performed on a single surface is equal to the number of chains observed per that surface.

(b) *Predict in-the-money.* Predict implied volatilities for below-average moneyness (i.e. moneyness $m \leq 1$) based on the out-of-the-money observations (moneyness $m > 1$).

(c) *Predict out-of-the-money.* Predict implied volatilities for above-average moneyness (i.e. moneyness $m \geq 1$) based on the in-the-money observations (moneyness $m < 1$).

(d) *Predict short maturities.* Predict the implied volatility for options with the time to maturity $\tau < 183$ [days] based on the implied volatility of the options with the time to maturity $\tau \geq 183$ [days].

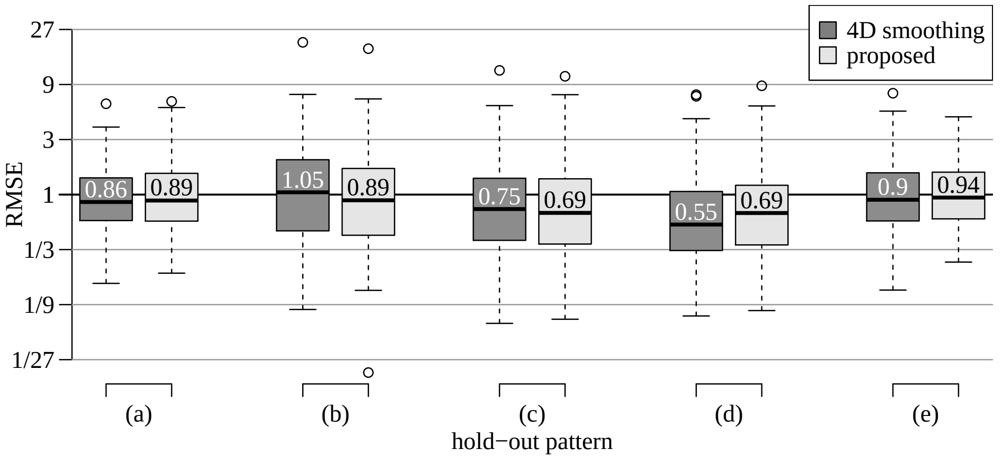
(e) *Predict long maturities.* Predict the implied volatility for options with the time to maturity $\tau > 183$ [days] based on the implied volatility of the options with the time to maturity $\tau \leq 183$ [days].

All the prediction strategies are performed only for those surfaces where both the discarded part and the kept part are non-empty. We measure the prediction error on surface with the index n (in the test partition within the K-fold cross-validation) by the following root mean square error criterion, relative to the pre-smoothing benchmark:

$$RMSE^{\text{method}}(n) = \sqrt{\frac{\sum_{m \in M_n^{\text{discarded}}} \left((\widehat{\Pi}^{\text{method}}(X(t_{nm}, s_{nm}) | \mathbb{Y}_n^{\text{kept}})) - Y_{nm} \right)^2}{\sum_{m \in M_n^{\text{discarded}}} \left((\widehat{\Pi}^{\text{pre-smooth}}(X(t_{nm}, s_{nm}) | \mathbb{Y}_n^{\text{kept}})) - Y_{nm} \right)^2}} \tag{4.32}$$

where $Y_{nm}, m = 1, \ldots, M_n$ are the implied volatility observations on the n-th surface, $M_n^{\text{discarded}} \subset \{1, \ldots, M_n\}$ denotes the set of observations' indices discarded for the n-th surface, $\mathbb{Y}_n^{\text{kept}}$ are the vectorized implied volatility observations that were kept to be conditioned on. The predictor $\Pi^{\text{pre-smooth}}$ denotes the pre-smoothing based on the observations $\mathbb{Y}_n^{\text{kept}}$). The predictors Π^{method} for method being either the separable smoother or the 4D smoother constitute the proposed predictors in this article with the covariance structure estimated by either of the two smoothers. These predictors are always trained only on the training partition (90 % of the surfaces) within the 10-fold cross-validation scheme. Note that this out-of-sample comparison is adversarial

Figure 4.9: Boxplots of root mean square errors relative to the pre-smoothing benchmark (4.32) for the prediction method of Section 4.6 with the covariance estimated by 4D smoothing or the proposed separable approach, and different hold-out patterns: **(a)** leave one chain out; **(b)** predict in-the-money; **(c)** predict out-of-the-money; **(d)** predict short maturities; and **(e)** predict long-maturities. Numbers inside the boxes provide numerical values of the median. For a given method, RMSE value 3 means that the given method is 3-times worse than the benchmark, while RMSE value 1/3 corresponds to a 3-fold improvement.



to the proposed approach, because when predicting a fixed surface, the measurements on that surface (and another 10% of measurements total) are not used for the mean and covariance estimation. In practice, we naturally utilize all available information. However, for the hold-out comparison study here, that would require frequent re-fitting of the covariance, which would not be computationally feasible, in particular for the 4D smoothing approach.

Figure 4.9 presents the results under the five hold-out patterns in form of boxplots created from the relative errors (4.32). We see that the prediction errors based on estimated covariances, be it the separable smoother or the 4D smoother, are typically smaller than the pre-smothing benchmark with the only exception of the 4D smoother in the hold-out pattern (b). The predictive performances of the separable and the 4D smoother are comparable, but they differ a lot in terms of runtime. It typically takes 30 seconds to calculate the separable smoother (including a cross-validation based selection of the smoothing bandwidths), while the 4D smoother takes around 3 hours. The latter runtime is moreover without considering any automatic selection of the bandwidths, because such would be computationally infeasible. Hence we use the bandwidths selected by the separable model, cf. Sections 4.5 and 4.8. The calculations are performed on a quite coarse grid of size 20 × 20. The calculations on a dense grid, such as 50 × 50 used in the qualitative analysis, c.f. Section 4.9, are not feasible for the 4D smoother.

Therefore, we conclude that the separable smoother approach enjoys a better predictive performance than the pre-smoothing benchmark and – while having having similar predictive performance as the predictor based on 4D smoothing – is computationally much faster than the said competitor. In fact, it requires two-dimensional smoothers only, just as the pre-smoothing benchmark.

Discussion and Future Directions

In practice, covariances are often non-separable (Aston et al., 2017; Bagchi and Dette, 2020; Constantinou et al., 2017) and assuming separability thus introduces a bias. The immense popularity of the separability assumption stems mainly from the computational advantages it entails. This thesis proposes several alternatives to separability, which enjoy the same computational advantages. Firstly, the separable-plus-diagonal model can be used to estimate a separable covariance from noisy observations without smoothing, which is useful especially when the noise is heteroscedastic and smoothing may not perform well. Secondly, the separable-plus-banded model may be useful when short-scale noise or additional weakly dependent signal is present. Finally, an R-separable covariance can be used to fit any covariance as long as the number of samples is high enough. While both the separable-plus-banded model and the notion of R-separability suggest ways in which estimation beyond separability can be conducted, they do so in a different spirit. The separable-plus-banded model is expected to hold if separability is violated only locally, i.e. via a short-range entanglement. R-separability, on the other hand, is expected to be useful whenever there are numerous separable effects overlaid in the data, e.g. when a process is a mixture of several processes with separable individual covariances.

The marginalization operators developed to estimate the proposed covariance structures, namely shifted partial tracing and the partial inner product, are interesting even beyond the scope of this thesis. Firstly, following Aston et al. (2017), partial tracing has become the method of choice for calculation of the marginal kernels. These marginal kernels are either used to form a separable proxy for the covariance or as building blocks for different models. For example, the partially traced marginal kernels are cornerstones of the weakly separable model of Lynch and Chen (2018). However, given the theoretical development and the practical evidence in this thesis, it seems that shifted partial tracing should be generally preferred due to its denoising properties. Secondly, understanding separability as an unconventional form of low-rankness, which can be exploited in a power iteration scheme consisting of a sequence of partial inner products, has many implications. One of these implications is the development of the methodology to estimate a separable covariance from sparse and noisy measurements. Even though the development of Chapter 4 is largely self-contained, it is needless to say that the estimation procedure is inspired by the generalized power iteration method of Chapter 3. In fact, one can understand a single step in the estimation procedure as a sparsified version of the partial inner product. Related and additional implications are discussed below.

While the computational emphasis of this thesis may seem excessive at first glance, computational issues are the main reason for considering separability in the first place. The merit of this computational emphasis quickly becomes apparent once one analyzes a large enough ensemble of random surfaces. For example, a single-electrode EEG datum sampled at 256 Hz can be transformed to a random surface via a short-time Fourier transform. This is an alternative way to how random surfaces can naturally arise when working with EEG data, and it is useful mostly when working with resting-state data, as opposed to the event-related potentials of Chapter 3. The standard is to keep frequencies lower than 60, leading to random surfaces with the spatial resolution $K_2 = 60$. When the temporal length of the signal is $T \in \mathbb{N}$ seconds, one runs out of memory with an unstructured covariance already when $T > 1$, while resting-state EEG signals are typically sampled over the course of several minutes. In general, brain imaging applications are abundant with densely sampled random surfaces, which can be very large compared to the one studied in Chapter 3. While the memory requirements are the main concern in the case of densely sampled data, runtime may cause the same concerns in the case of sparsely observed data, as demonstrated in the analysis of the implied volatility surfaces in Chapter 4.

From a high-level perspective, this thesis is mostly methodological, showing that there are many appealing alternatives to the standard and quite ubiquitous assumption of separability, which enjoy similar favorable computational properties. Abstracting from separability, this thesis suggests to model covariances using a superposition structure, where the superposed terms offer complementary forms of parsimony. There likely exist other forms of parsimony (which can be exploited in a similar manner) as well as other operators than shifted partial tracing, which have other than banded operators in their kernel, and thus can be used to deconvolve the terms in the superposition and allow for their isolated estimation. Before conducting search for such other models or covariance structures, however, it would be meaningful to verify usefulness of the already explored generalizations in more applied contexts.

Conceivable applications apart, we discuss several potential directions of future work and our final thoughts on them below. Many of these directions seem to be promising, and we may explore them in the future, while others seem to be blind alleys at the moment, but they still arise as very natural continuations of the presented work.

Matrix Completion

Descary and Panaretos (2019) proposed to model the covariance of a one-dimensional process as a superposition of two parts, a low-rank part and a banded part, with the following motivation in mind. The low-rank part of the covariance should capture the smooth, global behavior of the process, while the banded part should capture rough, local variations. The empirical covariance matrix was considered, its band of a certain size was deleted, and matrix completion techniques were utilized to reconstruct the low-rank part of the covariance, estimating the banded part in a subsequent step.

The basic idea of our model is similar to that of Descary and Panaretos (2019), allowing for a short-term structure atop some base structure. Thus one could say that our work generalizes that of Descary and Panaretos (2019) to higher dimensions, utilizing separability to achieve computational efficiency. In fact, one could marginalize via partial tracing (without shifting), and then use the matrix completion approach of Descary and Panaretos (2019) to estimate our separable-plus-banded model. However, matrix completion requires much stronger assumptions (namely analyticity and low-rankness – in the usual sense – of the separable part). Instead, we advocated for methodology that requires much weaker assumptions and can be implemented by means of careful use of linear operations. Still, the matrix completion approach of Descary and Panaretos (2019) can be useful in more general model structures, when combined with the methodology from this thesis.

Firstly, separability corresponds to a Kronecker product structure in finite-dimensional spaces, where the longing for non-separability is sometimes facilitated via a Kronecker sum model instead of the Kronecker product model (Park et al., 2017; Greenewald et al., 2019). A functional version of the Kronecker sum structure of the covariance could be $C = C_1 \tilde{\otimes} U + V \tilde{\otimes} C_2$, with U and V being banded. Consider adding this structure into the separable-plus-banded model as

$$C = A_1 \tilde{\otimes} A_2 + B + C_1 \tilde{\otimes} U + V \tilde{\otimes} C_2$$

in order to obtain a more general model, including both Kronecker product (i.e. separability) and Kronecker sum as special cases in finite dimensions. Assuming that U and V as well as B are banded by δ, we have

$$\mathrm{Tr}_1^{\delta}(C) = \mathrm{Tr}^{\delta}(A_2) A_1 + \mathrm{Tr}^{\delta}(C_2) V.$$

Now, one can deconvolve A_1 and V by the matrix completion approach of Descary and Panaretos (2019). Similarly, A_2 and U can be estimated by shifted partial tracing w.r.t. the second argument, followed by matrix completion using the previously obtained estimators of A_1, A_2, U and V. Secondly, C_1 and C_2 can be estimated by (non-shifted) partial tracing followed by matrix completion. Finally, B can be estimated as the remainder. Thus shifted partial tracing can be potentially combined with the matrix completion methodology of Descary and Panaretos (2019) to facilitate estimation of a model, which generalizes separability even further.

Secondly, following the development of Chapter 3, (R-)separability can be understood as an alternative form of low-rankness. Hence an R-separable-plus-banded model could be potentially tackled by matrix completion too. This goal seems to be entirely reachable on the theoretical level. However, there is an important computational question: would it be possible for a specific matrix completion algorithm to work directly on the level of data? In other words, would it be possible to avoid again the construction of the empirical estimator in order to be computationally as efficient as when working with a separable model?

We believe these generalizations could be particularly interesting with a suitable data set at hand.

Weak Separability and Weak Bandedness

The separable-plus-banded model of Chapter 2 can be generalized in two ways, which are both relatively straightforward and potentially interesting. The first is to replace separability with weak separability of Lynch and Chen (2018). The second is to weaken our bandedness assumption and utilize the matrix completion approach of Descary and Panaretos (2019).

It follows from Lemma 1 that a separable covariance $A_1 \tilde{\otimes} A_2$ has eigendecomposition

$$A_1 \tilde{\otimes} A_2 = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \lambda_j \rho_k (e_j \otimes f_k) \otimes_2 (e_j \otimes f_k),$$

where $A_1 = \sum \lambda_j e_j \otimes_2 e_j$ and $A_2 = \sum \rho_j f_j \otimes_2 f_j$ are eigendecompositions of the marginal operators. Thus we see that separability is characterized by two conditions: all eigenfunctions have a product structure (i.e. they are themselves separable), and the array of eigenvalues $\Gamma = (\gamma_{jk})_{j,k=1}^{\infty}$, where $\gamma_{jk} = \alpha_j \beta_k$, must be rank-1. The weakly separable model of Lynch and Chen (2018) assumes the product structure of the eigenfunctions, but allows for the array of eigenvalues to have a rank higher than one. Such a model can be estimated by utilizing a marginalization operator (Lynch and Chen, 2018, opt for partial tracing) on the empirical covariance to estimate the eigenfunctions, and then estimating the eigenvalues in the second step as $\gamma_{jk} = \langle \widehat{C}_N, e_j \otimes f_k \rangle$ for $j, k = 1, 2, \ldots$.

By utilizing shifted partial tracing as the marginalization operator, one can afford to estimate the weakly-separable-plus-banded model $C = A + B$, where A is weakly separable and B is banded. Such a model can be again estimated and manipulated with ease comparable to that of a separable model. Notably, numerical inversion of such a covariance can be achieved by the ADI scheme of Section 2.3.3 coupled together with the PCG scheme of Section 3.4. Specifically, when inverting the weakly-separable-plus-banded model, the separable equation, i.e. the first equation in (2.35), is now only weakly separable and no longer has a closed form solution. However, one can utilize the PCG scheme of Section 3.4 to find a solution iteratively. The overall inversion algorithm will thus consist of the outer ADI scheme and two embedded PCG schemes. Even though this may seem complicated, implementation of such an algorithm is trivial based on the development in this thesis. And again, one can also use shifted partial tracing to estimate a weakly separable model observed under the heteroscedastic white noise.

Moreover, the bandedness assumption can also be weakened. Our assumption of bandedness of B is manifested on the level of its kernel as follows. We assume that $b(t,s,t',s')=0$ whenever $|t-t'|>\delta$ *or* $|s-s'|>\delta$, which is a strong form of bandedness, assuming two locations are separated either in space *or* in time in order for the banded kernel to vanish. In particular, bandedness needs to hold marginally both in space and time. A weaker form of the assumption can be obtained by replacing the logical conjunction for the intersection:

$$b(t,s,t',s')=0 \quad \text{if} \quad |t-t'|>\delta \text{ and } |s-s'|>\delta.$$

With this weaker assumption, two locations have to be separated both in space and in time in order for the banded part to vanish. In particular, bandedness may not hold marginally in neither dimension. On the other hand, (non-marginal) weak dependence in time only suffices for this weak bandedness to hold. Such a model could be particularly interesting for applications where time is the important component, and could be useful even when there is no linear ordering in the spatial dimension, e.g. in portfolio optimization where time series of prices of different assets on a market are observed (Chen et al., 2020a).

Obviously, shifted partial tracing is not sufficient for bypassing the banded part of such a separable-plus-weakly-banded model. However, under the assumptions of analyticity and low-rankness on the separable constituents A_1 and A_2, one can marginalize the covariance using shifted partial tracing, and subsequently perform matrix completion in both the marginalized dimensions to deconvolve the separable constituents from the weakly banded kernel, which must be banded (in the strong and only sense) after marginalization.

Sparsely Observed Surface-valued Time Series

The analysis of implied volatility surfaces in Chapter 4 provides some insights into the statistical dependencies of such sparsely observed data, and our quantitative comparison demonstrates the prediction performance of our prediction method over the pre-smoothing benchmark. We have formed our data set collecting options across various symbols and timestamps with the reasoning that the volatility across these come also from one population, which is debatable. This study should be seen rather as a proof-of-concept on how to “borrow strength” across the data set made available by DeltaNeutral, and how this approach can be used to decrease the prediction error.

We believe that our methodology can provide even better results when considering a more homogeneous population such as the time series of the implied volatility surfaces related to a single fixed symbol/asset. In this case, pre-smoothing is usually required (Cont and Da Fonseca, 2002; Kearney et al., 2018) for forecasting of such time series. Our methodology could avoid the pre-smoothing step in this case by predicting the surfaces while borrowing the information across the entire data set. Furthermore, our methodology could be easily tailored to predicting principal components scores by conditional expectation (similarly to Yao et al., 2005a), which could be used for forecasting by a vector autoregression.

Combining separable covariances and time series of sparsely observed surfaces hints at other directions of future work. For example, Rubín and Panaretos (2020) showed how to estimate the spectral density operator non-parametrically from sparsely observed functional time series. Estimating a separable spectral density operator for sparsely observed surface-valued time series seems to be within reach and is likely to provide predictions that benefit from the information across time, and thus reducing the prediction error even more.

Statistical Tests

Lately, multiple tests of separability were developed in the FDA context, two notable instances being the tests of Aston et al. (2017) and Bagchi and Dette (2020). Aston et al. (2017) work in the trace-norm topology and utilize partial tracing to construct a separable proxy of the covariance, while Bagchi and Dette (2020) operate in the Hilbert-Schmidt topology and implicitly use the partial inner product.

The Hilbert-Schmidt topology facilitates more straightforward analysis. The asymptotic distribution of the test statistic leads to an asymptotic test, while a bootstrap test is also proposed. The asymptotic test is reportedly more powerful, while the bootstrap test is reportedly much faster. The latter is true because bootstrap avoids calculation of the asymptotic variance, which is very costly. But still, the bootstrap test of Bagchi and Dette (2020) is significantly slower than the bootstrap test of Aston et al. (2017), which appears to be (based on the simulation results reported by Bagchi and Dette, 2020) the major advantage of the latter. The computational efficiency of the bootstrap test of Aston et al. (2017) is achieved by dual means: 1) bootstrap is used again to avoid the expensive calculation of the asymptotic variance, 2) projection of the test statistic onto the leading principal components (both in space and time) is considered. On the other hand, the principal component projection, which allows for cheaper computation, can be criticized from a conceptual perspective: the tests of Aston et al. (2017) do not test separability in full, they only test it on a subspace, whose size has to be chosen by the user. Moreover, this subspace is a proxy to the principal subspace only if the hypothesis of separability holds. Also, if the first few eigenfunctions happen to be separable, the test by design tends to conclude that separability holds. While this issue is rather philosophical, the test of Bagchi and Dette (2020) could be preferable simply as a tuning-free alternative, but – in our personal experience – the test of Aston et al. (2017) is both faster and more powerful, likely due to the projection-based regularization.

The test statistic of Bagchi and Dette (2020) happens to be the Hilbert-Schmidt norm of the difference between the empirical covariance and the separable proxy of the covariance obtained by a single step of the generalized power iteration method. Chapter 3 of this thesis offers two ways in which the test of Bagchi and Dette (2020) could be improved. Firstly, one could replace the one-step proxy of Bagchi and Dette (2020) by the fully iterated proxy of Chapter 3. This would complicate the theory of Bagchi and Dette (2020), which would basically have to be replaced by some version of principal component asymptotics, while the performance of the resulting test would likely remain similar (judging based on the fact that the performance of the test of Bagchi and Dette, 2020, is reported to be very similar regardless of the starting point). On the other hand, the test statistic described above is the sum of all the separable component scores barring the first one, which corresponds to separability. This sum is zero if and only if the covariance is separable, but also if and only if the second score is zero. Replacing the test statistic by the estimated magnitude of the second score would introduce an alternative regularization to the problem, which could potentially improve the performance. Moreover, this proposed test could be easily extended to test R-separability.

Finally, Chapter 4 of this thesis develops sparsified version of the partial inner product. In the case of the bootstrap test of Bagchi and Dette (2020), the partial inner product operation is essentially all one needs to perform the test. Hence we should be immediately able to test separability also for sparsely observed random surfaces. On the other hand, finding the asymptotic variance of the statistic (which is not needed to perform the test, but rather to show its validity) based on sparsely observed data would require a substantial work.

Domains of Higher Dimensions

A very natural question is whether our methodology can be generalized to more than two-dimensional domains, for example when there is one temporal and two (or more) spatial dimensions. This question has to be addressed separately based on whether separability should be manifested between the spatial dimensions. Differences also appear between different methodologies proposed in this thesis.

Firstly, suppose that we seek no separation between the spatial dimensions. In that case, the separable component decomposition, for example, can be utilized directly after vectorization of the spatial domain. Note that this is possible due to the fact that our methodology has a peculiar relationship with continuity. We assume all data are coming in as matrices, but we think about some latent continuous surfaces from which the discrete measurements are sampled. Yet, apart from the sparse observations in Chapter 4, continuity is not needed for our methodology. The methodology can be in fact used as a multivariate methodology and the presented theory is valid as well. Only once we wish to relate the inherently discrete estimators with the latent objects on continuous domains, the functional aspect comes into play. And even at this point, the assumption of continuity can be largely avoided, depending on the sampling scheme relating the discrete measurements and the latent objects. Given how heavily the methodology developed in this thesis is inspired by the FDA viewpoint, avoiding continuity may seem like a useless mathematical exercise at first glance. However, there are occasional benefits of avoiding the continuity assumption. For example, when we have one temporal and two spatial dimensions, it is easy to linearly order the spatial dimensions, use the methodology in this thesis exactly as it stands, and transform back the results. This is possible for the separable component decomposition or the separable-plus-diagonal model, but not for the separable-plus-banded model in general, because bandedness is effectively lost after linearization of the domain. On the other hand, the methodology of Chapter 4, designed to handle a sparsely observed separable model, should be straightforward to generalize without the need to transform the spatial domain.

Secondly, assume that separability should hold between all the dimensions, i.e. for example the separable-plus-banded model would look like

$$\begin{aligned} C &= A_1 \tilde{\otimes} A_2 \tilde{\otimes} A_3 + B \\ c(t,s,x,t',s',x') &= a_1(t,t')a_2(s,s')a_3(x,x') + b(t,s,x,t',s',x'). \end{aligned}$$

In this case, problems can arise wherever there is an iterative scheme. For example, the generalized power iteration method may not be guaranteed to converge or it may converge but not to the global minimizer of (3.12), because it is a well-known fact in the literature on tensor decompositions (see Kolda and Bader, 2009, for a review) that alternating minimization schemes are not globally convergent. This may not pose an issue per se, we have no reason to believe that the resulting estimator of the generalized power iteration method in more than two dimensions would be poor for practical purposes. However, such an estimator, which would neither have an explicit representation nor would it satisfy some optimality conditions, might prove hard to be analyzed from the theoretical perspective. Also, it is hard to anticipate what would be the behavior of the iterative inversion algorithms proposed in Sections 2.3.3 and 3.4. On the other hand, there should be again no issues with the separable-plus-banded model of Chapter 2 and the sparsely observed separable model of Chapter 4.

Links between the Proposed Methods

There are obvious links between the three main methodological developments of this thesis: estimation of a separable-plus-banded model, an R-separable covariance, and a sparsely observed separable model.

Firstly, the impact of noise in the estimation procedure for an R-separable covariance is described in Section 3.5.3. Could we incorporate some form of shifting in the generalized power iteration method to suppress noise in the same way shifted partial tracing does, when working with a separable-plus banded model? Could we more generally estimate an R-separable-plus-banded model? The answer is, sadly, negative. While we could discard the diagonal (or around-diagonal) raw covariances in every iteration (like we do in Chapter 4), this would lead to loss of orthogonality between the separable components. Also, the resulting algorithm would no longer have an optimization formulation, so it would not be provably convergent.

Secondly, could we iterate the estimation procedure in Chapter 4 until convergence, then somehow deflate the raw covariances, and iterate again to effectively have a way of estimating an R-separable covariance from sparsely observed data? The answer is, again, negative for similar reasons. With every smoothing step, such an algorithm jumps away from the Krylov space, and hence iterating until convergence may not work. In Chapter 4, we do not have to iterate at all due to separability being assumed as a model. Since iteration is in principle not needed, we have consistent estimators after a single iteration. In practice, it makes sense to iterate twice: to estimate the weights consistently in the first step, and then use the consistent weights in the second step. The theory of Chapter 4 works for any fixed number of iterations, but is doesn't suggest that opting for more iterations is better, since everything is hidden in the constants (in the big O notation). We can say that when we fix the number of steps, say at $l \in \mathbb{N}$, and let the number of samples grow to infinity, the l-th iteration of the sparse scheme converges to the l-th iteration of the generalized power iteration method of Chapter 3. But for reasons described above, it is impossible to show uniformity across the iterations. The underlying reason is that in Chapter 4 we have to assume a model (separability), while in Chapter 3 we do not. The estimators in Chapter 3 are M-estimators, while the estimators in Chapter 4 are moment estimators, which are simpler and work due to separability being explicitly assumed.

It would be very interesting to see if the methodology in this thesis could be recast into a reproducing kernel Hilbert space setup, using another form of smoothing (with a fixed basis). There, optimization formulation could possibly be found even in the sparse case (cf. Cai and Yuan, 2010; Wong and Zhang, 2019; Wang et al., 2020), and one could potentially obtain a truly sparsified version of the methodology developed in Chapter 3.

Partial Tracing vs. Partial Inner Product

Assume we have densely observed random surfaces on a grid and want to estimate a separable proxy for the covariance. Should we adhere to (shifted) partial tracing or to the generalized power iteration, i.e. the partial inner product? Partial tracing has the advantage of being (almost) linear while the generalized power iteration has certain optimality properties, which is why some authors (e.g. Genton, 2007) speak of the best separable approximation. Actually, it would seem that the best separable approximation should outperform the partial tracing proxy just because of its name. However, the separable approximation obtained via generalized power iteration is *best* with respect to the empirical covariance, not the true covariance.

In our experience with simulations (not all simulation we conducted are reported in this thesis), the partial tracing proxy is superior when the true covariance is close to being separable (in terms of very small separable component scores), and there is no noise in the data. When there is noise, shifted partial tracing is superior as long as the true covariance is close to being separable and smooth enough (relatively to the grid size). On the other hand, if there is noise and the covariance is rough and noise level mediocre, the best separable approximation tends to outperform a (shifted) partial tracing proxy. But in practice, when the ground truth is unknown, it is hard to make a suggestion about which estimator to use.

Could we answer the question of partial tracing vs. the partial inner product at least theoretically? Working on a grid, (shifted) partial tracing can be regarded as a special case of the partial inner product (see Section 1.7.3), and even in the continuum (i.e. in theory) it can be regarded as its limiting case, cf. (1.12). As shown in Chapter 4, the rates of convergence are always the same under separability, regardless of the starting point. Hence our analysis focusing on rates of convergence cannot really answer this question. While we have made some attempts to answer these questions based on the fourth order dependence structure (not included in this thesis), we were not successful.

Pseudo-MLE Approach

At the end of Chapter 1, we introduced the MLE algorithm for matrix-variate Gaussian distribution and hinted its conjunction with the generalized power iteration on a finite-dimensional domain. Basically, one obtains the MLE algorithm by taking $R = 1$ and replacing the second arguments (the weights) in the partial inner products by their (operator) inverses in Algorithm 3.1 (Section 3.2.1). In infinite-dimensional spaces, such inverses are naturally not Hilbert-Schmidt operators. However, we could generalize the partial inner product to require a trace-class operator in the first argument and only a bounded operator in the second argument.

For an arbitrary Hilbert space $\mathcal{H}$ and a conjugate pair of indices p and q, we can define Schwartz' duality product $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ as the linear functional on $\mathcal{S}_p(\mathcal{H}) \times \mathcal{S}_q(\mathcal{H})$ satisfying

$$\langle\!\langle A, B \rangle\!\rangle = \sum_{j=1}^{\infty} \langle A e_j, B e_j \rangle_{\mathcal{H}}\,, \quad A \in \mathcal{S}_p(\mathcal{H}_1), B \in \mathcal{S}_q(\mathcal{H}_1)\,,$$

where $\{e_j\}_{j=1}^{\infty}$ is an orthonormal basis in H_1. See Unser and Tafti (2014) for details. Schwartz' duality product is a generalization of the inner product. Interestingly, the development of Section 3.2 can be repeated with inner products replaced by Schwartz' duality products, resulting for $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ in the partial duality products

$$S_1 : \mathcal{S}_p(\mathcal{H}) \times \mathcal{S}_q(\mathcal{H}_2) \to \mathcal{S}_p(\mathcal{H}_1) \quad \& \quad S_2 : \mathcal{S}_p(\mathcal{H}) \times \mathcal{S}_q(\mathcal{H}_1) \to \mathcal{S}_p(\mathcal{H}_2)$$

Both partial tracing ($p = 1$ and $q = \infty$) and the partial inner product ($p = q = 2$) are special cases of this partial duality product. The MLE algorithm would also be a special case of generalized power iteration with the partial inner product replaced by the partial duality product (with $p = 1$ and $q = \infty$), if inverses of trace-class operator were bounded. That is naturally not the case, but one can envision a pseudo-MLE algorithm with regularized inverses used instead, to ensure boundedness.

The two common approaches to regularization of the inverse problem in FDA are the ridge regularization and spectral cut-off. In both cases, the hinted pseudo-MLE algorithm can be implemented easily as a modification of Algorithm 3.1, the spectral cut-off regularization in particular corresponds to using the Moore-Penrose pseudoinverse instead of the standard inverse. While we observed some encouraging simulation results with *adaptive* ridge regularization (not reported in this thesis), the conceptually simpler approach involving the pseudoinverse usually breaks down convergence of the generalized power iteration method.

Mixed Sampling of Random Surfaces

The methodology of Chapter 4 is aimed at the sparse sampling regime, as reflected by the theory presented in Section 2.4. It is true that, in practice, it can be hard to distinguish whether data are observed sparsely or which of the other asymptotic regimes could be more appropriate. Furthermore, some surfaces can be sampled rather densely, while others sparsely, which we call a *mixed* regime. While a mixed regime does not prevent one from using the methodology of Chapter 4 per se, a natural question arises whether the fact that a single observation on a very sparsely observed surface contains more information compared to a single observation on a more densely observed surface should be taken into account.

Based on this motivation, Li and Hsing (2010) were the first to consider a unifying approach to sparse, dense, or mixed (one-dimensional) functional data sets. Their estimators are based on kernel smoothing with an additional weight factor corresponding to the number of points being smoothed at each functional datum. While the weighting adopted by Yao et al. (2005a), which is also the one we use, puts equal weight per observation (per a single point), the weighting scheme of Li and Hsing (2010) puts equal weight per subject, i.e. down-weighting observations on densely observed curves. The resulting convergence rates of Li and Hsing (2010) exhibit different asymptotic behavior depending on the sparse vs. dense distinction. Surprisingly, however, the SUBJ weighting scheme of Li and Hsing (2010) is in practice outperformed by the OBS weighting scheme of Yao et al. (2005a), regardless of the sampling regime (Zhang and Wang, 2016, 2018).

Another line of work aiming at a unified theory for different asymptotic sampling schemes is due to Zhang and Wang (2016), who assume that the number of measurements per subject (a curve) is fixed, but arbitrary. They use again kernel smoothers, but this time with a general per-subject weights. Under some technical conditions on the numbers of measurements per subjects, they obtain asymptotic rates of convergence depending on these numbers as well as the per-subject weights, which are free to be specified. In fact, under the sparse sampling scheme of Yao et al. (2005a), taking equal weights per observation (OBS weighting scheme) reduces the results of Zhang and Wang (2016) to match those of Yao et al. (2005a), while taking equal weights per subject (SUBJ weighting scheme) reduces their results to those of Li and Hsing (2010). On the other hand, considering the general results of Zhang and Wang (2016), a valid strategy (regardless of the measurement design) is to specify the weights per subject such that the asymptotic rates of convergence are optimized, which has been suggested in case of the $\mathcal{L}^2$-rates in the follow-up paper of Zhang and Wang (2018).

A full adaptation of the theory of Zhang and Wang (2016) to our multi-dimensional case under separability would be hard to achieve. However, it might be worthwhile to see whether our methodology could be improved using similar ideas. For that, we need to find the asymptotic rates under a general sampling scheme, and optimize them to obtain the per-surface weights. The additional difficulty compared to Zhang and Wang (2016) lies in the fact that we already have an exogenous per-observation weighting scheme present in our methodology (and also in multi-dimensionality and multi-step nature of the estimation scheme).

Assume we know $b = b(s, s')$ and we are estimating only $a = a(t, t')$. Consider the 2D smoother (4.5), with explicit per-surface weights v_n and per-measurement-within-surface weights $b_{nmm'} = b(s_{nm}, s_{nm'})$:

$$\begin{aligned}(\widehat{\gamma_0}, \widehat{\gamma_1}, \widehat{\gamma_2}) = \underset{\gamma_0,\gamma_1,\gamma_2}{\arg\min} \sum_{n=1}^{N} v_n \sum_{m \neq m'} b_{nmm'} K\left(\frac{t - t_{nm}}{h}\right) K\left(\frac{t' - t_{nm'}}{h}\right) \cdot \\ \cdot \left[\operatorname{sign}(b_{nmm'}) G_{nmm'} - \gamma_0 - \gamma_1(t - t_{nm}) - \gamma_2(t' - t_{nm'})\right]^2, \end{aligned} \tag{4.33}$$

leading to $\widehat{a} = \widehat{\gamma_0}$. Carefully inspecting the proofs of Theorem 5.1 and Lemma 5 of Zhang and Wang (2016), one can see that the specific form of the asymptotic rates comes from the application of Bernstein's inequality in Lemma 5. Generalizing the proof strategy to the smoother above, the rate is given by $\left(\mathbb{E}[V_n^2]\log(n)/h^4\right)^{1/2}$, where

$$V_n = v_n \sum_{m \neq m'} |b_{nmm'}| K\left(\frac{t - t_{nm}}{h}\right) K\left(\frac{t' - t_{nm'}}{h}\right) U^+_{nmm'},$$

where $U_{nmm'} = \operatorname{sign}(b_{nmm'})[X(t_{nm}, s_{nm}) - \mu(t_{nm}, s_{nm})][X(t_{nm'}, s_{nm'}) - \mu(t_{nm'}, s_{nm'})]$ and the superscript denotes the positive part. Following similar steps to those in Zhang and Wang (2016), we can bound the second moment as

$$\mathbb{E}[V_n^2] \leq R\Big[v_n^2 h^2 \beta_{n,2} + v_n^2 h^3 \beta_{n,3} + v_n^2 h^4 \beta_{n,4}\Big],$$

where

$$\begin{aligned}
\beta_{n,2} &= \sum_{m \neq m'} |b_{nmm'}|^2, \\
\beta_{n,3} &= 2 \sum_{(m_1, m_2, m) \in \Omega_3} |b_{nm_1m}||b_{nm_2m}|, \\
\Omega_3 &= \Big\{(m_1, m_2, m_3) \in \{1, \ldots, M_n\}^3 | m_i \neq m_j \ \forall i, j = 1, 2, 3, 4, i \neq j\Big\}, \\
\beta_{n,4} &= \sum_{(m_1, m_2, m'_1, m'_2) \in \Omega_4} |b_{nm_1m'_1}||b_{nm_2m'_2}|, \\
\Omega_4 &= \Big\{(m_1, m_2, m_3, m_4) \in \{1, \ldots, M_n\}^4 | m_i \neq m_j \ \forall i, j = 1, 2, 3, 4, i \neq j\Big\},
\end{aligned} \tag{4.34}$$

and R is a universal constant.

Hence the asymptotically optimal per-surface weights are given by

$$\operatorname*{arg\,min}_{v_1, \ldots, v_N} \sum_{n=1}^N v_n^2 \left(\frac{\beta_{n,2}}{h^2} + \frac{\beta_{n,3}}{h} + \beta_{n,4}\right) \quad \text{subject to} \quad \sum_{n=1}^N v_n \beta_{n,1} = 1,$$

where $\beta_{n,1} = \sum_{m \neq m'} |b_{nmm'}|$. Using the method of Lagrange multipliers, the optimal weights are found to be

$$v_n = \frac{\beta_{n,1}\Big[h^{-2}\beta_{n,2} + h^{-1}\beta_{n,3} + \beta_{n,4}\Big]^{-1}}{\sum_{n=1}^N \beta_{n,1}^2 \Big[h^{-2}\beta_{n,2} + h^{-1}\beta_{n,3} + \beta_{n,4}\Big]^{-1}}. \tag{4.35}$$

This asymptotically optimal per-surface weighting scheme can be incorporated into every step of the estimation methodology proposed in Section 4.3.

In our experience, however, using these weights does not lead to an improvement. As comprehensive and elegant as the unified theory of Zhang and Wang (2016) is, the practical gains from adopting the asymptotically optimal weights are very small already in the one-dimensional case (Zhang and Wang, 2018). In fact, Zhang and Wang (2016) state: “A general guideline is to adopt the OBS scheme unless one believes that the data are ultra-dense or if the distribution of M_n suggests a heavy upper tail.” Zhang and Wang (2018) still recommend to use the weights, as there is no associated computational overhead.

However, in our two-dimensional case under separability, the situation is more complicated. Firstly, there is a computational overhead associated with calculating the weights, namely when calculating $\beta_{n,3}$ and $\beta_{n,4}$ in (4.34). Even though, when working on a grid, these can be calculated on marginalized data (see Section 4.5) and a trick involving a rank-one tensor can be used to calculate $\beta_{n,4}$ with the same cost as $\beta_{n,3}$, there are still non-negligible associated computational costs. Secondly, the weighting scheme above depends on an unknown parameter ($b = b(s, s')$ in the step, where $a = a(t, t')$ is estimated). In practice, one naturally uses the estimator available in the given step, but this significantly complicates fully generalizing the unified theory of Zhang and Wang (2016) to our case. And, more importantly, the finite sample performance can suffer from the additional variability introduced in the estimation scheme, unless the sample size is very large. In our experience, using the per-surface weighting scheme often leads to inferior performance, so we have to conclude in the same way as Zhang and Wang (2016), and recommend to use the OBS scheme of Yao et al. (2005a), i.e. the scheme used in Chapter 4. However, we obtained promising preliminary results with weighting schemes created as a biased version of the optimal weights, e.g. using $\sqrt{v_n}$ as the weights. Since the denominator in (4.35) is inessential due to standardization intrinsic to every kernel smoother, using $\sqrt{v_n}$ corresponds to a biased (flattened) version of the optimal weights. It would be interested in general (both for one- and multi-dimensional functional data), whether biased versions of the optimal weights lead to significant improvement over the commonly used OBS scheme.

Appendix A: Proofs of the Asymptotic Results

Separable-plus-Banded Model

Proof of Theorem 4

The proof will be done separately for the two sampling schemes. Moreover, the following auxiliary result will be needed.

Lemma 7. *Let $Z_1, \ldots, Z_K$ be i.i.d. random variables with finite second moments. Then*

$$\mathbb{E}\Big(\sum_{k=1}^{K} Z_k \Big)^2 \leq K \sum_{k=1}^{K} \mathbb{E} Z_k^2 \,.$$

Proof. The claim follows from the Cauchy-Schwarz inequality followed by the arithmetic-geometric mean inequality:

$$\begin{aligned} \mathbb{E}\Big(\sum_{k=1}^{K} Z_k \Big)^2 &= \sum_{k=1}^{K} \sum_{l=1}^{K} \mathbb{E} Z_k Z_l \leq \sum_{k=1}^{K} \sum_{l=1}^{K} \sqrt{\mathbb{E} Z_k^2} \sqrt{\mathbb{E} Z_l^2} \\ &\leq \sum_{k=1}^{K} \sum_{l=1}^{K} \frac{\mathbb{E} Z_k^2 + \mathbb{E} Z_l^2}{2} = K \sum_{k=1}^{K} \mathbb{E} Z_k^2. \end{aligned} \qquad \square$$

Proof of Theorem 4, pointwise sampling scheme S1. We begin with the bias-variance decomposition (i.e. the parallelogram law):

$$\Big|\Big|\Big| \widehat{A}_1^K \,\tilde{\otimes}\, \widehat{A}_2^K - A_1 \,\tilde{\otimes}\, A_2 \Big|\Big|\Big|_2^2 = 2 \Big|\Big|\Big| \widehat{A}_1^K \,\tilde{\otimes}\, \widehat{A}_2^K - A_1^K \,\tilde{\otimes}\, A_2^K \Big|\Big|\Big|_2^2 + 2 \Big|\Big|\Big| A_1^K \,\tilde{\otimes}\, A_2^K - A_1 \,\tilde{\otimes}\, A_2 \Big|\Big|\Big|_2^2 .$$

For the bias term, we first distribute the norm calculation over the grid:

$$\begin{aligned} &\Big|\Big|\Big| A_1^K \,\tilde{\otimes}\, A_2^K - A_1 \,\tilde{\otimes}\, A_2 \Big|\Big|\Big|_2^2 = \\ &= \sum_{i,j,k,l=1}^{K} \int_{I_{i,j}^K \times I_{k,l}^K} \Big[a_1^K(t,t') a_2^K(s,s') - a_1(t,t') a_2(s,s') \Big]^2 \mathrm{d}t \, \mathrm{d}s \, \mathrm{d}t' \, \mathrm{d}s'. \end{aligned}$$

Since $a_1^K(t,s)a_2^K(s,s') = a_1(t_i,t_k)a_2(t_j,s_l)$ on $I_{i,j}^K \times I_{k,l}^K$, it follows from Lipschitz continuity that

$$\begin{aligned} |a_1^K(t,t')a_2^K(s,s') - a_1(t,t')a_2(s,s')| &\le L \sup_{(t,s,t',s')\in I_{i,j}^K\times I_{k,l}^K} \|(t,s,t',s') - (t_i,s_j,t_k,s_l)\|_2 \\ &\le 4^{1/2}K^{-1}L\,, \end{aligned}$$

which implies the bound for the bias term. It remains to show that the variance term is $\mathcal{O}_P(N^{-1})$ uniformly in K.

Since $\mathrm{Tr}^{\delta}(A) > 0$ and $\delta_K = \lceil \delta K \rceil / K \searrow \delta$, due to continuity of kernel a there exist $K_0 \in \mathbb{N}$ such that $\mathrm{Tr}^{\delta_K}(A) > 0$ for any $K \ge K_0$. Assume from now on that $K \ge K_0$.

Using that $A_1^K \tilde{\otimes} A_2^K = \frac{\mathrm{Tr}_1^{\delta_K}(C^K) \tilde{\otimes} \mathrm{Tr}_2^{\delta_K}(C^K)}{\mathrm{Tr}^{\delta_K}(C^K)}$ in our model, it follows from the triangle inequality that

$$\begin{aligned} \left|\!\left|\!\left| \widehat{A}_1^K \tilde{\otimes} \widehat{A}_2^K - A_1^K \tilde{\otimes} A_2^K \right|\!\right|\!\right|_2 &= \left|\!\left|\!\left| \frac{\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K) \tilde{\otimes} \mathrm{Tr}_2^{\delta_K}(\widehat{C}_N^K)}{\mathrm{Tr}^{\delta_K}(\widehat{C}_N^K)} - \frac{\mathrm{Tr}_1^{\delta_K}(C^K) \tilde{\otimes} \mathrm{Tr}_2^{\delta_K}(C^K)}{\mathrm{Tr}^{\delta_K}(C^K)} \right|\!\right|\!\right|_2 \\ &\le \frac{\left|\!\left|\!\left| \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K) \right|\!\right|\!\right|_2}{\left| \mathrm{Tr}^{\delta_K}(\widehat{C}_N^K) \right|} \left|\!\left|\!\left| \mathrm{Tr}_2^{\delta_K}(\widehat{C}_N^K - C^K) \right|\!\right|\!\right|_2 \\ &\quad + \frac{\left|\!\left|\!\left| \mathrm{Tr}_2^{\delta_K}(C^K) \right|\!\right|\!\right|_2}{\left| \mathrm{Tr}^{\delta_K}(\widehat{C}_N^K) \right|} \left|\!\left|\!\left| \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K) \right|\!\right|\!\right|_2 \qquad \text{(A.36)} \\ &\quad + \frac{\left|\!\left|\!\left| A_1^K \tilde{\otimes} A_2^K \right|\!\right|\!\right|_2}{\left| \mathrm{Tr}^{\delta_K}(\widehat{C}_N^K) \right|} \left| \mathrm{Tr}^{\delta_K}(\widehat{C}_N^K - C^K) \right|. \end{aligned}$$

Now we treat different terms separately. The numerators will be shown to be $\mathcal{O}_P(1)$, as well as $1/\left|\mathrm{Tr}^{\delta_K}(\widehat{C}_N^K)\right|$, while the remaining terms will be shown to be $\mathcal{O}_P(N^{-1/2})$; all these rates being uniform in K. To simplify the notation, we denote $\widetilde{k} := k + \delta_K K$, $\widetilde{K} := (1-\delta_K)K$, and $d_K := \delta_K K$.

Firstly, we show that $\left|\!\left|\!\left| \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K) \right|\!\right|\!\right|_2 = \mathcal{O}_P(N^{-1/2})$ uniformly in K. To that end, since $\mathbf{C}^K = \mathbb{E}(\mathbf{X} \otimes \mathbf{X})$ and using Lemma 4, we have

$$\begin{aligned} &\mathbb{E}\left|\!\left|\!\left| \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K) \right|\!\right|\!\right|_2^2 = \\ &= K^{-4}\mathbb{E}\left\| \mathrm{Tr}_1^{d_K}(\widehat{\mathbf{C}}_N^K - \mathbf{C}^K) \right\|_F^2 = K^{-4}\sum_{i=1}^K\sum_{j=1}^K \mathbb{E}\left| \mathrm{Tr}_1^{d_K}(\widehat{\mathbf{C}}_N^K - \mathbf{C}^K)[i,j] \right|^2 \\ &= K^{-4}\sum_{i=1}^K\sum_{j=1}^K \mathbb{E}\left| \frac{1}{N}\sum_{n=1}^N\sum_{k=1}^{\widetilde{K}} \left(\widetilde{\mathbf{X}}_n^K[i,k]\widetilde{\mathbf{X}}_n^K[j,\widetilde{k}] - \mathbb{E}\mathbf{X}_n^K[i,k]\mathbf{X}_n^K[i,\widetilde{k}] \right) \right|^2 . \end{aligned}$$

If we denote

$$\begin{aligned}
Z_{n,i,j}^{K} &:= \sum_{k=1}^{\widetilde{K}} \left(\widetilde{\mathbf{X}}_n^K[i,k] \widetilde{\mathbf{X}}_n^K[j,\widetilde{k}] - \mathbb{E}\mathbf{X}_n^K[i,k]\mathbf{X}_n^K[i,\widetilde{k}] \right) \\
&= \sum_{k=1}^{\widetilde{K}} \Big(\mathbf{X}_n^K[i,k]\mathbf{X}_n^K[j,\widetilde{k}] - \mathbb{E}\mathbf{X}_n^K[i,k]\mathbf{X}_n^K[i,\widetilde{k}] \\
&\qquad + \mathbf{E}_n^K[i,k]\mathbf{X}_n^K[j,\widetilde{k}] + \mathbf{X}_n^K[i,k]\mathbf{E}_n^K[j,\widetilde{k}] + \mathbf{E}_n^K[i,k]\mathbf{E}_n^K[j,\widetilde{k}] \Big) ,
\end{aligned}$$

we see that, for any $i, j = 1, \ldots, K$, $\left\{ Z_{n,i,j}^{K} \right\}_{n=1}^{N}$ is a set of mean zero and i.i.d. random variables and thus

$$\mathbb{E} \left|\!\left|\!\left| \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K) \right|\!\right|\!\right|_2^2 \leq \frac{1}{N} K^{-4} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{E} \left| Z_{\cdot,i,j}^{K} \right|^2$$

which can be bounded, using the parallelogram law, by

$$\begin{aligned}
\frac{4}{N} K^{-4} \sum_{i=1}^{K} \sum_{j=1}^{K} \Bigg\{ & \mathbb{E} \left| \sum_{k=1}^{\widetilde{K}} \mathbf{X}^K[i,k]\mathbf{X}^K[j,\widetilde{k}] - \mathbb{E}\mathbf{X}^K[i,k]\mathbf{X}^K[i,\widetilde{k}] \right|^2 \\
& + \mathbb{E} \left| \sum_{k=1}^{\widetilde{K}} \mathbf{E}^K[i,k]\mathbf{X}^K[j,\widetilde{k}] \right|^2 + \mathbb{E} \left| \sum_{k=1}^{\widetilde{K}} \mathbf{X}^K[i,k]\mathbf{E}^K[j,\widetilde{k}] \right|^2 \\
& + \mathbb{E} \left| \sum_{k=1}^{\widetilde{K}} \mathbf{E}^K[i,k]\mathbf{E}^K[j,\widetilde{k}] \right|^2 \Bigg\} .
\end{aligned} \tag{A.37}$$

The four terms in the parentheses will be treated separately.

For the first term, it follows from Lemma 7 that

$$\mathbb{E} \left| \sum_{k=1}^{\widetilde{K}} \mathbf{X}^K[i,k]\mathbf{X}^K[j,\widetilde{k}] - \mathbb{E}\mathbf{X}^K[i,k]\mathbf{X}^K[i,\widetilde{k}] \right|^2 \leq \widetilde{K} \sum_{k=1}^{\widetilde{K}} \mathrm{Var}\left(\mathbf{X}^K[i,k]\mathbf{X}^K[j,\widetilde{k}] \right) \leq S_1 \widetilde{K}^2 ,$$

where

$$\begin{aligned}
\mathrm{Var}\left(\mathbf{X}^K[i,k]\mathbf{X}^K[j,\widetilde{k}] \right) &= \mathrm{Var}\left(X(t_i^K, s_k^K) X(t_j, s_{\widetilde{k}} \right) \\
&\leq \sup_{t,s,t',s' \in [0,1]} \mathrm{Var}\left(X(t,s) X(t',s') \right) =: S_1 < \infty.
\end{aligned}$$

Note that S_1 is finite, since X has finite fourth moment and continuous sample paths. Also, S_1 is uniform in K.

For the second term, we have (denoting $\widetilde{l} = l + \delta_k K$)

$$\begin{aligned}\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}} \mathbf{E}^K[i,k]\mathbf{X}^K[j,\widetilde{k}]\right|^2 &= \sum_{k=1}^{\widetilde{K}}\sum_{l=1}^{\widetilde{K}} \mathbb{E}\Big(\mathbf{E}^K[i,k]\mathbf{X}^K[j,\widetilde{k}]\mathbf{E}^K[i,l]\mathbf{X}^K[j,\widetilde{l}]\Big) \\ &= \sum_{k=1}^{\widetilde{K}}\sum_{l=1}^{\widetilde{K}} \mathbb{E}\Big(\mathbf{E}^K[i,k]\mathbf{E}^K[i,l]\Big)\mathbb{E}\Big(\mathbf{X}^K[j,\widetilde{k}]\mathbf{X}^K[j,\widetilde{l}]\Big)\,.\end{aligned}$$

Since $\mathbb{E}\Big(\mathbf{E}^K[i,k]\mathbf{E}^K[i,l]\Big) = \sigma^2\mathbb{1}_{[k=l]}$, one of the sums vanishes, while $\mathbb{E}\Big|\mathbf{X}^K[j,\widetilde{k}]\Big|^2$ is bounded uniformly in K by $S_2 := \sup_{t,s\in[0,1]}\mathbb{E}\Big|X(t,s)\Big|^2 \leq \infty$. Hence the second term is bounded by $\widetilde{K}S_2\sigma^2$. The third term is dealt with similarly.

For the fourth and final term, we have

$$\begin{aligned}\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}} \mathbf{E}^K[i,k]\mathbf{E}^K[j,\widetilde{k}]\right|^2 &= \sum_{k=1}^{\widetilde{K}}\sum_{l=1}^{\widetilde{K}} \mathbb{E}\Big(\mathbf{E}^K[i,k]\mathbf{E}^K[j,\widetilde{k}]\mathbf{E}^K[i,l]\mathbf{E}^K[j,\widetilde{l}]\Big) \\ &= \sum_{k=1}^{\widetilde{K}}\sum_{l=1}^{\widetilde{K}} \sigma^4\mathbb{1}_{[k=l]} = \widetilde{K}\sigma^4\,.\end{aligned}$$

Upon collecting the bounds for the four terms and importing them back to bound (A.37), we obtain

$$\mathbb{E}\Big|\Big|\Big|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\Big|\Big|\Big|_2^2 \leq \frac{4}{N}\Big[S_1 + S_2K^{-1}\sigma^2 + K^{-1}\sigma^4\Big]\,. \tag{A.38}$$

This shows that if $\sigma^2 = \mathcal{O}(\sqrt{K})$, $\Big|\Big|\Big|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\Big|\Big|\Big|_2 = \mathcal{O}_P(N^{-1/2})$ uniformly in K.

The term $\Big|\Big|\Big|\mathrm{Tr}_2^{\delta_K}(\widehat{C}_N^K - C^K)\Big|\Big|\Big|_2$ from bound (A.36) can be treated similarly. Now we focus on the final stand-alone term $\Big|\mathrm{Tr}^{\delta_K}(\widehat{C}_N^K - C^K)\Big|$:

$$\begin{aligned}\mathbb{E}\left|\mathrm{Tr}^{\delta}(\widehat{C}_N^K - C^K)\right|^2 &= K^{-4}\mathbb{E}\left|\mathrm{Tr}^{\delta}(\widehat{\mathbf{C}}_N^K - \mathbf{C}^K)\right|^2 \\ &= K^{-4}\mathbb{E}\left|\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\left(\widetilde{\mathbf{X}}_n^K[i,j]\widetilde{\mathbf{X}}_n^K[\widetilde{i},\widetilde{j}] - \mathbb{E}\mathbf{X}_n^K[i,j]\mathbf{X}_n^K[\widetilde{i},\widetilde{j}]\right)\right|^2 \\ &= \frac{1}{N}K^{-4}\mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\left(\widetilde{\mathbf{X}}^K[i,j]\widetilde{\mathbf{X}}^K[\widetilde{i},\widetilde{j}] - \mathbb{E}\mathbf{X}^K[i,j]\mathbf{X}^K[\widetilde{i},\widetilde{j}]\right)\right|^2\end{aligned}$$

From the parallelogram law we have

$$
\begin{aligned}
\mathbb{E}\left|\operatorname{Tr}^{\delta}(\widehat{C}_N^K - C^K)\right|^2 \leq \frac{4}{N}K^{-4}\Bigg\{ & \mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\Big(\mathbf{X}^K[i,j]\mathbf{X}^K[\widetilde{i},\widetilde{j}] - \mathbb{E}\mathbf{X}^K[i,j]\mathbf{X}^K[\widetilde{i},\widetilde{j}]\Big)\right|^2 \\
& + \mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbf{E}^K[i,j]\mathbf{X}^K[\widetilde{i},\widetilde{j}]\right|^2 \\
& + \mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbf{X}^K[i,j]\mathbf{E}^K[\widetilde{i},\widetilde{j}]\right|^2 \\
& + \mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbf{E}^K[i,j]\mathbf{E}^K[\widetilde{i},\widetilde{j}]\right|^2 \Bigg\}.
\end{aligned}
$$

Using Lemma 7 to take the sums out of the expectation, the first term in the parentheses is again bounded by $K^4 S_1$. For the second term,

$$
\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbf{E}^K[i,j]\mathbf{X}^K[\widetilde{i},\widetilde{j}]\right|^2 &= \sum_{i,j,k,l=1}^{\widetilde{K}} \mathbb{E}\Big(\mathbf{E}^K[i,j]\mathbf{X}^K[\widetilde{i},\widetilde{j}]\mathbf{E}^K[k,l]\mathbf{X}^K[\widetilde{k},\widetilde{l}]\Big) \\
&= \sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbb{E}\left|X^K[\widetilde{i},\widetilde{j}]\right|^2 \leq K^2\sigma^2 S_2\,.
\end{aligned}
$$

The third term can be treated similarly, while for the fourth and final term we have

$$
\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbf{E}^K[i,j]\mathbf{E}^K[\widetilde{i},\widetilde{j}]\right|^2 &= \sum_{i,j,k,l=1}^{\widetilde{K}} \mathbb{E}\Big(\mathbf{E}^K[i,j]\mathbf{E}^K[\widetilde{i},\widetilde{j}]\mathbf{E}^K[k,l]\mathbf{E}^K[\widetilde{k},\widetilde{l}]\Big) \\
&= \sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbb{E}\left|\mathbf{E}^K[i,j]\right|^2\mathbb{E}\left|\mathbf{E}^K[\widetilde{i},\widetilde{j}]\right|^2 \leq K^2\sigma^4\,.
\end{aligned}
$$

Hence we obtain

$$
\mathbb{E}\left|\operatorname{Tr}^{\delta}(\widehat{C}_N^K - C^K)\right|^2 \leq \frac{4}{N}\Big[S_1 + S_2 K^{-2}\sigma^2 + K^{-2}\sigma^4\Big]\,. \tag{A.39}
$$

Note the different powers of K in (A.38) and (A.39). This reflects that the concentration of measurement error is weaker when averaging is performed over both time and space (when shifted tracing is used) in comparison to averaging only over either time or space (when shifted partial tracing is used).

Now let us focus on the numerators in (A.36), for example:

$$\left|\!\left|\!\left|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K)\right|\!\right|\!\right|_2 \leq \left|\!\left|\!\left|\mathrm{Tr}_1^{\delta_K}(C^K)\right|\!\right|\!\right|_2 + \left|\!\left|\!\left|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\right|\!\right|\!\right|_2,$$

where the second term is $\mathcal{O}_P(N^{-1/2})$ uniformly in K, while the first term is clearly bounded by $\sup_{t,s,t',s' \in [0,1]} c(t,s,t',s') < \infty$, hence $\left|\!\left|\!\left|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K)\right|\!\right|\!\right|_2$ is $\mathcal{O}_P(1)$ uniformly in K. Similarly for the other two numerator terms.

Finally, we consider the denominators in (A.36). The reverse triangle inequality implies

$$\left|\mathrm{Tr}^{\delta_K}(\widehat{C}_N^K)\right| \geq \left|\mathrm{Tr}^{\delta_K}(C^K)\right| - \left|\mathrm{Tr}^{\delta_K}(\widehat{C}_N^K - C^K)\right|,$$

where the second term is again $\mathcal{O}_P(N^{-1/2})$ uniformly in K as shown above, and the first term is bounded away from 0 uniformly in K (for large enough K) due to continuity of the kernel a of the separable part A and the assumption $\mathrm{Tr}^{\delta_K}(A) > 0$, because $\mathrm{Tr}^{\delta_K}(A) = \mathrm{Tr}^{\delta_K}(C)$. Hence $1/\left|\mathrm{Tr}^{\delta_K}(\widehat{C}_N^K)\right|$ is $\mathcal{O}_P(1)$ uniformly in K.

The proof of the rates for the separable estimator is complete upon collecting the rates for the different terms in (A.36).

The rate for the eigenvalues follows from the perturbation bounds (Bosq, 2012, Lemma 4.2):

$$|\widehat{\lambda}_i^K \widehat{\rho}_j^K - \lambda_i \rho_j|^2 \leq \left|\!\left|\!\left|\widehat{A}_1^K \,\tilde{\otimes}\, \widehat{A}_2^K - A_1 \,\tilde{\otimes}\, A_2\right|\!\right|\!\right|_2^2.$$

To show the rates for the eigenvectors, we will use again the perturbation bounds (Bosq, 2012, Lemma 4.3):

$$\|\widehat{e}_j^K - \mathrm{sign}(\langle \widehat{e}_j^K, e_j \rangle e_j)\|_2 \leq \alpha \|\widehat{A}_1^K - A_1\|_2,$$

where α is a constant depending on spacing between the eigenvalues. We cannot use this result directly, since we do not have consistency of $\widehat{A}_1^K$ (this is because of the scaling issues: $A_1 \,\tilde{\otimes}\, A_2 = (\alpha A_1) \,\tilde{\otimes}\, (A_2)/\alpha$ for any α). Hence similar bounds always have to be used in the product space. However, this poses no issues due to Lemma 1. We have

$$\begin{aligned}
\|\widehat{e}_j^K - \mathrm{sign}(\langle \widehat{e}_j^K, e_j \rangle e_j)\|_2 &= \|f_j\|_2 \|\widehat{e}_j^K - \mathrm{sign}(\langle \widehat{e}_j^K, e_j \rangle e_j)\|_2 \\
&= \|\widehat{e}_j^K \otimes f_j - \mathrm{sign}(\langle \widehat{e}_j^K, e_j \rangle) e_j \otimes f_j\|_2 \\
&\leq \|\widehat{e}_j^K \otimes \widehat{f}_j^K - \mathrm{sign}(\langle \widehat{e}_j^K, e_j \rangle)\, \mathrm{sign}(\langle \widehat{f}_j^K, f_j \rangle) e_j \otimes f_j\|_2.
\end{aligned}$$

The previous inequality follows from the Cauchy-Schwarz inequality and the fact that the left-hand side of the inequality equal to $2 - 2\,\mathrm{sign}(\langle \widehat{e}_j^K, e_j \rangle) \langle \widehat{e}_j^K, e_j \rangle e_j \rangle \langle f_j, f_j \rangle$ while the right-hand side is equal to $2 - 2\,\mathrm{sign}(\langle \widehat{e}_j^K, e_j \rangle)\, \mathrm{sign}(\langle \widehat{f}_j^K, f_j \rangle) \langle \widehat{e}_j^K, e_j \rangle e_j \rangle \langle \widehat{f}_j^K, f_j \rangle$. Altogether, the rate for $\widehat{A}_1^K \,\tilde{\otimes}\, \widehat{A}_2^K$ translates to the eigenvectors of $\widehat{A}_1^K$, and similarly for the eigenvectors of $\widehat{A}_2^K$. $\square$

The proof of the theorem in the case of pixel-wise sampling scheme (S2) is in many regards similar, but some arguments are slightly more subtle.

Proof of Theorem 4, pixel-wise sampling scheme S2. We begin again the by the bias-variance decomposition and bound the bias term in the same manner. For the variance term, we use the triangle inequality treat all the terms in (A.36) separately. The fractions are also treated the same way as before and the conclusion of the proof will follow similarly, once it is established that $\left\vert\!\left\vert\!\left\vert \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\right\vert\!\right\vert\!\right\vert_2$, $\left\vert\!\left\vert\!\left\vert \mathrm{Tr}_2^{\delta_K}(\widehat{C}_N^K - C^K)\right\vert\!\right\vert\!\right\vert_2$ and $\left\vert \mathrm{Tr}^{\delta_K}(\widehat{C}_N^K - C^K)\right\vert$ are all $\mathcal{O}_P(N^{-1/2})$ uniformly in K. Establishing these rates for the pointwise sampling scheme (S1) was the bulk of the previous proof, and now we will establish the same for the pixel-wise sampling scheme (S2).

We begin with $\left\vert\!\left\vert\!\left\vert \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\right\vert\!\right\vert\!\right\vert_2$. Exactly as in the previous proof, we obtain the bound (A.37) here as well:

$$
\begin{aligned}
\mathbb{E}\left\vert\!\left\vert\!\left\vert \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\right\vert\!\right\vert\!\right\vert_2^2 \leq \frac{4}{N}\Bigg\{ & K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{X}^K[i,k]\mathbf{X}^K[j,\widetilde{k}]\right|^2 \\
& + K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{E}^K[i,k]\mathbf{X}^K[j,\widetilde{k}]\right|^2 \\
& + K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{X}^K[i,k]\mathbf{E}^K[j,\widetilde{k}]\right|^2 \\
& + K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{E}^K[i,k]\mathbf{E}^K[j,\widetilde{k}]\right|^2 \Bigg\} \\
=: \frac{4}{N}\Bigg\{ & (I) + (II) + (III) + (IV) \Bigg\},
\end{aligned}
$$

and again we treat the four terms in the parentheses (labeled by Roman numbers) separately.

For the first term, we drop the inner expectation only increasing the term and obtaining

$$
\begin{aligned}
(I) &= K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{X}^K[i,k]\mathbf{X}^K[j,\widetilde{k}]\right|^2 \\
&= K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{\widetilde{K}}\sum_{l=1}^{\widetilde{K}}\mathbb{E}\Big(\mathbf{X}^K[i,k]\mathbf{X}^K[j,\widetilde{k}]\mathbf{X}^K[i,l]\mathbf{X}^K[j,\widetilde{l}]\Big) \\
&= \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{\widetilde{K}}\sum_{l=1}^{\widetilde{K}}\mathbb{E}\langle X, g_{i,k}^K\rangle\langle X, g_{j,\widetilde{k}}^K\rangle\langle X, g_{i,l}^K\rangle\langle X, g_{j,\widetilde{l}}^K\rangle\,,
\end{aligned}
$$

where we used that $\mathbf{X}^K[i,j] = K\langle X, g^K_{i,j}\rangle$ for the function $g_{i,j}$ defined in (2.40). If we now denote $\Gamma = \mathbb{E} X \otimes X \otimes X \otimes X$, it follows from the outer product algebra (or can be verified explicitly using integral representations) that (recall that we denote $\widetilde{k} = k + \delta_K K$ and $\widetilde{l} = l + \delta_K K$)

$$\begin{aligned}\mathbb{E}\langle X, g^K_{i,k}\rangle\langle X, g^K_{j,\widetilde{k}}\rangle\langle X, g^K_{i,l}\rangle\langle X, g^K_{j,\widetilde{l}}\rangle &= \mathbb{E}\langle X \otimes X \otimes X \otimes X, g^K_{i,k} \otimes g^K_{j,\widetilde{k}} \otimes g^K_{i,l} \otimes g^K_{j,\widetilde{l}}\rangle \\ &= \langle \Gamma, g^K_{i,k} \otimes g^K_{j,\widetilde{k}} \otimes g^K_{i,l} \otimes g^K_{j,\widetilde{l}}\rangle \\ &= \langle \Gamma(g^K_{i,k} \otimes g^K_{j,\widetilde{l}}), g^K_{j,\widetilde{k}} \otimes g^K_{i,l}\rangle\,.\end{aligned}$$

Due to positive semi-definiteness of Γ, the last expression is bounded by

$$\frac{1}{2}\Big[\langle \Gamma(g^K_{i,k} \otimes g^K_{j,\widetilde{l}}), g^K_{i,k} \otimes g^K_{j,\widetilde{l}}\rangle + \langle \Gamma(g^K_{j,\widetilde{k}} \otimes g^K_{i,l}), g^K_{j,\widetilde{k}} \otimes g^K_{i,l}\rangle\Big]$$

which gives us the bound

$$(I) \leq \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{(1-\delta_K)K}\sum_{l=1}^{(1-\delta_K)K}\Big[\langle \Gamma(g^K_{i,k} \otimes g^K_{j,\widetilde{l}}), g^K_{i,k} \otimes g^K_{j,\widetilde{l}}\rangle + \langle \Gamma(g^K_{j,\widetilde{k}} \otimes g^K_{i,l}), g^K_{j,\widetilde{k}} \otimes g^K_{i,l}\rangle\Big]\,.$$

Since Γ is positive semi-definite, we can add terms into the bound to symmetrize it:

$$(I) \leq \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{l=1}^{K}\langle \Gamma(g^K_{i,k} \otimes g^K_{j,l}), g^K_{i,k} \otimes g^K_{j,l}\rangle\,.$$

Finally, note that $\langle g^k_{i,j}, g_{k,l}\rangle = \mathbb{1}_{[i=k,j=l]}$ for $i,j,k,l = 1,\dots,K$, hence $\{g^K_{i,j}\}_{i,j=1}^{K}$ can be completed to an orthonormal basis of $\mathcal{L}^2([0,1]^2)$ denoted as $\{g^K_{i,j}\}_{i,j=1}^{\infty}$. We can add some more extra terms due to positive semi-definiteness of Γ to obtain

$$(I) \leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\langle \Gamma(g^K_{i,k} \otimes g^K_{j,l}), g^K_{i,k} \otimes g^K_{j,l}\rangle = |||\Gamma|||_1\,.$$

Note that even though the orthonormal basis used changes with every K, the final equality holds for any orthonormal basis (Hsing and Eubank, 2015, p. 114), and hence we obtain uniformity in K.

The strategy is similar for the remaining terms (II), (III) and (IV). For the second one:

$$\begin{aligned}(II) &= K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{E}^K[i,k]\mathbf{X}^K[j,\widetilde{k}]\right|^2 = K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{\widetilde{K}}\mathbb{E}\left|\mathbf{E}^K[i,k]\right|^2\mathbb{E}\left|\mathbf{X}^K[j,\widetilde{k}]\right|^2 \\ &= K^{-3}\sigma^2\sum_{j=1}^{K}\sum_{k=1}^{\widetilde{K}}\mathbb{E}|\mathbf{X}^K[j,\widetilde{k}]|^2 = K^{-1}\sigma^2\sum_{j=1}^{K}\sum_{k=1}^{\widetilde{K}}\mathbb{E}\langle X, g^K_{j,\widetilde{k}}\rangle^2 \\ &= K^{-1}\sigma^2\sum_{j=1}^{K}\sum_{k=1}^{\widetilde{K}}\langle C(g^K_{j,\widetilde{k}}), g^K_{j,\widetilde{k}}\rangle^2 \leq K^{-1}\sigma^2|||C|||_1\,.\end{aligned}$$

The third term can be treated exactly like the second one, and for the final term we have

$$\begin{aligned}
(IV) &= K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{E}^K[i,k]\mathbf{E}^K[j,\widetilde{k}]\right|^2 \\
&= K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{\widetilde{K}}\sum_{l=1}^{\widetilde{K}}\mathbb{E}\Big(\mathbf{E}^K[i,k]\mathbf{E}^K[j,\widetilde{k}]\mathbf{E}^K[i,l]\mathbf{E}^K[j,\widetilde{l}]\Big) \\
&= K^{-4}\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{\widetilde{K}}\mathbb{E}\Big|\mathbf{E}^K[i,k]\Big|^2\mathbb{E}\Big|\mathbf{E}^K[j,\widetilde{k}]\Big|^2 \leq K^{-1}\sigma^4\,,
\end{aligned}$$

Piecing things together, we have

$$\mathbb{E}\Big|\!\Big|\!\Big|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\Big|\!\Big|\!\Big|_2^2 \leq \frac{4}{N}\Big[|\!|\!|\Gamma|\!|\!|_1 + 2K^{-1}\sigma^2|\!|\!|C|\!|\!|_1 + K^{-1}\sigma^4\Big]\,.$$

Thus we have shown that $\Big|\!\Big|\!\Big|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\Big|\!\Big|\!\Big|_2 = \mathcal{O}_P(N^{-1/2})$ uniformly in K, since $\sigma^2 = \mathcal{O}(\sqrt{K})$. It can be shown in an analogous way that $\Big|\!\Big|\!\Big|\mathrm{Tr}_2^{\delta_K}(\widehat{C}_N^K - C^K)\Big|\!\Big|\!\Big|_2 = \mathcal{O}_P(N^{-1/2})$ uniformly in K, and it remains to show the same for $\Big|\mathrm{Tr}^{\delta_K}(\widehat{C}_N^K - C^K)\Big|$.

Similarly to before we obtain the following bound:

$$\begin{aligned}
\mathbb{E}\left|\mathrm{Tr}^{\delta_K}(\widehat{C}_N^K - C^K)\right|^2 \leq \frac{4}{N}\Bigg\{ & K^{-4}\mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbf{X}^K[i,j]\mathbf{X}^K[\widetilde{i},\widetilde{j}]\right|^2 \\
& + K^{-4}\mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbf{E}^K[i,j]\mathbf{X}^K[\widetilde{i},\widetilde{j}]\right|^2 \\
& + K^{-4}\mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbf{X}^K[i,j]\mathbf{E}^K[\widetilde{i},\widetilde{j}]\right|^2 \\
& + K^{-4}\mathbb{E}\left|\sum_{i=1}^{\widetilde{K}}\sum_{j=1}^{\widetilde{K}}\mathbf{E}^K[i,j]\mathbf{E}^K[\widetilde{i},\widetilde{j}]\right|^2 \Bigg\} \\
=: \frac{4}{N}\Bigg\{ & (I) + (II) + (III) + (IV) \Bigg\}\,,
\end{aligned}$$

in which we will treat again the four terms separately.

For the first term:

$$\begin{aligned}
(I) &= K^{-4} \sum_{i,j,k,l=1}^{\widetilde{K}} \mathbb{E}\Big(\mathbf{X}^K[i,j]\mathbf{X}^K[\widetilde{i},\widetilde{j}]\mathbf{X}^K[k,l]\mathbf{X}^K[\widetilde{k},\widetilde{l}] \Big) \\
&= \sum_{i,j,k,l=1}^{\widetilde{K}} \mathbb{E}\Big(\langle X, g^K_{i,j}\rangle \langle X, g^K_{\widetilde{i},\widetilde{j}}\rangle \langle X, g^K_{k,l}\rangle \langle X, g^K_{\widetilde{k},\widetilde{l}}\rangle \Big) \\
&= \sum_{i,j,k,l=1}^{\widetilde{K}} \mathbb{E}\langle X \otimes X \otimes X \otimes X, g^K_{i,j} \otimes g^K_{\widetilde{i},\widetilde{j}} \otimes g^K_{k,l} \otimes g^K_{\widetilde{k},\widetilde{l}} \rangle \\
&= \sum_{i,j,k,l=1}^{\widetilde{K}} \langle \Gamma(g^K_{i,j} \otimes g^K_{k,l}), g^K_{\widetilde{i},\widetilde{j}} \otimes g^K_{\widetilde{k},\widetilde{l}} \rangle \leq |||\Gamma|||_1 \,.
\end{aligned}$$

For the second term,

$$\begin{aligned}
(II) &= K^{-4} \sum_{i,j,k,l=1}^{\widetilde{K}} \mathbb{E}\Big(\mathbf{X}^K[i,j]\mathbf{E}^K[\widetilde{i},\widetilde{j}]\mathbf{X}^K[k,l]\mathbf{E}^K[\widetilde{k},\widetilde{l}] \Big) \\
&= K^{-4} \sum_{i,j,k,l=1}^{\widetilde{K}} \mathbb{E}\Big(\mathbf{X}^K[i,j]\mathbf{X}^K[k,l] \Big) \mathbb{E}\Big(\mathbf{E}^K[\widetilde{i},\widetilde{j}]\mathbf{E}^K[\widetilde{k},\widetilde{l}] \Big)
\end{aligned}$$

and since $\mathbb{E}\Big(\mathbf{E}^K[\widetilde{i},\widetilde{j}]\mathbf{E}^K[\widetilde{k},\widetilde{l}] \Big) = \sigma^2 \mathbb{1}_{[i=k, j=l]}$, we have

$$(II) = \sigma^2 K^{-4} \sum_{i,j=1}^{\widetilde{K}} \mathbb{E}\Big| \mathbf{X}^K[i,j] \Big|^2 = \sigma^2 K^{-2} \sum_{i,j=1}^{\widetilde{K}} \mathbb{E}\langle X, g^K_{i,j}\rangle^2 \leq \sigma^2 K^{-2} |||C|||_1 \,.$$

The third term is bounded similarly, and for the final term:

$$\begin{aligned}
(IV) &= K^{-4} \sum_{i,j,k,l=1}^{\widetilde{K}} \mathbb{E}\Big(\mathbf{E}^K[i,j]\mathbf{E}^K[\widetilde{i},\widetilde{j}]\mathbf{E}^K[k,l]\mathbf{E}^K[\widetilde{k},\widetilde{l}] \Big) \\
&= K^{-4} \sum_{i=1}^{\widetilde{K}} \sum_{j=1}^{\widetilde{K}} \mathbb{E}\Big| \mathbf{E}^K[i,j] \Big|^2 \mathbb{E}\Big| \mathbf{E}^K[\widetilde{i},\widetilde{j}] \Big|^2 \leq K^{-2}\sigma^4
\end{aligned}$$

In summary, we have obtained the following bound:

$$\mathbb{E}\left| \mathrm{Tr}^{\delta_K}(\widehat{C}^K_N - C^K) \right|^2 = \frac{4}{N}\Big[|||\Gamma|||_1 + 2K^{-2}\sigma^2 |||C|||_1 + K^{-2}\sigma^4 \Big]. \qquad \square$$

The proof for the eigenvalues and eigenvectors remains the same as with sampling scheme (S1).

Proof of Proposition 8

The proof is similar to the one of Theorem 4. To save space, we will use the notation $\| \cdot \|_\infty$ for the uniform norm, i.e. $\|C\|_\infty := \sup_{t,s,t',s'} |c(t,s,t',s')|$. This is not to be confused with the operator norm of C denoted as $|||C|||_\infty$.

We begin with the triangle inequality separating the bias and the variance:

$$\|\widehat{A}^K - A\|_\infty \leq \|\widehat{A}^K - A^K\|_\infty + \|A^K - A\|_\infty,$$

and we bound the bias first.

Under (S1), we have

$$\begin{aligned}\|A^K - A\|_\infty &= \sup_{i,j,k,l=1}^{K} \sup_{(t,s,t',s') \in I^K_{i,j} \times I^K_{k,l}} \left| a_1^K(t_i, t_k) a_2^K(s_j, s_l) - a_1(t,t') a_2(s,s') \right| \\ &\leq \sup_{i,j,k,l=1}^{K} 2LK^{-1} = 2LK^{-1},\end{aligned}$$

where we used the Lipschitz property of A. On the other hand, under (S2), we have

$$\begin{aligned}&\|A^K - A\|_\infty = \\ &= \sup_{\substack{i,j,k,l=1,\dots,K \\ (t,s,t',s') \in I^K_{i,j} \times I^K_{k,l}}}{}^{K} \left| \frac{1}{|I^K_{i,j}|} \frac{1}{|I^K_{k,l}|} \int_{I^K_{i,j} \times I^K_{k,l}} \Big[a_1(u,v) a_2(x,y) - a_1(t,s) a_2(t',s') \Big] \mathrm{d}u \, \mathrm{d}v \, \mathrm{d}x \, \mathrm{d}y \right| \\ &\leq \sup_{i,j,k,l=1}^{K} \sup_{(t,s,t',s') \in I^K_{i,j} \times I^K_{k,l}} K^4 \int_{I^K_{i,j} \times I^K_{k,l}} \left| a_1(u,v) a_2(x,y) - a_1(t,t') a_2(s,s') \right| \mathrm{d}u \, \mathrm{d}v \, \mathrm{d}x \, \mathrm{d}y \\ &\leq \sup_{i,j,k,l=1}^{K} \sup_{(t,s,t',s') \in I^K_{i,j} \times I^K_{k,l}} K^4 \int_{I^K_{i,j} \times I^K_{k,l}} 2LK^{-1} \leq 2LK^{-1}.\end{aligned}$$

Similarly to (A.36) in the proof of Theorem 4, we obtain

$$\begin{aligned}\|\widehat{A}^K - A^K\|_\infty \leq\; & \frac{\left\| \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K) \right\|_\infty}{\left| \mathrm{Tr}^{\delta_K}(\widehat{C}_N^K) \right|} \left\| \mathrm{Tr}_2^{\delta_K}(\widehat{C}_N^K - C^K) \right\|_\infty \\ &+ \frac{\left\| \mathrm{Tr}_2^{\delta_K}(C^K) \right\|_\infty}{\left| \mathrm{Tr}^{\delta_K}(\widehat{C}_N^K) \right|} \left\| \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K) \right\|_\infty \\ &+ \frac{\left\| A_1^K \tilde{\otimes} A_2^K \right\|_\infty}{\left| \mathrm{Tr}^{\delta_K}(\widehat{C}_N^K) \right|} \left| \mathrm{Tr}^{\delta_K}(\widehat{C}_N^K - C^K) \right|,\end{aligned}$$

and we will again show that the numerators and denominators are $\mathcal{O}_P(1)$, while the remaining terms are $\mathcal{O}_P(N^{-1/2})$. In fact, the term that has to be treated is $\left\| \mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K) \right\|_\infty$.

Once we show that this term is $\mathcal{O}_P(N^{-1/2})$, exactly the same arguments like in the proof of Theorem 4 can be used to conclude.

We calculate

$$
\begin{aligned}
&\mathbb{E}\left\|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\right\|_\infty^2 = \\
&\qquad = K^{-2}\mathbb{E}\left\|\mathrm{Tr}_1^{d_K}(\widehat{\mathbf{C}}_N^K - \mathbf{C}^K)\right\|_\infty^2 = K^{-2}\sup_{i,j=1}^{K}\mathbb{E}\left|\mathrm{Tr}_1^{d_K}(\widehat{\mathbf{C}}_N^K - \mathbf{C}^K)[i,j]\right|^2 \\
&\qquad = K^{-2}\sup_{i,j=1}^{K}\mathbb{E}\left|\frac{1}{N}\sum_{n=1}^{N}\underbrace{\sum_{k=1}^{\widetilde{K}}\left(\widetilde{\mathbf{X}}_n^K[i,k]\widetilde{\mathbf{X}}_n^K[j,\widetilde{k}] - \mathbb{E}\mathbf{X}_n^K[i,k]\mathbf{X}_n^K[i,\widetilde{k}]\right)}_{=:Z_{n,ij}}\right|^2.
\end{aligned}
$$

Again, for a fixed i, j, $Z_{n,ij}$ is a set of mean-zero i.i.d. random variables and hence

$$
\mathbb{E}\left|\!\left|\!\left|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\right|\!\right|\!\right|_\infty^2 = N^{-1}K^{-2}\sup_{i,j=1}^{K}\mathbb{E}\left|Z_{n,ij}\right|^2,
$$

so it suffices to show that $K^{-2}\mathbb{E}\left|Z_{n,ij}\right|^2$ is uniformly bounded.

Under (S1), we proceed similarly as in (A.37):

$$
\begin{aligned}
K^{-2}\mathbb{E}\left|Z_{n,ij}\right|^2 \leq K^{-2}\Bigg\{&\mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{X}^K[i,k]\mathbf{X}^K[j,\widetilde{k}] - \mathbb{E}\mathbf{X}^K[i,k]\mathbf{X}^K[i,\widetilde{k}]\right|^2 \\
&+ \mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{E}^K[i,k]\mathbf{X}^K[j,\widetilde{k}]\right|^2 + \mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{X}^K[i,k]\mathbf{E}^K[j,\widetilde{k}]\right|^2 \\
&+ \mathbb{E}\left|\sum_{k=1}^{\widetilde{K}}\mathbf{E}^K[i,k]\mathbf{E}^K[j,\widetilde{k}]\right|^2\Bigg\}.
\end{aligned}
$$

The first term in the parentheses is bounded again by S_1K^2, the second term is bounded by $KS_2\sigma^2$, and the third term by $K\sigma^4$. Collecting the bounds together, we obtain under (S1) that

$$
\mathbb{E}\left|\!\left|\!\left|\mathrm{Tr}_1^{\delta_K}(\widehat{C}_N^K - C^K)\right|\!\right|\!\right|_\infty^2 \leq \frac{4}{N}\Big[S_1 + S_2K^{-1}\sigma^2 + K^{-1}\sigma^4\Big],
$$

from which the claim of follows.

Under (S2), the proof is an equivalent modification to the proof of Theorem 4.

Proof of Theorem 5

Here, we provide the rates of convergence for the adaptive bandwidth choice of Section 2.2.2. For that, we first need to study the behavior of the empirical objective function in (2.22).

Let us denote the empirical objective as

$$\widehat{\Xi}(\delta) = \left|\!\left|\!\left| \widehat{C}(\delta) \right|\!\right|\!\right|_2^2 - \frac{2}{N} \sum_{n=1}^{N} \langle X_n, \widehat{C}_{-n}(\delta) X_n \rangle$$

and the theoretical objective as

$$\Xi(\delta) = |||C(\delta) - C|||_2^2.$$

Recall that $\widehat{C}(\delta) = \widehat{A}(\delta) + \widehat{B}(\delta)$ is our separable-plus-banded estimator, while $C(\delta)$ is its limit version with infinite number of samples, i.e. $C(\delta) = A(\delta) + B(\delta)$ with

$$A(\delta) = \frac{\mathrm{Tr}_1^{\delta}(C) \,\tilde{\otimes}\, \mathrm{Tr}_2^{\delta}(C)}{\mathrm{Tr}^{\delta}(C)}, \quad B(\delta) = \mathrm{Ta}\Big(C - A(\delta)\Big).$$

The following linearization of our estimators will often allow us to develop suitable bounds:

$$\begin{aligned}
\widehat{A}(\delta_1) - A(\delta_2) = &\frac{\mathrm{Tr}_1^{\delta_1}(\widehat{C}_N)}{\mathrm{Tr}^{\delta_1}(\widehat{C}_N)} \,\tilde{\otimes}\, \Big[\mathrm{Tr}_2^{\delta_1}(\widehat{C}_N) - \mathrm{Tr}_2^{\delta_2}(C)\Big] \\
&+ \Big[\mathrm{Tr}_1^{\delta_1}(\widehat{C}_N) - \mathrm{Tr}_1^{\delta_2}(C)\Big] \,\tilde{\otimes}\, \frac{\mathrm{Tr}_2^{\delta_2}(C)}{\mathrm{Tr}^{\delta_1}(\widehat{C}_N)} \\
&+ \frac{\mathrm{Tr}_1^{\delta_2}(C) \,\tilde{\otimes}\, \mathrm{Tr}_2^{\delta_2}(C)}{\mathrm{Tr}^{\delta_2}(C)\mathrm{Tr}^{\delta_1}(\widehat{C}_N)} \Big[\mathrm{Tr}^{\delta_2}(C) - \mathrm{Tr}^{\delta_1}(\widehat{C}_N)\Big].
\end{aligned} \tag{A.40}$$

Proposition 14. *Let δ be such that $Tr^{\delta}(C) \neq 0$ and let assumption (A1) hold, then $\left|\!\left|\!\left| \widehat{C}(\delta) - C(\delta) \right|\!\right|\!\right|_2^2 = \mathcal{O}_P(N^{-1})$.*

Proof. Firstly, note that

$$\left|\!\left|\!\left| \widehat{C}(\delta) - C(\delta) \right|\!\right|\!\right|_2 \leq \left|\!\left|\!\left| \widehat{A}(\delta) - A(\delta) \right|\!\right|\!\right|_2 + \left|\!\left|\!\left| \mathrm{Ta}\Big(\widehat{A}(\delta) - A(\delta)\Big) \right|\!\right|\!\right|_2 + \left|\!\left|\!\left| \widehat{C}_N - C \right|\!\right|\!\right|_2.$$

Note that $\left|\!\left|\!\left| \mathrm{Ta}\Big(\widehat{A}(\delta) - A(\delta)\Big) \right|\!\right|\!\right|_2 \leq \left|\!\left|\!\left| \widehat{A}(\delta) - A(\delta) \right|\!\right|\!\right|_2$ due to Toeplitz averaging being a linear projection. Hence we only need to bound the difference between the separable parts, for which we use the linearization formula (A.40):

$$\begin{aligned}\left|\!\left|\!\left|\widehat{A}(\delta) - A(\delta)\right|\!\right|\!\right|_2 &\leq \left|\!\left|\!\left|\widehat{A}(\delta) - A(\delta)\right|\!\right|\!\right|_1 \\ &\leq \frac{\left|\!\left|\!\left|\mathrm{Tr}_1^\delta(\widehat{C}_N)\right|\!\right|\!\right|_1}{|\mathrm{Tr}^\delta(\widehat{C}_N)|}\left|\!\left|\!\left|\mathrm{Tr}_2^\delta(\widehat{C}_N - C)\right|\!\right|\!\right|_1 + \left|\!\left|\!\left|\mathrm{Tr}_1^\delta(\widehat{C}_N - C)\right|\!\right|\!\right|_1 \frac{\left|\!\left|\!\left|\mathrm{Tr}_2^\delta(C)\right|\!\right|\!\right|_1}{|\mathrm{Tr}^\delta(\widehat{C}_N)|} \\ &\qquad + \frac{\left|\!\left|\!\left|\mathrm{Tr}_1^\delta(C)\right|\!\right|\!\right|_1 \left|\!\left|\!\left|\mathrm{Tr}_2^\delta(C)\right|\!\right|\!\right|_1}{|\mathrm{Tr}^\delta(C)\mathrm{Tr}^\delta(\widehat{C}_N)|}\left|\mathrm{Tr}^\delta(\widehat{C}_N - C)\right|.\end{aligned} \tag{A.41}$$

From (2.7), we have

$$\left|\!\left|\!\left|\mathrm{Tr}_2^\delta(\widehat{C}_N - C)\right|\!\right|\!\right|_1 \leq \left|\!\left|\!\left|\widehat{C}_N - C\right|\!\right|\!\right|_1 = \mathcal{O}_P(N^{-1/2})$$

since the CLT for $\widehat{C}_N$ holds (Mas, 2006), and similarly for $\left|\!\left|\!\left|\mathrm{Tr}_1^\delta(\widehat{C}_N - C)\right|\!\right|\!\right|_1$ and $\left|\mathrm{Tr}^\delta(\widehat{C}_N - C)\right|$. The statement then follows upon noticing that the numerators on the right hand side of (A.41) are obviously bounded while the denominators are bounded away from zero for N large enough. □

Proposition 15. *Let δ be such that $Tr^\delta(C) \neq 0$ and let assumption (A1) hold, then $\widehat{\Xi}(\delta) = \Xi(\delta) - |||C|||_2^2 + \mathcal{O}_P(N^{-1/2})$.*

Proof. Instead of the empirical objective, we will first work with a slightly modified, biased version of it:

$$\widetilde{\Xi}(\delta) = \left|\!\left|\!\left|\widehat{C}(\delta)\right|\!\right|\!\right|_2^2 - \frac{2}{N}\sum_{n=1}^N \langle X_n, \widehat{C}(\delta) X_n\rangle$$

By adding and subtracting $\left|\!\left|\!\left|\widehat{C}(\delta) - C\right|\!\right|\!\right|_2^2 = \left|\!\left|\!\left|\widehat{C}(\delta)\right|\!\right|\!\right|_2^2 - 2\langle \widehat{C}(\delta), C\rangle - |||C|||_2^2$, we obtain

$$\begin{aligned}&\left|\widetilde{\Xi}(\delta) + |||C|||_2^2 - \Xi(\delta)\right| = \\ &\quad = \left|\frac{2}{N}\sum_{n=1}^N \langle \widehat{C}(\delta), X_n \otimes X_n\rangle - 2\langle \widehat{C}(\delta), C\rangle + \left|\!\left|\!\left|\widehat{C}(\delta) - C\right|\!\right|\!\right|_2^2 - |||C(\delta) - C|||_2^2\right|,\end{aligned}$$

and from the triangle inequality we now have

$$\begin{aligned}\left|\widetilde{\Xi}(\delta) + \|C\|_2 - \Xi(\delta)\right| &\leq \left|\frac{2}{N}\sum_{n=1}^N \langle \widehat{C}(\delta), X_n \otimes X_n\rangle - 2\langle \widehat{C}(\delta), C\rangle\right| \\ &\qquad + \left|\left|\!\left|\!\left|\widehat{C}(\delta) - C\right|\!\right|\!\right|_2^2 - |||C(\delta) - C|||_2^2\right| =: (I) + (II).\end{aligned}$$

The first term can be bounded by Cauchy-Schwarz inequality

$$(I) = 2\left|\langle \widehat{C}(\delta), \widehat{C}_N - C\rangle\right| \leq 2\left|\left|\left|\widehat{C}(\delta)\right|\right|\right|_2 \left|\left|\left|\widehat{C}_N - C\right|\right|\right|_2 = \mathcal{O}_P(N^{-1/2}),$$

whereas the second term can be bounded similarly after using the mean value theorem:

$$(II) = 2\langle \Gamma - C, \widehat{C}(\delta) - C(\delta)\rangle \leq 2|||\Gamma - C|||_2 \left|\left|\left|\widehat{C}(\delta) - C(\delta)\right|\right|\right|,$$

where Γ is between $\widehat{C}(\delta)$ and $C(\delta)$. Hence the term (II) is also $\mathcal{O}_P(N^{-1/2})$ according to the previous proposition.

Now it remains to show that the bias introduced by working with $\widetilde{\Xi}(\delta)$ instead of $\widehat{\Xi}(\delta)$ is asymptotically negligible. For that, it suffices to show that

$$\left|\frac{1}{N}\sum_{n=1}^{N}\langle \widehat{C}(\delta), X_n \otimes X_n\rangle - \frac{1}{N}\sum_{n=1}^{N}\langle \widehat{C}_{-n}(\delta), X_n \otimes X_n\rangle\right| = \mathcal{O}_P(N^{-1}). \tag{A.42}$$

The previous expression can be bounded as

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N}\langle \widehat{C}(\delta) - \widehat{C}_{-n}(\delta), X_n \otimes X_n\rangle\right| &\leq \frac{1}{N}\sum_{n=1}^{N}\left|\langle \widehat{C}(\delta) - \widehat{C}_{-n}(\delta), X_n \otimes X_n\rangle\right| \\ &\leq \frac{1}{N}\sum_{n=1}^{N}\left|\left|\left|\widehat{C}(\delta) - \widehat{C}_{-n}(\delta)\right|\right|\right|_2 \|X_n\|_2^2.\end{aligned}$$

Using the linearization argument (A.40) again like in the proof of the previous proposition, we obtain

$$\left|\left|\left|\widehat{C}(\delta) - \widehat{C}_{-n}(\delta)\right|\right|\right|_2 \leq \left[const + \mathcal{O}_P(N^{-1/2})\right]\left|\left|\left|\widehat{C}_N - \widehat{C}_{-n}\right|\right|\right|_1$$

where $\widehat{C}_{-n}$ is the empirical covariance estimator without the n-th observation. Since

$$\widehat{C}_N - \widehat{C}_{-n} = \frac{1}{N}X_n \otimes X_n + \frac{1}{N(N-1)}\sum_{j\neq n} X_j \otimes X_j$$

for any $n = 1, \ldots, N$, we have from the triangle inequality that

$$\left|\left|\left|\widehat{C}_N - \widehat{C}_{-n}\right|\right|\right|_1 \leq \frac{1}{N}\|X_n\|_2^2 + \frac{1}{N(N-1)}\sum_{j\neq n}\|X_j\|_2^2.$$

Overall, we have that the left-hand size of (A.42) is bounded by $\left[const + \mathcal{O}_P(N^{-1/2})\right]\frac{1}{N}D_N$, where

$$D_N = \frac{1}{N}\sum_{n=1}^{N}\|X_n\|_2^4 + \frac{1}{N(N-1)}\sum_{n=1}^{N}\sum_{j\neq N}\|X_n\|_2^2\|X_j\|_2^2$$

which is $o_P(1)$ from the law of large numbers. Hence we get (A.42) and the proof is complete. □

According to the previous proposition, the empirical objective is consistent for the theoretical objective up to a constant. And the constant, though unknown, does not affect the bandwidth choice. This leads to the rates of convergence of our estimators with adaptively chosen bandwidth as stated in Theorem 5.

Proof of Theorem 5. We have from the triangle inequality

$$\vert\vert\vert \widehat{A}(\widehat{\delta}) - A \vert\vert\vert_2 \leq \vert\vert\vert \widehat{A}(\widehat{\delta}) - A(\widehat{\delta}) \vert\vert\vert_2 + \vert\vert\vert A(\widehat{\delta}) - A \vert\vert\vert_2 .$$

Due to our assumptions the separable-plus-banded model holds with a certain $\delta^\star$ and there exists at least one $\delta \in \Delta$ such that the separable-plus-banded model holds with δ. On the other hand, for any $\delta < \delta^\star$ the separable-plus-banded model does not hold and hence $\Xi(\delta) > \Xi(\delta^\star)$. Therefore, due to Proposition 15, there exists N_0 such that for all $N \geq N_0$ we have $\widehat{\delta} \geq \delta^\star$. Thus $\vert\vert\vert A(\widehat{\delta}) - A \vert\vert\vert_2 = 0$ for all $N \geq N_0$.

Secondly, we observe from the proof of Proposition 14 that $\vert\vert\vert \widehat{A}(\delta) - A(\delta) \vert\vert\vert_2 = \mathcal{O}_p(N^{-1/2})$ for any $\delta \in \Delta$ such that $\delta \geq \delta^\star$, hence also for $\widehat{\delta}$, and the proof is complete.

The assertion for the banded part follows easily using the previous part of the proof and triangle inequalities:

$$\vert\vert\vert \widehat{B}(\widehat{\delta}) - B \vert\vert\vert_2 \leq \vert\vert\vert \widehat{B}(\widehat{\delta}) - B(\widehat{\delta}) \vert\vert\vert_2 + \vert\vert\vert B(\widehat{\delta}) - B \vert\vert\vert_2 ,$$

where

$$\begin{aligned} \vert\vert\vert \widehat{B}(\widehat{\delta}) - B(\widehat{\delta}) \vert\vert\vert_2 &= \vert\vert\vert \mathrm{Ta}\Big(\widehat{C}_N - \widehat{A}(\widehat{\delta}) - C + A(\widehat{\delta})\Big) \vert\vert\vert_2 \leq \vert\vert\vert \widehat{C}_N - \widehat{A}(\widehat{\delta}) - C + A(\widehat{\delta}) \vert\vert\vert_2 \\ &\leq \vert\vert\vert \widehat{C}_N - C \vert\vert\vert_2 + \vert\vert\vert \widehat{A}(\widehat{\delta}) - A(\widehat{\delta}) \vert\vert\vert_2 = \mathcal{O}_P(N^{-1/2}), \end{aligned}$$

and similarly

$$\vert\vert\vert B(\widehat{\delta}) - B \vert\vert\vert_2 \leq \vert\vert\vert \widehat{C}_N - C \vert\vert\vert_2 + \vert\vert\vert A(\widehat{\delta}) - A \vert\vert\vert_2 = \mathcal{O}_P(N^{-1/2}).$$

□

In case that the separable-plus-banded model does not hold, i.e. C does not posses the separable-plus-banded structure for any δ, there still exists $\delta_0 \in \Delta$ such that

$$\Xi(\delta_0) = \min_{\delta \in \Delta} \Xi(\delta),$$

and the same argument as the one in the previous proof yields that $\widehat{C}(\widehat{\delta})$ is root-n consistent for $C(\delta_0)$. In this instance, $C(\delta_0) \neq C$, but $C(\delta_0)$ is the best separable-plus-banded proxy to C, which can be obtained by the proposed estimation methodology based on shifted partial tracing.

Proof of Proposition 9

We begin with the asymptotic distribution for $\widehat{A}(\widetilde{\delta})$, modifying the proof of Theorem 3.

Let us denote by $\delta_m^\star$ the smallest of such bandwidths in Δ which is larger than $\delta^\star$. By Theorem 3, we know that $\sqrt{N}\Big(\widehat{A}(\delta_m^\star) - A\Big)$ is asymptotically Gaussian and mean-zero. Since

$$\sqrt{N}\Big(\widehat{A}(\widetilde{\delta}) - A\Big) = \sqrt{N}\Big(\widehat{A}(\widetilde{\delta}) - \widehat{A}(\delta_m^\star)\Big) + \sqrt{N}\Big(\widehat{A}(\delta_m^\star) - A\Big),$$

we only need to show that $\sqrt{N}\Big(\widehat{A}(\widehat{\delta}) - \widehat{A}(\delta_m^\star)\Big)$ converges to zero in probability.

Let us denote $\widetilde{\Delta} := \{\delta \in \Delta; \delta < \delta_m^\star\}$. Since the separable-plus-banded model holds, and $\delta_m^\star$ is the smallest bandwidth in Δ such that B is banded by this bandwidth, it must be $|\!|\!| C(\delta) - C |\!|\!|_2 > 0$ for all $\delta \in \widetilde{\Delta}$.

Now, fix any $\epsilon > 0$, and observe that

$$P\left(\Big|\sqrt{N}\Big(\widehat{A}(\widetilde{\delta}) - \widehat{A}(\delta_m^\star)\Big)\Big| > \epsilon\right) \leq P\left(\widetilde{\delta} \neq \delta_m^\star\right) = P\left(\operatorname*{arg\,min}_{\delta\in\Delta} \widehat{\Xi}_\tau(\delta) \neq \operatorname*{arg\,min}_{\delta\in\Delta} \Xi_\tau(\delta)\right). \tag{A.43}$$

Let $\alpha > 0$ be arbitrary. For any j such that $\delta_j \neq \delta_m^\star$ there exists $N_j \in \mathbb{N}$ such that for all $N \geq N_j$ we have

$$\Big|\widehat{\Xi}_\tau(\delta_j) + |\!|\!| C |\!|\!|_2^2 - \Xi_\tau(\delta_j)\Big| < \alpha.$$

Taking $N_0 := \max N_j$ and $\alpha := \tau$ we obtain that the probability in (A.43) is equal to zero for any $N \geq N_0$ and the proof is thus complete.

The proof for $\widehat{B}(\widetilde{\delta})$ is an equivalent modification of the proof of Theorem 3.

Proof of Theorem 6

We begin by proving several claims. All of the four claims below hold uniformly in K for any δ such that $\mathrm{Tr}^\delta(A) \neq 0$, and are proven sequentially.

Claim 1: $\left|\!\left|\!\left| \widehat{C}_N^K - C^K \right|\!\right|\!\right|_\star^2 = \mathcal{O}_P(N^{-1})$

Similarly to the proof of Theorem 4, we calculate

$$\mathbb{E}\left|\!\left|\!\left| \widehat{C}_N^K - C^K \right|\!\right|\!\right|_\star^2 = \frac{1}{K^4} \sum_{(i,j)\neq(k,l)} \mathbb{E}\left| \frac{1}{N} \sum_{n=1}^N \Big(\underbrace{\widetilde{\mathbf{X}}_n^K[i,j]\widetilde{\mathbf{X}}_n^K[k,l] - \mathbb{E}\mathbf{X}_n^K[i,j]\mathbf{X}_n^K[k,l]}_{=:Z_{n,ijkl}} \Big) \right|^2.$$

For a fixed i,j,k,l, $Z_{n,ijkl}$ are zero-mean (this is the reason why we need to remove the diagonal from the norm) i.i.d. random variables and thus

$$\mathbb{E}|\!|\!|\widehat{C}_N^K - C^K|\!|\!|_\star^2 = \frac{1}{N}\frac{1}{K^4}\sum_{(i,j)\neq(k,l)} \mathbb{E}\left|Z_{n,ijkl}\right|^2,$$

which is from the parallelogram law equal to

$$\frac{4}{N}\frac{1}{K^4}\sum_{(i,j)\neq(k,l)}\left\{\mathbb{E}\left|\mathbf{X}^K[i,j]\mathbf{X}^K[k,l] - \mathbb{E}\mathbf{X}^K[i,j]\mathbf{X}^K[k,l]\right|^2\right.$$
$$\left. + \mathbb{E}\left|\mathbf{E}^K[i,j]\mathbf{X}^K[k,l]\right|^2 + \mathbb{E}\left|\mathbf{X}^K[i,j]\mathbf{E}^K[k,l]\right|^2 + \mathbb{E}\left|\mathbf{E}^K[i,j]\mathbf{E}^K[k,l]\right|^2\right\}.$$

The first term in the parentheses is bounded by S_1, the second and third are bounded by $\sigma^2 S_2$, and the final term is bounded by σ^4, which yields the claim.

Claim 2: $|\!|\!|\widehat{A}^K(\delta) - A^K(\delta)|\!|\!|_2^2 = \mathcal{O}_P(N^{-1})$

Here we use the linearization argument (A.40) and proceed exactly like in Theorem 4.

Claim 3: $|\!|\!|\widehat{C}^K(\delta) - C^K(\delta)|\!|\!|_\star^2 = \mathcal{O}_P(N^{-1})$

We have

$$\begin{aligned}
&|\!|\!|\widehat{C}^K(\delta) - C^K(\delta)|\!|\!|_\star \leq \\
&\qquad \geq |\!|\!|\widehat{A}^K(\delta) - A^K(\delta)|\!|\!|_\star + |\!|\!|\mathrm{Ta}\left(\widehat{C}_N^K - \widehat{A}^K(\delta) - C^K + A^K(\delta)\right)|\!|\!|_\star \\
&\qquad \leq |\!|\!|\widehat{A}^K(\delta) - A^K(\delta)|\!|\!|_\star + |\!|\!|\widehat{C}_N^K - \widehat{A}^K(\delta) - C^K + A^K(\delta)|\!|\!|_\star \\
&\qquad \leq 2|\!|\!|\widehat{A}^K(\delta) - A^K(\delta)|\!|\!|_\star + |\!|\!|\widehat{C}_N^K - C^K|\!|\!|_\star,
\end{aligned}$$

where we utilized the triangle inequality in the first and last inequality, while the second inequality follows from $\mathrm{Ta}(\cdot)$ being a linear projection. Now, the second term is bounded by Claim 2, while the first term is bounded by Claim 1 and the fact that

$$|\!|\!|\widehat{A}^K(\delta) - A^K(\delta)|\!|\!|_\star \leq |\!|\!|\widehat{A}^K(\delta) - A^K(\delta)|\!|\!|_2.$$

Claim 4: $\widehat{\Xi}^K(\delta) = \Xi^K(\delta) - |\!|\!|C^K|\!|\!|_\star^2 + \mathcal{O}_P(N^{-1/2})$

We first work with a biased version of the empirical objective $\widehat{\Xi}^K$, i.e.

$$\widetilde{\Xi}^K(\delta) = |\!|\!|\widehat{C}^K(\delta)|\!|\!|_\star^2 - \frac{2}{N}\sum_{n=1}^{N}\langle X_n^K, \widehat{C}^K(\delta)X_n^K\rangle_\star,$$

and show Claim 4 with $\widehat{\Xi}^K$ replaced by $\widetilde{\Xi}^K$. For this, we bound similarly to the proof of Proposition 15:

$$\left|\widetilde{\Xi}^K(\delta) + \left|\!\left|\!\left| C^K \right|\!\right|\!\right|_\star^2 - \Xi^K(\delta)\right| \leq 2\left|\langle \widehat{C}^K(\delta), \widehat{C}_N^K - C\rangle_\star\right| \\ + \left| \left|\!\left|\!\left| \widehat{C}^K(\delta) - C^K \right|\!\right|\!\right|_\star - \left|\!\left|\!\left| C^K(\delta) - C^K \right|\!\right|\!\right|_\star \right|$$

The Cauchy-Schwarz inequality still holds for the semi-inner-product (Conway, 2019), which allows us to bound the first term using Claim 1. For the second term, note that the mean value theorem can still be used, since the Fréchet derivative is a linear operation and the semi-norm is *consistent* with the semi-inner-product. Hence using the mean value theorem, and the Cauchy-Schwarz inequality, we have

$$\left| \left|\!\left|\!\left| \widehat{C}^K(\delta) - C^K \right|\!\right|\!\right|_\star - \left|\!\left|\!\left| \widehat{C}^K(\delta) - C^K \right|\!\right|\!\right|_\star \right| = 2\langle \Gamma - C^K, \widehat{C}^K(\delta) - C^K(\delta)\rangle_\star \\ \leq 2 \left|\!\left|\!\left| \Gamma - C^K \right|\!\right|\!\right|_\star \left|\!\left|\!\left| \widehat{C}^K(\delta) - C^K(\delta) \right|\!\right|\!\right|_\star.$$

Hence the bound follows from Claim 3.

It remains to show that the introduced bias is asymptotically negligible, i.e. to bound $|\widehat{\Xi}^K(\delta) - \widetilde{\Xi}^K(\delta)|$. For this, we use triangle and Cauchy-Schwarz inequality:

$$\left| \frac{1}{N} \sum_{n=1}^{N} \langle \widehat{C}(\delta) - \widehat{C}_{-n}(\delta), \widetilde{X}_n \otimes \widetilde{X}_n \rangle_\star \right| \leq \frac{1}{N} \sum_{n=1}^{N} \left| \langle \widehat{C}(\delta) - \widehat{C}_{-n}(\delta), \widetilde{X}_n \otimes \widetilde{X}_n \rangle_\star \right| \\ \leq \frac{1}{N} \sum_{n=1}^{N} \left|\!\left|\!\left| \widehat{C}(\delta) - \widehat{C}_{-n}(\delta) \right|\!\right|\!\right|_\star \| \widetilde{X}_n \otimes \widetilde{X}_n \|_\star.$$

Now, since $|\!|\!|\cdot|\!|\!|_\star \leq |\!|\!|\cdot|\!|\!|_2$, the remainder of the proof is exactly the same as the end of the proof of Proposition 15.

Now we can prove the Theorem itself. Using the parallelogram law, we have

$$\left|\!\left|\!\left| \widehat{A}^K(\widehat{\delta}) - A \right|\!\right|\!\right|_2^2 \leq 4\left[\left|\!\left|\!\left| \widehat{A}^K(\widehat{\delta}) - A^K(\widehat{\delta}) \right|\!\right|\!\right|_2^2 + \left|\!\left|\!\left| A^K(\widehat{\delta}) - A^K \right|\!\right|\!\right|_2^2 + \left|\!\left|\!\left| A^K - A \right|\!\right|\!\right|_2^2 \right].$$

The first term in the brackets is bounded by Claim 2, while the last term in the brackets correspond to the bias and can be treated the same as in the proof of Theorem 4. It remains to show that the middle term in the brackets is equal to zero for all sufficiently large N. But this the same way as the first paragraph of the proof of Theorem 5.

Separable Component Decomposition

Proof of Theorem 7

Before proving the asymptotic theorems, we need perturbation bounds for the separable component decomposition. From (3.4), $C \in \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$ has the separable component decomposition (SCD)

$$C = \sum_{i=1}^{\infty} \sigma_i A_i \tilde{\otimes} B_i,$$

where $\{A_i\}_{i\geq 1}$ (resp. $\{B_i\}_{i\geq 1}$) is an orthonormal basis (ONB) of $\mathcal{S}_2(\mathcal{H}_1)$ (resp. $\mathcal{S}_2(\mathcal{H}_2)$), and $|\sigma_1| \geq |\sigma_2| \geq \cdots \geq 0$. We use the notation $\mathcal{C}$ to indicate the element of $\mathcal{S}_2(\mathcal{H}_2 \otimes \mathcal{H}_2, \mathcal{H}_1 \otimes \mathcal{H}_1)$ that is isomorphic to C (see (3.2)–(3.4)). The following lemma gives perturbation bounds for the components of the SCD.

Lemma 8 (Perturbation Bounds for SCD)**.**
Let $C = \sum_{i=1}^{\infty} \sigma_i A_i \tilde{\otimes} B_i$ *and* $\widetilde{C} = \sum_{i=1}^{\infty} \widetilde{\sigma}_i \widetilde{A}_i \tilde{\otimes} \widetilde{B}_i$ *be SCDs of* C *and* $\widetilde{C}$*. Also suppose that* $\sigma_1 > \sigma_2 > \cdots \geq 0$*, and* $\left\langle A_i, \widetilde{A}_i \right\rangle_{\mathcal{S}_2(\mathcal{H}_1)}, \left\langle B_i, \widetilde{B}_i \right\rangle_{\mathcal{S}_2(\mathcal{H}_2)} \geq 0$ *for every* $i = 1, 2, \ldots$ *(adjust the sign of* $\widetilde{\sigma}_i$ *as required). Then,*

(a) $\sup_{i\geq 1} \left|\sigma_i - \widetilde{\sigma}_i\right| \leq \left|\left|\left| C - \widetilde{C} \right|\right|\right|_2$.

(b) For every $i \geq 1$*,*

$$\left|\left|\left| A_i - \widetilde{A}_i \right|\right|\right|_{\mathcal{S}_2(\mathcal{H}_1)} \leq \frac{2\sqrt{2}}{\alpha_i} \left|\left|\left| C - \widetilde{C} \right|\right|\right|_2 \left(|||C|||_2 + \left|\left|\left| \widetilde{C} \right|\right|\right|_2 \right),$$

$$\left|\left|\left| B_i - \widetilde{B}_i \right|\right|\right|_{\mathcal{S}_2(\mathcal{H}_2)} \leq \frac{2\sqrt{2}}{\alpha_i} \left|\left|\left| C - \widetilde{C} \right|\right|\right|_2 \left(|||C|||_2 + \left|\left|\left| \widetilde{C} \right|\right|\right|_2 \right),$$

where $\alpha_i = \min\{\sigma_{i-1}^2 - \sigma_i^2, \sigma_i^2 - \sigma_{i+1}^2\}$*. Here,* $|||\cdot|||_2$ *denotes the Hilbert-Schmidt norm.*

Proof. Note that σ_i (resp. $\widetilde{\sigma}_i$) is the i-th singular value of the operator $\mathcal{C}$ (resp. $\widetilde{\mathcal{C}}$). Following (Bosq, 2012, Lemma 4.2), we get that $\sup_{i\geq 1} \left|\sigma_i - \widetilde{\sigma}_i\right| \leq \left|\left|\left| \mathcal{C} - \widetilde{\mathcal{C}} \right|\right|\right| \leq \left|\left|\left| \mathcal{C} - \widetilde{\mathcal{C}} \right|\right|\right|_2$. Part (a) now follows by noting that $\left|\left|\left| \mathcal{C} - \widetilde{\mathcal{C}} \right|\right|\right|_2 = \left|\left|\left| C - \widetilde{C} \right|\right|\right|_2$, because of the isometry between $\mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$ and $\mathcal{S}_2(\mathcal{H}_2 \otimes \mathcal{H}_2, \mathcal{H}_1 \otimes \mathcal{H}_1)$.

For part (b), recall that A_i (resp. $\widetilde{A}_i$) is isomorphic to e_i (resp. $\widetilde{e}_i$), the i-th right singular element of $\mathcal{C}$ (resp. of $\widetilde{\mathcal{C}}$) (see (3.2)–(3.4)). Now, e_i (resp. $\widetilde{e}_i$) is the i-th eigen-element of $\mathcal{C}\mathcal{C}^\top$ (resp. $\widetilde{\mathcal{C}}\widetilde{\mathcal{C}}^\top$) with corresponding eigenvalue $\lambda_i = \sigma_i^2$ (resp. $\widetilde{\lambda}_i = \widetilde{\sigma}_i^2$). Here, $\mathcal{C}^\top$ (resp. $\widetilde{\mathcal{C}}^\top$) denotes the adjoint of $\mathcal{C}$ (resp. $\widetilde{\mathcal{C}}$). Also, $\langle e_i, \widetilde{e}_i \rangle_{\mathcal{H}_1 \otimes \mathcal{H}_1} = \langle A_i, \widetilde{A}_i \rangle_{\mathcal{S}_2(\mathcal{H}_1)} \geq 0$. Now, using a perturbation bound on the eigen-elements of operators (Bosq, 2012, Lemma 4.3), we get

$$\|e_i - \widetilde{e}_i\|_{\mathcal{H}_1 \otimes \mathcal{H}_1} \leq \frac{2\sqrt{2}}{\alpha_i} \left|\!\left|\!\left| \mathcal{C}\mathcal{C}^\top - \widetilde{\mathcal{C}}\widetilde{\mathcal{C}}^\top \right|\!\right|\!\right| \leq \frac{2\sqrt{2}}{\alpha_i} \left|\!\left|\!\left| \mathcal{C}\mathcal{C}^\top - \widetilde{\mathcal{C}}\widetilde{\mathcal{C}}^\top \right|\!\right|\!\right|_2,$$

where $\alpha_i = \min\{\sigma_{i-1}^2 - \sigma_i^2, \sigma_i^2 - \sigma_{i+1}^2\}$. Now,

$$\begin{aligned} \left|\!\left|\!\left| \mathcal{C}\mathcal{C}^\top - \widetilde{\mathcal{C}}\widetilde{\mathcal{C}}^\top \right|\!\right|\!\right|_2 = \left|\!\left|\!\left| \mathcal{C}\left(\mathcal{C} - \widetilde{\mathcal{C}}\right)^\top + \left(\mathcal{C} - \widetilde{\mathcal{C}}\right)\widetilde{\mathcal{C}}^\top \right|\!\right|\!\right|_2 &\leq \left|\!\left|\!\left| \mathcal{C} - \widetilde{\mathcal{C}} \right|\!\right|\!\right|_2 \left(|||\mathcal{C}|||_2 + \left|\!\left|\!\left| \widetilde{\mathcal{C}} \right|\!\right|\!\right|_2 \right) \\ &= \left|\!\left|\!\left| C - \widetilde{C} \right|\!\right|\!\right|_2 \left(|||C|||_2 + \left|\!\left|\!\left| \widetilde{C} \right|\!\right|\!\right|_2 \right), \end{aligned}$$

where we have used: (i) the triangle inequality for the Hilbert-Schmidt norm, (ii) the fact that the Hilbert-Schmidt norm of an operator and its adjoint are the same, and (iii) the isometry between $\mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2)$ and $\mathcal{S}_2(\mathcal{H}_2 \otimes \mathcal{H}_2, \mathcal{H}_1 \otimes \mathcal{H}_1)$. By noting that $\left|\!\left|\!\left| A_i - \widetilde{A}_i \right|\!\right|\!\right|_{\mathcal{S}_2(\mathcal{H}_1)} = \|e_i - \widetilde{e}_i\|_{\mathcal{H}_1 \otimes \mathcal{H}_1}$, the upper bound on $\left|\!\left|\!\left| A_i - \widetilde{A}_i \right|\!\right|\!\right|_{\mathcal{S}_2(\mathcal{H}_1)}$ follows. The bound on $\left|\!\left|\!\left| B_i - \widetilde{B}_i \right|\!\right|\!\right|_{\mathcal{S}_2(\mathcal{H}_2)}$ can be proved similarly. □

The following lemma gives us perturbation bound for the best R-separable approximation of Hilbert-Schmidt operators.

Lemma 9 (Perturbation Bound for Best R-separable Approximation)**.** *Let C and $\widetilde{C}$ be two Hilbert-Schmidt operators on $\mathcal{H}_1 \otimes \mathcal{H}_2$, with SCD $C = \sum_{r=1}^\infty \sigma_r A_r \tilde{\otimes} B_r$ and $\widetilde{C} = \sum_{r=1}^\infty \widetilde{\sigma}_r \widetilde{A}_r \tilde{\otimes} \widetilde{B}_r$, respectively. Let C_R and $\widetilde{C}_R$ be the best R-separable approximations of C and $\widetilde{C}$, respectively. Then,*

$$\left|\!\left|\!\left| C_R - \widetilde{C}_R \right|\!\right|\!\right|_2 \leq \left\{ 4\sqrt{2}\left(|||C|||_2 + \left|\!\left|\!\left| \widetilde{C} \right|\!\right|\!\right|_2 \right) \sum_{r=1}^R \frac{\sigma_r}{\alpha_r} + 1 \right\} \left|\!\left|\!\left| C - \widetilde{C} \right|\!\right|\!\right|_2,$$

where $\alpha_r = \min\{\sigma_{r-1}^2 - \sigma_r^2, \sigma_r^2 - \sigma_{r+1}^2\}$.

Proof. Note that C_R and $\widetilde{C}_R$ have SCD $C_R = \sum_{i=1}^R \sigma_i A_i \tilde{\otimes} B_i$ and $\widetilde{C}_R = \sum_{i=1}^R \widetilde{\sigma}_i \widetilde{A}_i \tilde{\otimes} \widetilde{B}_i$, respectively. W.l.o.g. we assume that $\left\langle A_i, \widetilde{A}_i \right\rangle_{\mathcal{S}_2(\mathcal{H}_1)}, \left\langle B_i, \widetilde{B}_i \right\rangle_{\mathcal{S}_2(\mathcal{H}_2)} \geq 0$ for every $i = 1, \ldots, R$ (if not, one can change the sign of $\widetilde{A}_i$ or $\widetilde{B}_i$, and adjust the sign of $\widetilde{\sigma}_i$ as required). Thus,

$$\begin{aligned} \left|\!\left|\!\left| C_R - \widetilde{C}_R \right|\!\right|\!\right|_2 &= \left|\!\left|\!\left| \sum_{i=1}^R \sigma_i A_i \tilde{\otimes} B_i - \sum_{i=1}^R \widetilde{\sigma}_i \widetilde{A}_i \tilde{\otimes} \widetilde{B}_i \right|\!\right|\!\right|_2 \\ &= \left|\!\left|\!\left| \sum_{i=1}^R \sigma_i \left(A_i \tilde{\otimes} B_i - \widetilde{A}_i \tilde{\otimes} \widetilde{B}_i \right) + \sum_{i=1}^R \left(\sigma_i - \widetilde{\sigma}_i \right) \widetilde{A}_i \tilde{\otimes} \widetilde{B}_i \right|\!\right|\!\right|_2 \\ &\leq \left|\!\left|\!\left| \sum_{i=1}^R \sigma_i \left(A_i \tilde{\otimes} B_i - \widetilde{A}_i \tilde{\otimes} \widetilde{B}_i \right) \right|\!\right|\!\right|_2 + \left|\!\left|\!\left| \sum_{i=1}^R \left(\sigma_i - \widetilde{\sigma}_i \right) \widetilde{A}_i \tilde{\otimes} \widetilde{B}_i \right|\!\right|\!\right|_2. \qquad \text{(A.44)} \end{aligned}$$

Now, $\left|\!\left|\!\left|\sum_{i=1}^{R}\left(\sigma_i-\widetilde{\sigma}_i\right)\widetilde{A}_i\,\tilde{\otimes}\,\widetilde{B}_i\right|\!\right|\!\right|_2^2=\sum_{i=1}^{R}\left(\sigma_i-\widetilde{\sigma}_i\right)^2\leq\sum_{i=1}^{\infty}\left(\sigma_i-\widetilde{\sigma}_i\right)^2$ which, by von Neumann's trace inequality, is bounded by $\left|\!\left|\!\left|C-\widetilde{C}\right|\!\right|\!\right|_2^2$ (Hsing and Eubank, 2015, Theorem 4.5.3). On the other hand,

$$
\begin{aligned}
\left|\!\left|\!\left|\sum_{i=1}^{R}\sigma_i\left(A_i\,\tilde{\otimes}\,B_i-\widetilde{A}_i\,\tilde{\otimes}\,\widetilde{B}_i\right)\right|\!\right|\!\right|_2 &\leq\sum_{i=1}^{R}\sigma_i\left|\!\left|\!\left|A_i\,\tilde{\otimes}\,B_i-\widetilde{A}_i\,\tilde{\otimes}\,\widetilde{B}_i\right|\!\right|\!\right|_2\\
&=\sum_{i=1}^{R}\sigma_i\left|\!\left|\!\left|A_i\,\tilde{\otimes}\left(B_i-\widetilde{B}_i\right)+\left(A_i-\widetilde{A}_i\right)\tilde{\otimes}\,\widetilde{B}_i\right|\!\right|\!\right|_2\\
&\leq\sum_{i=1}^{R}\sigma_i\left(\left|\!\left|\!\left|A_i-\widetilde{A}_i\right|\!\right|\!\right|_{\mathcal{S}_2(\mathcal{H}_1)}+\left|\!\left|\!\left|B_i-\widetilde{B}_i\right|\!\right|\!\right|_{\mathcal{S}_2(\mathcal{H}_2)}\right)\\
&\leq 4\sqrt{2}\left|\!\left|\!\left|C-\widetilde{C}\right|\!\right|\!\right|_2\left(|\!|\!|C|\!|\!|_2+\left|\!\left|\!\left|\widetilde{C}\right|\!\right|\!\right|_2\right)\sum_{i=1}^{R}\frac{\sigma_i}{\alpha_i}
\end{aligned}
$$

where the last inequality follows by part (b) of Lemma 8. The lemma is proved upon using these inequalities in conjunction with (A.44). □

Now, we can use the perturbation bounds to prove the asymptotic theorems. We begin with the fully observed case.

Proof of Theorem 7. To bound the error of the estimator, we use the following *bias-variance-type* decomposition

$$
\left|\!\left|\!\left|\widehat{C}_{R,N}-C\right|\!\right|\!\right|_2\leq\left|\!\left|\!\left|\widehat{C}_{R,N}-C_R\right|\!\right|\!\right|_2+|\!|\!|C_R-C|\!|\!|_2, \tag{A.45}
$$

where C_R is the best R-separable approximation of C. If C has SCD $C=\sum_{i=1}^{\infty}\sigma_iA_i\,\tilde{\otimes}\,B_i$, then C_R has SCD $C_R=\sum_{i=1}^{R}\sigma_iA_i\,\tilde{\otimes}\,B_i$, and

$$
|\!|\!|C-C_R|\!|\!|_2^2=\sum_{i=R+1}^{\infty}\sigma_i^2. \tag{A.46}
$$

For the first part, we use the perturbation bound from Lemma 9, to get

$$
\left|\!\left|\!\left|\widehat{C}_{R,N}-C_R\right|\!\right|\!\right|_2\leq\left\{4\sqrt{2}\left(\left|\!\left|\!\left|\widehat{C}_N\right|\!\right|\!\right|_2+|\!|\!|C|\!|\!|_2\right)\sum_{r=1}^{R}\frac{\sigma_r}{\alpha_r}+1\right\}\left|\!\left|\!\left|\widehat{C}_N-C\right|\!\right|\!\right|_2, \tag{A.47}
$$

where $\alpha_r=\min\{\sigma_{r-1}^2-\sigma_r^2,\sigma_r^2-\sigma_{r+1}^2\}$.

Since $\mathbb{E}(\|X\|^4)$ is finite, $\left|\!\left|\!\left|\widehat{C}_N-C\right|\!\right|\!\right|_2=\mathcal{O}_{\mathbb{P}}(N^{-1/2})$ and $\left|\!\left|\!\left|\widehat{C}_N\right|\!\right|\!\right|_2=|\!|\!|C|\!|\!|_2+o_{\mathbb{P}}(1)$. Using these in (A.47), we get that

$$\left|\!\left|\!\left|\widehat{C}_{R,N} - C_R\right|\!\right|\!\right|_2 = \mathcal{O}_{\mathbb{P}}\left(\frac{a_R}{\sqrt{N}}\right), \tag{A.48}$$

where $a_R = |\!|\!|C|\!|\!|_2 \sum_{i=1}^{R}(\sigma_i/\alpha_i)$. The theorem follows by combining (A.46) and (A.48) in (A.45). □

Proof of Theorem 8

Next, we prove Theorem 8. Before doing that, we introduce a notational convention: when the operator A has a kernel that is piecewise constant on the rectangles $\{I_{i,j}^K\}$ (i.e. a "pixelated kernel"), we will write $\|\mathbf{A}\|_{\mathrm{F}}$ for the Frobenius norm of the corresponding tensor of pixel coefficients. This is proportional to the Hilbert-Schmidt norm $|\!|\!|A|\!|\!|_2$ of A. We summarise this in the lemma below, whose straightforward proof we omit, since it is similar to Lemma 4.

Lemma 10. *Let A be an operator with a pixelated kernel*

$$a(t,s,t',s') = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} \mathbf{A}[i,j,k,l]\, \mathbf{1}\{(t,s) \in I_{i,j}^K, (t',s') \in I_{k,l}^K\},$$

where $\mathbf{A} = (\mathbf{A}[i,j,k,l])_{i,j,k,l} \in \mathbb{R}^{K_1\times K_2\times K_1\times K_2}$ is the tensorized version of A. Then,

$$|\!|\!|A|\!|\!|_2 = \frac{1}{K_1K_2}\|\mathbf{A}\|_{\mathrm{F}},$$

where $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2}\mathbf{A}^2[i,j,k,l]}$ is the Frobenius norm of $\mathbf{A}$.

Proof of Theorem 8. We decompose the error of our estimator as

$$\left|\!\left|\!\left|\widehat{C}_{R,N}^K - C\right|\!\right|\!\right|_2 \le \left|\!\left|\!\left|\widehat{C}_{R,N}^K - C_R\right|\!\right|\!\right|_2 + |\!|\!|C_R - C|\!|\!|_2. \tag{A.49}$$

The second term $|\!|\!|C_R - C|\!|\!|_2$ equals $\sqrt{\sum_{r=R+1}^{\infty}\sigma_r^2}$, see (A.46). For the first term, we observe that $\widehat{C}_{R,N}^K$ and C_R are the best R-separable approximations of $\widehat{C}_N^K$ and C, respectively. So, using Lemma 9, we get

$$\left|\!\left|\!\left|\widehat{C}_{R,N}^K - C_R\right|\!\right|\!\right|_2 \le \left\{4\sqrt{2}\left(|\!|\!|C|\!|\!|_2 + \left|\!\left|\!\left|\widehat{C}_N^K\right|\!\right|\!\right|_2\right)\sum_{r=1}^{R}\frac{\sigma_r}{\alpha_r} + 1\right\}\left|\!\left|\!\left|\widehat{C}_N^K - C\right|\!\right|\!\right|_2, \tag{A.50}$$

with $\alpha_r = \min\{\sigma_{r-1}^2 - \sigma_r^2, \sigma_r^2 - \sigma_{r+1}^2\}$. Next, we derive bounds on $\left|\!\left|\!\left|\widehat{C}_N^K - C\right|\!\right|\!\right|_2$. We use the general bound

$$\left\|\left|\widehat{C}_N^K - C\right\|\right|_2 \leq \left\|\left|\widehat{C}_N^K - C^K\right\|\right|_2 + \left\|\left|C^K - C\right\|\right|_2. \tag{A.51}$$

We will now bound these two terms, separately under (S1) and (S2).

Under (S1), recall that C^K is the integral operator with kernel

$$c^K(t,s,t',s') = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} c(t_i^{K_1}, s_j^{K_2}, t_k^{K_1}, s_l^{K_2})\, \mathbf{1}\{(t,s) \in I_{i,j}^K, (t',s') \in I_{k,l}^K\}.$$

Using this, we get that

$$\left\|\left|C^K - C\right\|\right|_2^2 = \iiiint \left\{c^K(t,s,t',s') - c(t,s,t',s')\right\}^2 dtdsdt'ds'$$

is equal to

$$\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} \iiiint_{I_{i,j}^K \times I_{k,l}^K} \left\{c(t_i^{K_1}, s_j^{K_2}, t_k^{K_1}, s_l^{K_2}) - c(t,s,t',s')\right\}^2 dtdsdt'ds'. \tag{A.52}$$

Because of the Lipschitz condition, for $(t,s) \in I_{i,j}^K, (t',s') \in I_{k,l}^K$,

$$\begin{aligned} \left[c(t_i^{K_1}, s_j^{K_2}, t_k^{K_1}, s_l^{K_2}) - c(t,s,t',s')\right]^2 \leq& \\ &\leq L^2\{(t - t_i^{K_1})^2 + (s - s_j^{K_2})^2 + (t' - t_k^{K_1})^2 + (s' - s_l^{K_2})^2\} \\ &\leq L^2\left(\frac{1}{K_1^2} + \frac{1}{K_2^2} + \frac{1}{K_1^2} + \frac{1}{K_2^2}\right) = 2L^2\left(\frac{1}{K_1^2} + \frac{1}{K_2^2}\right). \end{aligned}$$

Plugging this into (A.52) yields

$$\left\|\left|C^K - C\right\|\right|_2^2 \leq 2L^2\left(\frac{1}{K_1^2} + \frac{1}{K_2^2}\right). \tag{A.53}$$

For the first part in (A.51), we observe that $\widehat{C}_N^K$ is the sample covariance of $X_1^K, \ldots, X_N^K \sim X^K$, which are i.i.d. with $\mathbb{E}(X^K) = 0$ and $\operatorname{Var}(X^K) = C^K$. Also, $\widehat{C}_N^K$ and C^K are pixelated operators with discrete versions $\widehat{\mathbf{C}}_N^K$ and $\mathbf{C}^K$, respectively, where $\widehat{\mathbf{C}}_N^K = N^{-1}\sum_{n=1}^N (\mathbf{X}_n^K - \overline{\mathbf{X}}_N^K) \otimes (\mathbf{X}_n^K - \overline{\mathbf{X}}_N^K)$ is the sample variance based on $\mathbf{X}_1^K, \ldots, \mathbf{X}_N^K$ and $\mathbf{C}^K[i,j,k,l] = \operatorname{Cov}\{\mathbf{X}_n^K[i,j], \mathbf{X}_N^K[k,l]\}$ is the discrete version of C^K. So, by Lemma 10,

$$\left\|\left|\widehat{C}_N^K - C^K\right\|\right|_2 = \frac{1}{K_1K_2}\left\|\widehat{\mathbf{C}}_N^K - \mathbf{C}^K\right\|_{\mathrm{F}}. \tag{A.54}$$

For the Frobenius norm, we can write

$$\begin{aligned}\left\|\widehat{\mathbf{C}}_N^K - \mathbf{C}^K\right\|_{\mathrm{F}} &= \left\|\frac{1}{N}\sum_{n=1}^N \mathbf{X}_n^K \otimes \mathbf{X}_n^K - \overline{\mathbf{X}}_N^K \otimes \overline{\mathbf{X}}_N^K - C^K\right\|_{\mathrm{F}} \\ &\leq \left\|\frac{1}{N}\sum_{n=1}^N \mathbf{X}_n^K \otimes \mathbf{X}_n^K - C^K\right\|_{\mathrm{F}} + \left\|\overline{\mathbf{X}}_N^K \otimes \overline{\mathbf{X}}_N^K\right\|_{\mathrm{F}}. \end{aligned} \tag{A.55}$$

Now, $\left\|\overline{\mathbf{X}}_N^K \otimes \overline{\mathbf{X}}_N^K\right\|_{\mathrm{F}} = \left\|\overline{\mathbf{X}}_N^K\right\|_{\mathrm{F}}^2 = \left\|N^{-1}\sum_{n=1}^N \mathbf{X}_n^K\right\|_F^2$, where $\mathbf{X}_n^K$ are i.i.d., zero-mean elements. So,

$$\mathbb{E}\left\|\overline{\mathbf{X}}_N^K \otimes \overline{\mathbf{X}}_N^K\right\|_{\mathrm{F}} = \mathbb{E}\left\|\frac{1}{N}\sum_{n=1}^N \mathbf{X}_n^K\right\|_{\mathrm{F}}^2 = \frac{1}{N}\mathbb{E}\left\|\mathbf{X}^K\right\|_{\mathrm{F}}^2.$$

Again, $\left\|\mathbf{X}^K\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\{\mathbf{X}^K[i,j]\}^2$, so

$$\mathbb{E}\left\|\mathbf{X}^K\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\mathbb{E}\Big(\mathbf{X}^K[i,j]\Big)^2 = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\mathrm{Var}\left(\mathbf{X}^K[i,j]\right).$$

Under our measurement scheme, $\mathbf{X}^K[i,j] = X(t_i^{K_1}, s_j^{K_2})$, so

$$\mathrm{Var}\left(\mathbf{X}^K[i,j]\right) = c(t_i^{K_1}, s_j^{K_2}, t_i^{K_1}, s_j^{K_2}) \leq \sup_{(t,s)\in[0,1]^2} c(t,s,t,s) =: S_1,$$

where S_1 is finite since we assume that X has continuous sample paths. This shows that

$$\mathbb{E}\left\|\overline{\mathbf{X}}_N^K \otimes \overline{\mathbf{X}}_N^K\right\|_{\mathrm{F}} \leq \frac{K_1K_2S_1}{N}. \tag{A.56}$$

Next, we define $\mathbf{Z}_n^K = \mathbf{X}_n^K \otimes \mathbf{X}_n^K - \mathbf{C}^K$. Then, $\mathbf{Z}_1^K, \ldots, \mathbf{Z}_N^K$ are i.i.d., mean centered, which gives

$$\mathbb{E}\left\|\frac{1}{N}\sum_{n=1}^N \mathbf{X}_n^K \otimes \mathbf{X}_n^K - \mathbf{C}^K\right\|_{\mathrm{F}}^2 = \mathbb{E}\left\|\frac{1}{N}\sum_{n=1}^N \mathbf{Z}_n^K\right\|_{\mathrm{F}}^2 = \frac{1}{N}\mathbb{E}\left\|\mathbf{Z}_1^K\right\|_{\mathrm{F}}^2.$$

Now,

$$\begin{aligned}\mathbb{E}\left\|\mathbf{Z}_1^K\right\|_{\mathrm{F}}^2 &= \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2}\mathbb{E}\Big(\mathbf{X}_1^K[i,j]\mathbf{X}_1^K[k,l] - \mathbf{C}^K[i,j,k,l]\Big)^2 \\ &= \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2}\mathrm{Var}\left(\mathbf{X}_1^K[i,j]\mathbf{X}_1^K[k,l]\right) \\ &= \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2}\mathrm{Var}\left(X(t_i^{K_1}, s_j^{K_2})X(t_k^{K_1} s_l^{K_2})\right) \\ &\leq K_1^2K_2^2 \underbrace{\sup_{(t,s,t',s')\in[0,1]^4}\mathrm{Var}\left(X(t,s)X(t',s')\right)}_{=:S_2},\end{aligned}$$

where S_2 is finite since we assume that X has continuous sample paths and finite fourth moment. Thus,

$$\mathbb{E}\left\|\frac{1}{N}\sum_{n=1}^{N}\mathbf{X}_n^K\otimes\mathbf{X}_n^K-\mathbf{C}^K\right\|_{\mathrm{F}}^2\leq\frac{K_1^2K_2^2S_2}{N},$$

which implies that

$$\mathbb{E}\left\|\frac{1}{N}\sum_{n=1}^{N}\mathbf{X}_n^K\otimes\mathbf{X}_n^K-\mathbf{C}^K\right\|_{\mathrm{F}}\leq\sqrt{\mathbb{E}\left\|\frac{1}{N}\sum_{n=1}^{N}\mathbf{X}_n^K\otimes\mathbf{X}_n^K-\mathbf{C}^K\right\|_{\mathrm{F}}^2}\leq\frac{K_1K_2\sqrt{S_2}}{\sqrt{N}}. \quad \text{(A.57)}$$

Combining (A.55), (A.56) and (A.57), we obtain

$$\mathbb{E}\left\|\widehat{\mathbf{C}}_N^K-\mathbf{C}^K\right\|_{\mathrm{F}}\leq\frac{K_1K_2\sqrt{S_2}}{\sqrt{N}}+\frac{K_1K_2S_1}{N}.$$

Finally, (A.54) yields

$$\mathbb{E}\left|\left|\left|\widehat{C}_N^K-C^K\right|\right|\right|_2\leq\frac{\sqrt{S_2}}{\sqrt{N}}+\frac{S_1}{N}=\mathcal{O}(N^{-1/2}),$$

uniformly in K_1, K_2, which shows

$$\left|\left|\left|\widehat{C}_N^K-C^K\right|\right|\right|_2=\mathcal{O}_{\mathbb{P}}(N^{-1/2}), \quad \text{(A.58)}$$

and the $\mathcal{O}_{\mathbb{P}}$ term is uniform in K_1, K_2. Using (A.53) and (A.58) in (A.51), we have

$$\left|\left|\left|\widehat{C}_N^K-C\right|\right|\right|_2=\mathcal{O}_{\mathbb{P}}(N^{-1/2})+L\sqrt{\frac{2}{K_1^2}+\frac{2}{K_2^2}}, \quad \text{(A.59)}$$

where the $\mathcal{O}_{\mathbb{P}}$ term is uniform in K_1, K_2.

Next, we consider the measurement scheme (S2). Observe that, under this scheme, C^K is the integral operator with kernel

$$\begin{aligned}c^K(t,s,t',s')&=\operatorname{Cov}\left\{X^K(t,s),X^K(t',s')\right\}\\&=\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2}\widetilde{c}^K(i,j,k,l)\,\mathbf{1}\{(t,s)\in I_{i,j}^K,(t',s')\in I_{k,l}^K\},\end{aligned} \quad \text{(A.60)}$$

where

$$\widetilde{c}^K(i,j,k,l)=\frac{1}{|I_{i,j}^K|\,|I_{k,l}^K|}\iint_{I_{i,j}^K\times I_{k,l}^K}c(u,v,u',v')dudvdu'dv'.$$

Now, as in (A.52), we get

$$\begin{aligned}\left\|\left\|C^K - C\right\|\right\|_2^2 &= \iiiint \left\{c^K(t,s,t',s') - c(t,s,t',s')\right\}^2 dtdsdt'ds' \\ &= \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} \iint_{I_{i,j}^K \times I_{k,l}^K} \left\{\widetilde{c}(i,j,k,l) - c(t,s,t',s')\right\}^2 dtdsdt'ds'. \qquad \text{(A.61)}\end{aligned}$$

Using (A.60) and the Lipschitz condition on c, given $(t,s) \in I_{i,j}^K, (t',s') \in I_{k,l}^K$, one has

$$\begin{aligned}&\left|\widetilde{c}(i,j,k,l) - c(t,s,t',s')\right| = \\ &\quad= \left|\frac{1}{|I_{i,j}^K|\,|I_{k,l}^K|} \iiiint_{I_{i,j}^K \times I_{k,l}^K} \left\{c(u,v,u',v') - c(t,s,t',s')\right\} dudvdu'dv'\right| \\ &\quad\le \frac{1}{|I_{i,j}^K|\,|I_{k,l}^K|} \iiiint_{I_{i,j}^K \times I_{k,l}^K} \left|c(u,v,u',v') - c(t,s,t',s')\right| dudvdu'dv' \\ &\quad\le \frac{1}{|I_{i,j}^K|\,|I_{k,l}^K|} \iiiint_{I_{i,j}^K \times I_{k,l}^K} L\sqrt{\frac{1}{K_1^2} + \frac{1}{K_2^2} + \frac{1}{K_1^2} + \frac{1}{K_2^2}}\, dudvdu'dv' \\ &\quad= L\sqrt{\frac{2}{K_1^2} + \frac{2}{K_2^2}}.\end{aligned}$$

Using this in (A.61) yields

$$\left\|\left\|C^K - C\right\|\right\|_2^2 \le 2L^2\left(\frac{1}{K_1^2} + \frac{1}{K_2^2}\right). \qquad \text{(A.62)}$$

For $\left\|\left\|\widehat{C}_K^N - C^K\right\|\right\|_2$, we proceed similarly as under the measurement scheme (S1). We need to get bounds on $\mathbb{E}\left\|\mathbf{X}^K\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \operatorname{Var}\left(\mathbf{X}^K[i,j]\right)$ and

$$\mathbb{E}\left\|\mathbf{Z}_1^K\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} \operatorname{Var}\left(\mathbf{X}_1^K[i,j]\mathbf{X}_1^K[k,l]\right).$$

Recall that under measurement scheme (S2),

$$\mathbf{X}^K[i,j] = \frac{1}{|I_{i,j}^K|}\int_{I_{i,j}^K} X(t,s)\,\mathrm{d}t\,\mathrm{d}s = \sqrt{K_1K_2}\langle X, g_{i,j}^K\rangle,$$

where $g_{i,j}^K(t,s) = \sqrt{K_1K_2}\,\mathbf{1}\{(t,s) \in I_{i,j}^K\}$. So,

$$\operatorname{Var}\left(\mathbf{X}^K[i,j]\right) = K_1K_2\operatorname{Var}\left(\langle X, g_{i,j}^K\rangle\right) = K_1K_2\langle Cg_{i,j}^K, g_{i,j}^K\rangle.$$

Observe that $\left(g_{i,j}^K\right)_{i=1,\ldots,K_1, j=1,\ldots,K_2}$ are orthonormal in $\mathcal{L}_2([0,1]^2)$ (i.e., $\langle g_{i,j}^K, g_{k,l}^K\rangle = \mathbf{1}\{(i,j) = (k,l)\}$). Thus, we can extend them to form a basis of $\mathcal{L}_2([0,1]^2)$ like in the proof of Theorem 4. We again denote this extended basis by $(g_{i,j}^K)_{i,j=1}^\infty$. Since C is positive semi-definite, $\langle C g_{i,j}^K, g_{i,j}^K\rangle \geq 0$ for every $i, j \geq 1$. Thus,

$$\begin{aligned}\sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \operatorname{Var}\left(\mathbf{X}^K[i,j]\right) &= K_1K_2\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\langle C g_{i,j}^K, g_{i,j}^K\rangle \\ &\leq K_1K_2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\langle C g_{i,j}^K, g_{i,j}^K\rangle \\ &= K_1K_2 |||C|||_1. \qquad \text{(A.63)}\end{aligned}$$

Again, $\mathbf{X}[i,j]\mathbf{X}[k,l] = K_1K_2\langle X \otimes X, g_{i,j}^K \otimes g_{k,l}^K\rangle$. This shows that

$$\operatorname{Var}\left(\mathbf{X}_1^K[i,j]\mathbf{X}_1^K[k,l]\right) = K_1^2K_2^2\langle \Gamma(g_{i,j}^K \otimes g_{k,l}^K), g_{i,j}^K \otimes g_{k,l}^K\rangle,$$

where $\Gamma = \mathbb{E}(X \otimes X \otimes X \otimes X) - C \otimes C$ is the covariance operator of $X \otimes X$. Note that the assumption $\mathbb{E}(\|X\|_2^4) < \infty$ ensures the existence of Γ, and further assures that Γ is a trace-class operator. Since the $(g_{i,j}^K)$ are orthornormal in $\mathcal{L}_2([0,1]^2)$, the $\left(g_{i,j}^K \otimes g_{k,l}^K\right)_{i,j,k,l}$ are orthonormal in $\mathcal{L}_2([0,1]^4)$. So, we can extend them to a basis $(g_{i,j}^K \otimes g_{k,l}^K)_{i,j,k,l=1}^{\infty}$ of $\mathcal{L}_2([0,1]^4)$. Since Γ is positive semi-definite

$$\begin{aligned}\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} \operatorname{Var}\left(\mathbf{X}_1^K[i,j]\mathbf{X}_1^K[k,l]\right) &= K_1^2K_2^2\sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2}\langle \Gamma(g_{i,j}^K \otimes g_{k,l}^K), g_{i,j}^K \otimes g_{k,l}^K\rangle \\ &\leq K_1^2K_2^2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\langle \Gamma(g_{i,j}^K \otimes g_{k,l}^K), g_{i,j}^K \otimes g_{k,l}^K\rangle \\ &= K_1^2K_2^2 |||\Gamma|||_1. \qquad \text{(A.64)}\end{aligned}$$

Using (A.63) and (A.64), and proceeding as in the case of (S1), we get

$$\mathbb{E}\left|\left|\left|\widehat{C}_N^K - C^K\right|\right|\right|_2 \leq \frac{\sqrt{|||\Gamma|||_1}}{\sqrt{N}} + \frac{|||C|||_1}{N},$$

that is,

$$\left|\left|\left|\widehat{C}_N^K - C^K\right|\right|\right|_2 = \mathcal{O}_{\mathbb{P}}(N^{-1/2}), \qquad \text{(A.65)}$$

where the $\mathcal{O}_{\mathbb{P}}$ term is uniform in K_1, K_2. Finally, using (A.62) and (A.65) in (A.51), we get

$$\left|\left|\left|\widehat{C}_N^K - C\right|\right|\right|_2 = \mathcal{O}_{\mathbb{P}}(N^{-1/2}) + L\sqrt{\frac{2}{K_1^2} + \frac{2}{K_2^2}}, \qquad \text{(A.66)}$$

where the $\mathcal{O}_{\mathbb{P}}$ term is uniform in K_1, K_2. Observe that we obtain the same rate under both (S1) and (S2). Also observe that, under both the schemes,

$$\left|\left|\left|\widehat{C}_N^K\right|\right|\right|_2 = |||C|||_2 + \mathcal{O}_{\mathbb{P}}(N^{-1/2}) + L\sqrt{\frac{2}{K_1^2} + \frac{2}{K_2^2}}.$$

Using these in (A.50), we get

$$\begin{aligned}\left|\left|\left|\widehat{C}^K_{R,N} - C_R\right|\right|\right|_2 &\le \left\{4\sqrt{2}\left(|||C|||_2 + |||C|||_2 + \mathcal{O}_{\mathbb{P}}(N^{-1/2}) + L\sqrt{\frac{2}{K_1^2}+\frac{2}{K_2^2}}\right)\sum_{r=1}^{R}\frac{\sigma_r}{\alpha_r} + 1\right\} \\ &\qquad \times \left\{\mathcal{O}_{\mathbb{P}}(N^{-1/2}) + L\sqrt{\frac{2}{K_1^2}+\frac{2}{K_2^2}}\right\} \\ &= \mathcal{O}_{\mathbb{P}}\left(\frac{a_R}{\sqrt{N}}\right) + \left(16a_R + \sqrt{2}\right)L\sqrt{\frac{1}{K_1^2}+\frac{1}{K_2^2}} + \frac{8\sqrt{2}L^2}{|||C|||_2}\left(\frac{1}{K_1^2}+\frac{1}{K_2^2}\right)a_R,\end{aligned}$$

where the $\mathcal{O}_{\mathbb{P}}$ term is uniform in K_1, K_2. □

Proof of Theorem 9

Finally, we consider the case where the surfaces are observed at irregular (and possibly different number of) locations.

We start by using the inequality

$$\left|\left|\left|\widetilde{C}_{R,N} - C\right|\right|\right|_2 \le \left|\left|\left|\widetilde{C}_{R,N} - C_R\right|\right|\right|_2 + |||C_R - C|||_2. \tag{A.67}$$

For the second part on the right-hand side, using (A.46), we get that

$$|||C_R - C|||_2 = \Big(\sum_{r=1}^{R}\sigma_r^2\Big)^{1/2}.$$

For the first part, we use Lemma 9 to get

$$\left|\left|\left|\widetilde{C}_{R,N} - C_R\right|\right|\right|_2 \le \left\{4\sqrt{2}\Big(|||C|||_2 + \left|\left|\left|\widetilde{C}_N\right|\right|\right|_2\Big)\sum_{r=1}^{R}\frac{\sigma_r}{\alpha_r} + 1\right\}\left|\left|\left|\widetilde{C}_N - C\right|\right|\right|_2. \tag{A.68}$$

Now,

$$\left|\left|\left|\widetilde{C}_N - C\right|\right|\right|_2 \le \left|\left|\left|\widetilde{C}_N - \widehat{C}_N\right|\right|\right|_2 + \left|\left|\left|\widehat{C}_N - C\right|\right|\right|_2.$$

The first part on the right is $\mathcal{O}_{\mathbb{P}}(b_N)$ by the assumption of the theorem, while the second part is $\mathcal{O}_{\mathbb{P}}(N^{-1/2})$ since $\mathbb{E}(\|X\|^4) < \infty$. Hence, we get $\left|\left|\left|\widetilde{C}_N - C\right|\right|\right|_2 \le \mathcal{O}_{\mathbb{P}}(b_N)$, which also shows

$$\left|\left|\left|\widetilde{C}_N\right|\right|\right|_2 \le |||C|||_2 + \left|\left|\left|\widetilde{C}_N - C\right|\right|\right|_2 \le |||C|||_2 + \mathcal{O}_{\mathbb{P}}(b_N).$$

Using these in (A.68), we get $\left|\left|\left|\widetilde{C}_{R,N} - C_R\right|\right|\right|_2 = \mathcal{O}_{\mathbb{P}}(a_R b_N)$ where $a_R = |||C|||_2 \sum_{r=1}^{R}\sigma_r/\alpha_r$. The result now follows by substituting this in (A.67).

Separability under Sparse Measurements

Explicit Formula for Local Polynomial Regression

Our estimators introduced in Section 4.3 are based on local polynomial regression techniques and are defined as minimizers of (weighted) least squares problems (4.5). It turns out that the minimizers for these point-wise optimization problems admit a unique solution given by an explicit formula. In this section we recall this formula for a general local linear polynomial regression with possibly exogenous weights which will be later used in the proofs of the asymptotic behaviour of our estimators.

The local linear surface smoother of the generic set $\{(x_k, y_k, z_k) \mid k = 1, \dots, M\} \subset \mathbb{R}^3$ given weights $\{w_k \mid k = 1, \dots, M\}$ is defined as the solution of the least squares problem (4.5). It turns out that this minimizer to this optimization problem admits a unique solution:

$$\widehat{\gamma}_{0(x,y)} = \Psi_1(x, y)\,[\Psi_2(x, y)]^{-1}\,, \qquad (x, y) \in [0, 1]^2, \tag{A.69}$$

where for each $(x, y) \in [0, 1]^2$ and $p, q \in \mathbb{N}_0$ we define

$$\begin{aligned}
\Phi_1(x, y) &= S_{20}(x, y)S_{02}(x, y) - [S_{11}(x, y)]^2\,,\\
\Phi_2(x, y) &= S_{10}(x, y)S_{02}(x, y) - S_{01}(x, y)S_{11}(x, y),\\
\Phi_3(x, y) &= S_{01}(x, y)S_{20}(x, y) - S_{10}(x, y)S_{11}(x, y),\\
\Psi_1(x, y) &= \Phi_1(x, y)Q_{00}(x, y) - \Phi_2(x, y)Q_{10}(x, y) - \Phi_3(x, y)Q_{01}(x, y),\\
\Psi_2(x, y) &= \Phi_1(x, y)S_{00}(x, y) - \Phi_2(x, y)S_{10}(x, y) - \Phi_3(x, y)S_{01}(x, y),
\end{aligned}$$

and

$$S_{pq}(x, y) = \frac{1}{M}\sum_{k=1}^{M}\left(\frac{x - x_k}{h_1}\right)^p\left(\frac{y - y_k}{h_2}\right)^q\frac{1}{h_1h_2}\mathcal{K}\left(\frac{x - x_k}{h_1}\right)\mathcal{K}\left(\frac{y - y_k}{h_2}\right)w_m \tag{A.70}$$

for $0 \le p + q \le 2$, and

$$Q_{pq}(x, y) = \frac{1}{M}\sum_{k=1}^{M}\left(\frac{x - x_k}{h_1}\right)^p\left(\frac{y - y_k}{h_2}\right)^q\frac{1}{h_1h_2}\mathcal{K}\left(\frac{x - x_k}{h_1}\right)\mathcal{K}\left(\frac{y - y_k}{h_2}\right)w_kz_k \tag{A.71}$$

for $0 \le p + q \le 1$. In the above, $h_1 > 0$ and $h_2 > 0$ are smoothing bandwidths in the first and the second dimension respectively.

The formula (A.69) is derived by differentiating the weighted least squares (4.5) and finding the solution to the normal equations. It is based on the standard steps used in the local regression literature, see e.g. Fan and Gijbels (1996)[§3.1] or Rubín and Panaretos (2020)[§B.2].

Kernel Averages of m-dependent Data

Thanks the explicit formula (A.69) we may reduce the asymptotic behaviour assessment to the investigation of the terms (A.70) and (A.71). In this section we review the general asymptotic framework for the asymptotics of these kernel averages and hence the framework for the asymptotics of (A.69). We shall use the general theory developed by Hansen (2008) who derived a toolbox for strong mixing time series data where the regressors attain values in possibly unbounded sets. Here we recall this result and write down a simplified version sufficient for our data.

Theorem 12 (Hansen, 2008)**.** *Let $\{(U_k, V_k, Z_k)\}_{k\in\mathbb{Z}} \in \mathbb{R}^3$ be a strictly stationary sequence of random vectors and consider the averages of the form*

$$\Xi_k(u,v) = \frac{1}{kh_1h_2}\sum_{i=1}^{k}\left(\frac{u-U_i}{h_1}\right)^p\left(\frac{v-V_i}{h_2}\right)^q\mathcal{K}\left(\frac{u-U_i}{h_1}\right)\mathcal{K}\left(\frac{v-V_i}{h_2}\right)Z_i \qquad \text{(A.72)}$$

where $p, q \in \mathbb{N}_0$.

- *(C1) $\mathcal{K}(\cdot)$ is the Epanechnikov kernel, i.e. $\mathcal{K}(u) = (3/4)(1-u^2)\mathbb{1}_{[|u|<1]}$.*
- *(C2) (U_k, V_k) attain values in the set $[0,1]^2$.*
- *(C3) $\{(U_k, V_k, Z_k)\}$ is an m-dependent sequence, i.e. for each $k \in \mathbb{Z}$, the random vectors $(\dots, U_k, V_k, Z_k)$ and $(U_{k+m}, V_{k+m}, Z_{k+m}, U_{k+m+1}, V_{k+m+1}, Z_{k+m+1}, \dots)$ are independent.*
- *(C4) There exists $s > 2$ such that $(u,v) \mapsto \mathbb{E}[|Z_1|^s | U_1 = u, V_1 = v]$ is bounded.*
- *(C5) The probability density function of (U_1, V_1) is twice continuously differentiable.*
- *(C6) The smoothing bandwidth satisfies $(\log k)/(kh_1h_2) = o(1)$ as $k \to \infty$.*

Then the kernel averages (A.72) *satisfy*

$$\sup_{(u,v)\in[0,1]^2}|\Xi_k(u,v) - \mathbb{E}\Xi_k(u,v)| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log k}{kh_1h_2}}\right), \qquad as \quad k\to\infty. \qquad \text{(A.73)}$$

Theorem 12 is essentially a special case of Hansen (2008, Thm 2), where the considered sequence is defined on a bounded domain and is m-dependent. We also assume that the two dimensional smoothing kernel (A.72) is the product of two one dimensional Epanechnikov kernels. Note that the function $(u,v) \mapsto u^p v^q K(u)K(v)$ satisfies Hansen's conditions on the smoothing kernel.

The only generalisation where Theorem 12 deviates from Hansen (2008, Thm 2) is that we allow the smoothing bandwidth (h_1, h_2) to attain different values in different directions.

The proof of such generalisation, while having the smoothing kernel as a product of one-dimensional kernels, follows the lines of the proof Hansen (2008, Thm 2) where h^d, with $d = 2$, is replaced by $h_1 h_2$.

The following corollary goes one step further and incorporates the convergence of $\mathbb{E}\Xi_k(u, v)$ into the statement (A.73).

Corollary 5. *Under assumption of Theorem 12 consider the function*

$$M(u,v) = c_p c_q \mathbb{E}\left[Z_1 | U_1 = u, V_1 = v\right] f_{U_1,V_1}(u,v), \qquad x, y \in [0,1], \tag{A.74}$$

where $f_{U_1,V_1}(\cdot,\cdot)$ denotes the probability density function of (U_1, V_1) and $c_r = \int x^r \mathcal{K}(x)\,\mathrm{d}x$ for $r \in \mathbb{N}_0$. Moreover:

(D1) Assume that the function $M(\cdot,\cdot)$ is twice continuously differentiable on $[0,1]^2$.

Then the kernel averages (A.72) *satisfy*

$$\sup_{(u,v)\in[0,1]^2} \left|\Xi_k(u,v) - M(u,v)\right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log k}{k h_1 h_2}} + h_1^2 + h_2^2\right), \qquad \text{as} \quad k \to \infty. \tag{A.75}$$

Proof. We start by decomposing the supremum (A.75) into a stochastic and a deterministic part

$$\begin{aligned} \sup_{(u,v)\in[0,1]^2} \left|\Xi_k(u,v) - M(u,v)\right| \leq & \sup_{(u,v)\in[0,1]^2} \left|\Xi_k(u,v) - \mathbb{E}\Xi_k(u,v)\right| \\ & + \sup_{(u,v)\in[0,1]^2} \left|\mathbb{E}\Xi_k(u,v) - M(u,v)\right|. \end{aligned} \tag{A.76}$$

The first term on the right-hand side of (A.76) is of order $O_{\mathbb{P}}(\sqrt{(\log n)/(n h_1 h_2)})$ by Theorem 12. The expectation in the second term on the right-hand side of (A.76) is developed as

$$\begin{aligned} \mathbb{E}\Xi_k(u,v) &= \mathbb{E}\left[\left(\frac{u-U_1}{h_1}\right)^p \left(\frac{v-V_1}{h_2}\right)^q \frac{1}{h_1 h_2} \mathcal{K}\left(\frac{u-U_1}{h_1}\right) \mathcal{K}\left(\frac{v-V_1}{h_2}\right) \mathbb{E}\left[Z_{11} | U_1, V_1\right]\right] \\ &= \mathbb{E}\left[\left(\frac{u-U_1}{h_1}\right)^p \left(\frac{v-V_1}{h_2}\right)^q \frac{1}{h_1 h_2} \mathcal{K}\left(\frac{u-U_1}{h_1}\right) \mathcal{K}\left(\frac{v-V_1}{h_2}\right) M(U_1, V_1)\right] \\ &= \iint \left(\frac{u-x}{h_1}\right)^p \left(\frac{v-y}{h_2}\right)^q \frac{1}{h_1 h_2} \mathcal{K}\left(\frac{u-x}{h_1}\right) \mathcal{K}\left(\frac{v-y}{h_2}\right) M(x,y) f_{U_1,V_1}(x,y) dx dy \\ &= \iint (\tilde{x})^p (\tilde{y})^q \mathcal{K}(\tilde{x}) \mathcal{K}(\tilde{y}) M(u + h_1\tilde{x}, v + h_2\tilde{y}) f_{U_1,V_1}(u + h_1\tilde{x}, v + h_2\tilde{y}) d\tilde{x} d\tilde{y}. \end{aligned} \tag{A.77}$$

Applying the Taylor expansion of order 2 in the right-hand side of (A.77) and using assumptions (C5) and (D1), the second term on the right-hand side of (A.76) is of order $O(h_1^2 + h_2^2)$. ☐

Proof of Proposition 12

Tailoring the generic smoother (4.5) to the mean surface estimator (4.6), we arrive at the customized versions of (A.70) and (A.71):

$$S_{pq}^{\mu}(t,s) = \frac{1}{\sum_{n=1}^{N} M_n} \sum_{n=1}^{N} \sum_{m=1}^{M_n} \left(\frac{t-t_{nm}}{h_{\mu,1}}\right)^p \left(\frac{s-s_{nm}}{h_{\mu,2}}\right)^q \cdot \tag{A.78}$$

$$\cdot \frac{1}{h_{\mu,1}h_{\mu,2}} \mathcal{K}\left(\frac{t-t_{nm}}{h_{\mu,1}}\right) \mathcal{K}\left(\frac{s-s_{nm}}{h_{\mu,2}}\right), \quad 0 \le p+q \le 2. \tag{A.79}$$

$$Q_{pq}^{\mu}(t,s) = \frac{1}{\sum_{n=1}^{N} M_n} \sum_{n=1}^{N} \sum_{m=1}^{M_n} \left(\frac{t-t_{nm}}{h_{\mu,1}}\right)^p \left(\frac{s-s_{nm}}{h_{\mu,2}}\right)^q \cdot \tag{A.80}$$

$$\cdot \frac{1}{h_{\mu,1}h_{\mu,2}} \mathcal{K}\left(\frac{t-t_{nm}}{h_{\mu,1}}\right) \mathcal{K}\left(\frac{s-s_{nm}}{h_{\mu,2}}\right) Y_{nm}, \quad 0 \le p+q \le 1. \tag{A.81}$$

We assess the asymptotic behaviour of (A.78) and (A.80) in the following lemmas.

Lemma 11. *Under assumptions (B1) – (B6),*

$$\sup_{(t,s)\in[0,1]^2} \left| Q_{pq}^{\mu}(t,s) - M_{[Q_{pq}^{\mu}]}(t,s) \right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N h_1 h_2}} + h_{\mu,1}^2 + h_{\mu,2}^2 \right), \qquad \text{as} \quad N \to \infty, \tag{A.82}$$

for each $0 \le p+q \le 1$ *and where*

$$M_{[Q_{00}^{\mu}]}(t,s) = \mu(t,s) f_{(t,s)}(t,s), \qquad M_{[Q_{10}^{\mu}]}(t,s) = M_{[Q_{01}^{\mu}]}(t,s) = 0, \qquad t,s \in [0,1]. \tag{A.83}$$

Proof. Define the sequence of random vectors $\{(U_k, V_k, Z_k)\}_{k=1}^{\infty}$ by putting $\{t_{nm}\}$, $\{s_{nm}\}$ and $\{Y_{nm}\}$ in order such that

$$\begin{aligned} \{U_1, U_2, \ldots\} &= \{t_{11}, t_{12}, \ldots, t_{1m_1}, t_{21}, \ldots, t_{2m_2}, t_{31}, \ldots\}, \\ \{V_1, V_2, \ldots\} &= \{s_{11}, s_{12}, \ldots, s_{1m_1}, s_{21}, \ldots, s_{2m_2}, s_{31}, \ldots\}, \\ \{Z_1, Z_2, \ldots\} &= \{Y_{11}, Y_{12}, \ldots, Y_{1m_1}, Y_{21}, \ldots, Y_{2m_2}, Y_{31}, \ldots\}. \end{aligned} \tag{A.84}$$

The sequence$\{(U_k, V_k, Z_k)\}_{k=1}^{\infty}$ satisfies the assumption of Theorem 12 and Corollary 5, namely strict stationarity is by assumptions (B1) – (B3), is M^{max}-dependent by assumption (B1), and the conditions (C4), (C5), (C6), (D1) are satisfied by assumptions (B5), (B2), (B6), (B5) respectively. Therefore the sequence of kernel averages

$$\Xi_{pq,k}^{\mu}(t,s) = \frac{1}{k h_{\mu,1} h_{\mu,2}} \sum_{i=1}^{n} \left(\frac{t-U_i}{h_1}\right)^p \left(\frac{s-V_i}{h_2}\right)^q \mathcal{K}\left(\frac{t-U_i}{h_1}\right) \mathcal{K}\left(\frac{s-V_i}{h_2}\right) Z_i$$

satisfies

$$\sup_{(t,s)\in[0,1]^2} \left|\Xi^{\mu}_{pq,k}(t,s) - M_{[Q^{\mu}_{pq}]}(t,s)\right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log k}{kh_1h_2}} + h^2_{\mu,1} + h^2_{\mu,2}\right), \qquad \text{as} \quad k\to\infty,$$

and the formulae (A.83) follow from the definition of $\{(U_k, V_k, Z_k)\}_{k=1}^{\infty}$ and definition (A.74). Since the sequence $\{Q^{\mu}_{pq}(t,s)\}_{N=1}^{\infty}$ is a subsequence of $\{\Xi^{\mu}_{pq,k}(t,s)\}_{k=1}^{\infty}$ and $k = k(N) \asymp N$ as $N\to\infty$, the convergence rate (A.82) holds as well. ☐

Lemma 12. *Under assumptions (B1) – (B3),*

$$\sup_{(t,s)\in[0,1]^2} \left|S^{\mu}_{pq}(t,s) - M_{[S^{\mu}_{pq}]}(t,s)\right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh_1h_2}} + h^2_{\mu,1} + h^2_{\mu,2}\right), \qquad \textit{as} \quad N\to\infty, \tag{A.85}$$

for each $0 \le p+q \le 1$ *and where*

$$M_{[S^{\mu}_{00}]}(t,s) = \mu(t,s)f_{(t,s)}(t,s), \qquad M_{[S^{\mu}_{11}]}(t,s) = M_{[S^{\mu}_{10}]}(t,s) = M_{[S^{\mu}_{01}]}(t,s) = 0, \tag{A.86}$$

$$M_{[S^{\mu}_{20}]}(t,s) = M_{[S^{\mu}_{02}]}(t,s) = c_2 f_{(t,s)}(t,s), \qquad c_2 = \int x^2\mathcal{K}(x)\,\mathrm{d}x. \qquad t,s\in[0,1]. \tag{A.87}$$

Proof. The proof of this lemma follows essentially the same lines as the proof of Lemma 11. In the definition of the sequence $\{(U_k, V_k, Z_k)\}_{k=1}^{\infty}$ we put $Z_k = 1$, for all $k\in\mathbb{N}$, on the line (A.84). The formulae (A.86) and (A.87) are verified analogously by the definition (A.74). ☐

Proof of Proposition 12. We are now ready to combine the above and prove Proposition 12. Following the explicit formulae for local linear smoothers presented above, we have for

$$\begin{aligned}
\Phi^{\mu}_1(t,s) &= S^{\mu}_{20}(t,s)S^{\mu}_{02}(t,s) - \left[S^{\mu}_{11}(t,s)\right]^2,\\
\Phi^{\mu}_2(t,s) &= S^{\mu}_{10}(t,s)S^{\mu}_{02}(t,s) - S^{\mu}_{01}(t,s)S^{\mu}_{11}(t,s),\\
\Phi^{\mu}_3(t,s) &= S^{\mu}_{01}(t,s)S^{\mu}_{20}(t,s) - S^{\mu}_{10}(t,s)S^{\mu}_{11}(t,s),\\
\Psi^{\mu}_1(t,s) &= \Phi_1(t,s)Q^{\mu}_{00}(t,s) - \Phi_2(t,s)Q^{\mu}_{10}(t,s) - \Phi_3(t,s)Q^{\mu}_{01}(t,s),\\
\Psi^{\mu}_2(t,s) &= \Phi_1(t,s)S^{\mu}_{00}(t,s) - \Phi_2(t,s)S^{\mu}_{10}(t,s) - \Phi_3(t,s)S^{\mu}_{01}(t,s),
\end{aligned}$$

their asymptotic behaviour

$$\begin{aligned}
\Phi_1^\mu(t,s) &= \left(c_2 f_{(t,s)}(t,s)\right)^2 + \mathcal{O}_{\mathbb{P}}\left(r_N^\mu\right),\\
\Phi_2^\mu(t,s) &= \mathcal{O}_{\mathbb{P}}\left(r_N^\mu\right),\\
\Phi_3^\mu(t,s) &= \mathcal{O}_{\mathbb{P}}\left(r_N^\mu\right),\\
\Psi_1^\mu(t,s) &= (c_2)^2\left(f_{(t,s)}(t,s)\right)^3\mu(t,s) + \mathcal{O}_{\mathbb{P}}\left(r_N^\mu\right),\\
\Psi_2^\mu(t,s) &= (c_2)^2\left(f_{(t,s)}(t,s)\right)^3 + \mathcal{O}_{\mathbb{P}}\left(r_N^\mu\right),
\end{aligned}$$

uniformly in $(t,s)\in[0,1]^2$, as $N\to\infty$, where $r_N^\mu = \sqrt{(\log N)/(N h_{\mu,1} h_{\mu,2})} + h_{\mu,1}^2 + h_{\mu,2}^2$. Hence

$$\widehat{\mu}(t,s) = \Psi_1^\mu(t,s)/\left[\Psi_2^\mu(t,s)\right]^{-1} = \mu(t,s) + \mathcal{O}_{\mathbb{P}}\left(r_N^\mu\right), \qquad \text{as} \quad N\to\infty,$$

by the uniform version of Slutsky's theorem and by the fact that $f_{(t,s)}(\cdot,\cdot)\neq 0$ on $[0,1]^2$ by (B2). Thanks to assumption (B6), the convergence rate simplifies to the common rate h and

$$\widehat{\mu}(t,s) = \mu(t,s) + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N h^2}} + h^2\right), \qquad \text{as} \quad N\to\infty. \qquad \square$$

Proof of Theorem 10 and Corollary 4

Lemma 13. *Assume the conditions (A1), (B1) – (B9) and fix a deterministic twice continuously differentiable kernel $\beta(s,s')$, $s,s'\in[0,1]$ such that $\iint[\beta(s,s')]^2 ds ds' > 0$. Then the smoother $\widehat{\alpha}(t,t')$, $t,t'\in[0,1]$, obtained by smoothing the set*

$$\left\{\left(t_{nm}, t_{nm'}, \frac{G_{nmm'}}{\beta(s_{nm},s_{nm'})}\right) \,\middle|\, m,m'=1,\dots,M_n,\ m\neq m',\ n=1\dots,N\right\} \tag{A.88}$$

using weights $\{\beta^2(s_{nm},s_{nm'})\}$ admits the following asymptotics

$$\widehat{\alpha}(t,t') = a(t,t')\frac{\iint \beta(s,s')b(s,s')f_s(s)f_s(s')dsds'}{\iint\left[\beta(s,s')\right]^2 f_s(s)f_s(s')dsds'} + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N h^2}} + h^2\right) \tag{A.89}$$

uniformly in $(t,t')\in[0,1]^2$ as $N\to\infty$. We recall that $a(t,t')$, $t,t'\in[0,1]$, and $b(s,s')$, $s,s'\in[0,1]$, on the right-hand side of (A.89) *are the true separable covariance structure components* (4.1)*.*

Proof. We start the proof by the analysis of a simplified case. Suppose that we know the mean surface $\mu(\cdot,\cdot)$ and define the raw covariances accordingly

$$\tilde{G}_{nmm'} = (Y_{nm} - \mu(t_{nm}, s_{nm}))(Y_{nm'} - \mu(t_{nm'}, s_{nm'})). \tag{A.90}$$

Construct the smoother of the set (A.88) where we replace $G_{nmm'}$ by $\tilde{G}_{nmm'}$. Such smoother, denoted as $\tilde{\alpha}(\cdot,\cdot)$, is given by the formula

$$\begin{aligned}
\tilde{\alpha}(t,t') &= \Psi_1^{\tilde{\alpha}}(t,t') / \left[\Psi_2^{\tilde{\alpha}}(t,t')\right]^{-1}, \qquad\qquad (A.91)\\
\Psi_1^{\tilde{\alpha}}(t,t') &= \Phi_1(t,t')Q_{00}^{\tilde{\alpha}}(t,t') - \Phi_2(t,t')Q_{10}^{\tilde{\alpha}}(t,t') - \Phi_3(t,t')Q_{01}^{\tilde{\alpha}}(t,t'),\\
\Psi_2^{\tilde{\alpha}}(t,t') &= \Phi_1(t,t')S_{00}^{\tilde{\alpha}}(t,t') - \Phi_2(t,t')S_{10}^{\tilde{\alpha}}(t,t') - \Phi_3(t,t')S_{01}^{\tilde{\alpha}}(t,t'),\\
\Phi_1^{\tilde{\alpha}}(t,t') &= S_{20}^{\tilde{\alpha}}(t,t')S_{02}^{\tilde{\alpha}}(t,t') - \left[S_{11}^{\tilde{\alpha}}(t,t')\right]^2,\\
\Phi_2^{\tilde{\alpha}}(t,t') &= S_{10}^{\tilde{\alpha}}(t,t')S_{02}^{\tilde{\alpha}}(t,t') - S_{01}^{\tilde{\alpha}}(t,t')S_{11}^{\tilde{\alpha}}(t,t'),\\
\Phi_3^{\tilde{\alpha}}(t,t') &= S_{01}^{\tilde{\alpha}}(t,t')S_{20}^{\tilde{\alpha}}(t,t') - S_{10}^{\tilde{\alpha}}(t,t')S_{11}^{\tilde{\alpha}}(t,t'),\\
S_{pq}^{\tilde{\alpha}}(t,t') &= \frac{1}{\sum_{n=1}^{N} M_n(M_n-1)} \sum_{n=1}^{N} \sum_{\substack{m,m'=1\\ m\neq m'}}^{M_n} \left(\frac{t-t_{nm}}{h_a}\right)^p \left(\frac{t'-t_{nm'}}{h_a}\right)^q \cdot\\
&\quad \cdot \frac{1}{h_a^2}\mathcal{K}\left(\frac{t-t_{nm}}{h_a}\right)\mathcal{K}\left(\frac{t'-t_{nm'}}{h_a}\right)\left[\beta(s_{nm}, s_{nm'})\right]^2, \qquad 0 \leq p+q \leq 2, \qquad (A.92)\\
Q_{pq}^{\tilde{\alpha}}(t,t') &= \frac{1}{\sum_{n=1}^{N} M_n(M_n-1)} \sum_{n=1}^{N} \sum_{\substack{m,m'=1\\ m\neq m'}}^{M_n} \left(\frac{t-t_{nm}}{h_a}\right)^p \left(\frac{t'-t_{nm'}}{h_a}\right)^q \cdot\\
&\quad \cdot \frac{1}{h_a^2}\mathcal{K}\left(\frac{t-t_{nm}}{h_a}\right)\mathcal{K}\left(\frac{t'-t_{nm'}}{h_a}\right)\beta(s_{nm}, s_{nm'})\tilde{G}_{nmm'}, \qquad 0 \leq p+q \leq 1. \qquad (A.93)
\end{aligned}$$

The asymptotic behaviour of $Q_{pq}^{\tilde{\alpha}}$ and $S_{pq}^{\tilde{\alpha}}(t,t')$ is assessed similarly as the surface smoother in Lemmas 11 and 12. We proceed again with defining the $[M^{max}]^2$-dependent sequences $\{(U_k, V_k, Z_k)\}_{k=1}^{\infty}$ by putting the pairs $(t_{nm}, t_{nm'})$, $m, m' = 1, \ldots, M_n$, $m \neq m'$, $n = 1, \ldots, N$ into the sequence $\{(U_k, V_k)\}_{k=1}^{\infty}$ such that we set $U_k = t_{nm}$ and $V_k = t_{nm'}$ while starting from the data from the first surface ($n = 1$), then proceeding with $n = 2$ etc.

For the asymptotics of $Q_{pq}^{\tilde{\alpha}}$, define

$$Z_k^Q = \beta(s_{nm}, s_{nm'})\tilde{G}_{nmm'} \tag{A.94}$$

where $s_{nm}, s_{nm'}, \tilde{G}_{nmm'}$ correspond to that sparse observation which was assigned to (U_k, V_k). We use Theorem 12 and Corollary 5 thanks to assumptions (B7) – (B9).

Moreover, we verify

$$\begin{aligned}
\mathbb{E}\left[Z_k^Q \middle| U_k = t, V_k = t'\right] = \\
&= \mathbb{E}\left[\beta(s_{nm}, s_{nm'})\left(X_n(t, s_{nm}) + \varepsilon_{nm}\right)\left(X_n(t, s_{nm'}) + \varepsilon_{nm'}\right)\right)\middle| U_k = t, V_k = t'] \\
&= \mathbb{E}\left[\beta(s_{nm}, s_{nm'})a(t,t')b(s_{nm}, s_{nm'})\right] \\
&= a(t,t')\iint \beta(s,s')b(s,s')f_s(s)f_s(s')dsds',
\end{aligned}$$

where $f_s(s) = \int f_{(t,s)}(t,s)\,\mathrm{d}t$ is the marginal density of the random position s_{11}. Likewise, denote $f_t(t) = \int f_{(t,s)}(t,s)\,\mathrm{d}s$ is the marginal density of the random position t_{11}. Then

$$Q_{00}^{\tilde{\alpha}}(t,t') = a(t,t')f_t(t)f_t(t')\iint \beta(s,s')b(s,s')f_s(s)f_s(s')dsds' + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh_a^2}} + h_a^2\right),$$

uniformly in $(t,t') \in [0,1]^2$ as $N \to \infty$ and where $c_r = \int x^r \mathcal{K}(x)\,\mathrm{d}x$, $r \in \mathbb{N}$.

Similarly to the analysis above we assess the asymptotics of $S_{pq}^{\tilde{\alpha}}$. Instead of the definition in (A.94) we set here $Z_k^S = [\beta(s_{nm}, s_{nm'})]^2$ and calculate

$$\mathbb{E}\left[Z_k^S \middle| U_k = t, V_k = t'\right] = \iint [\beta(s,s')]^2 f_s(s)f_s(s')dsds'.$$

Hence

$$S_{pq}^{\tilde{\alpha}}(t,t') = c_p c_q f_t(t)f_t(t')\iint [\beta(s,s')]^2 f_s(s)f_s(s')dsds' + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh_a^2}} + h_a^2\right)$$

uniformly in $(t,t') \in [0,1]^2$ as $N \to \infty$. By the assumptions on the kernel $\beta(\cdot,\cdot)$, the uniform Slutsky theorem, the formula (A.91), and the fact that $h_a \asymp h$ as in assumption (B9):

$$\tilde{\alpha}(t,t') = a(t,t')\frac{\iint \beta(s,s')b(s,s')f_s(s)f_s(s')dsds'}{\iint \left[\beta(s,s')\right]^2 f_s(s)f_s(s')dsds'} + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right)$$

uniformly in $(t,t') \in [0,1]^2$ as $N \to \infty$.

It remains to comment on the difference $\tilde{\alpha}(t,t')$ and $\widehat{\alpha}(t,t')$, i.e. when the empirical mean $\widehat{\mu}(\cdot,\cdot)$ is supplied into the raw covariances $G_{nmm'}$. Since

$$\begin{aligned}
G_{nmm'} = \tilde{G}_{nmm'} \\
&+ \left(\mu(t_{nm}, s_{nm}) - \widehat{\mu}(t_{nm}, s_{nm})\right)\left(Y_{nm'} - \widehat{\mu}(t_{nm'}, s_{nm'})\right) \\
&+ \left(\mu(t_{nm'}, s_{nm'}) - \widehat{\mu}(t_{nm'}, s_{nm'})\right)\left(Y_{nm} - \widehat{\mu}(t_{nm}, s_{nm})\right) \\
&+ \left(\mu(t_{nm}, s_{nm}) - \widehat{\mu}(t_{nm}, s_{nm})\right)\left(\mu(t_{nm'}, s_{nm'}) - \widehat{\mu}(t_{nm'}, s_{nm'})\right)
\end{aligned}$$

we conclude by Proposition 12 that

$$G_{nmm'} = \tilde{G}_{nmm'} + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right) \tag{A.95}$$

uniformly across all n, m, m' as $N \to \infty$. Therefore the claim (A.89) follows. $\square$

Corollary 6. *Assume the conditions (A1), (B1) – (B9) and consider a random kernel* $\widehat{\beta}(\cdot,\cdot)$ *such that*

$$\widehat{\beta}(s,s') = \beta(s,s') + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right) \tag{A.96}$$

uniformly in $(s,s') \in [0,1]^2$ *as* $N \to \infty$*, where* $\beta(s,s')$*,* $s,s' \in [0,1]$ *is a deterministic twice continuously differentiable kernel such that* $\iint [\beta(s,s')]^2 ds ds' > 0$*. Then the smoother* $\widehat{\alpha}(t,t')$*,* $t,t' \in [0,1]$*, obtained by smoothing the set*

$$\left\{\left(t_{nm}, t_{nm'}, \frac{G_{nmm'}}{\widehat{\beta}(s_{nm}, s_{nm'})}\right) \;\middle|\; m,m' = 1,\dots,M_n,\, m \neq m',\; n = 1\dots,N\right\}$$

using weights $\{\widehat{\beta}^2(s_{nm}, s_{nm'})\}$ *admits the same asymptotics as in the previous lemma:*

$$\widehat{\alpha}(t,t') = a(t,t')\frac{\iint \beta(s,s')b(s,s')f_s(s)f_s(s')dsds'}{\iint \left[\beta(s,s')\right]^2 f_s(s)f_s(s')dsds'} + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right)$$

uniformly in $(t,t') \in [0,1]^2$ *as* $N \to \infty$*.*

Proof. The proof of this corollary follows the same lines as the proof of Lemma 13. We define

$$\begin{aligned}
S^{\widehat{\alpha}}_{pq}(t,t') = {} & \frac{1}{\sum_{n=1}^N M_n(M_n-1)} \sum_{n=1}^N \sum_{\substack{m,m'=1\\ m\neq m'}}^{M_n} \left(\frac{t-t_{nm}}{h_a}\right)^p \left(\frac{t'-t_{nm'}}{h_a}\right)^q \cdot \\
& \cdot \frac{1}{h_a^2}\mathcal{K}\left(\frac{t-t_{nm}}{h_a}\right)\mathcal{K}\left(\frac{t'-t_{nm'}}{h_a}\right)\left[\widehat{\beta}(s_{nm}, s_{nm'})\right]^2, \qquad 0 \le p+q \le 2, \\
Q^{\widehat{\alpha}}_{pq}(t,t') = {} & \frac{1}{\sum_{n=1}^N M_n(M_n-1)} \sum_{n=1}^N \sum_{\substack{m,m'=1\\ m\neq m'}}^{M_n} \left(\frac{t-t_{nm}}{h_a}\right)^p \left(\frac{t'-t_{nm'}}{h_a}\right)^q \cdot \\
& \cdot \frac{1}{h_a^2}\mathcal{K}\left(\frac{t-t_{nm}}{h_a}\right)\mathcal{K}\left(\frac{t'-t_{nm'}}{h_a}\right)\widehat{\beta}(s_{nm}, s_{nm'})\tilde{G}_{nmm'}, \qquad 0 \le p+q \le 1,
\end{aligned}$$

as analogues of (A.92) and (A.93). Thanks to assumption (A.96), the difference in asymptotically negligible

$$
\begin{aligned}
S^{\widehat{\alpha}}_{pq}(t,t') &= S^{\tilde{\alpha}}_{pq}(t,t') + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right), \\
Q^{\widehat{\alpha}}_{pq}(t,t') &= Q^{\tilde{\alpha}}_{pq}(t,t') + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right),
\end{aligned}
$$

uniformly in $(t,t') \in [0,1]^2$ as $N \to \infty$. The rest of the proof follows from the proof of Lemma 13. □

We are now ready to prove our main result.

Proof of Theorem 10. The proof is now quite a simple application of Lemma 13 and Corollary 6. First note that even though these results are formulated for the estimation of the covariance kernel $a(\cdot,\cdot)$, they can be likewise applied for the estimation of $b(\cdot,\cdot)$ due to their symmetry in the separable model (4.1).

The estimator $\widehat{a}_0(\cdot,\cdot)$ is realised by smoothing the raw covariances $G_{nmm'}$ without any weights, thus corresponding to the initial guess $\beta(s,s') \equiv 1$, $s,s' \in [0,1]$. Therefore its asymptotic behaviour is by Lemma 13:

$$
\widehat{a}_0(t,t') = \Theta a(t,t') + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right)
$$

uniformly in $(t,t') \in [0,1]^2$ as $N \to \infty$ where Θ is defined in (4.26).

Now, applying Corollary 6 three times and by assumption (B10) we obtain

$$
\begin{aligned}
\widehat{b}_0(s,s') &= \frac{1}{\Theta} b(s,s') + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right), \\
\widehat{a}(t,t') &= \Theta a(t,t') + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right), \\
\widehat{a}(s,s') &= \frac{1}{\Theta} b(s,s') + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right),
\end{aligned}
$$

uniformly in $(t,t') \in [0,1]^2$ or $(s,s') \in [0,1]^2$, as $N \to \infty$. □

Proof of Corollary 4. This corollary follows directly by applying Theorem 10 onto the right hand side of:

$$\left|\widehat{a}(t,t')\widehat{b}(s,s') - a(t,t')b(s,s')\right| \le \left|\widehat{a}(t,t') - \Theta\widehat{a}(t,t')\right| \left|\widehat{b}(s,s')\right| \\ + \left|a(t,t')\right| \Theta \left|\widehat{b}(s,s') - \frac{1}{\Theta} b(s,s')\right|. \qquad \square$$

Proof of Proposition 13

The noise level estimator asymptotic behaviour is treated analogously to previous estimators of the mean surface $\mu(\cdot,\cdot)$ and the covariance kernels $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$.

The estimator $\widehat{V}(t,s)$ is formed by smoothing the raw covariances G_{nmm} agains (t_{nm}, s_{nm}). for $m = 1,\dots,M_n$, $n = 1,\dots,N$. Therefore we form the sequence of vectors $\{(U_k, V_k, Z_k)\}_{k=1}^{\infty}$ by putting $\{t_{nm}\}$, $\{s_{nm}\}$ and $\{\tilde{G}_{nmm}\}$ (defined in (A.90)) in order such that

$$\begin{aligned}
\{U_1, U_2, \dots\} &= \{t_{11}, t_{12}, \dots, t_{1m_1}, t_{21}, \dots, t_{2m_2}, t_{31}, \dots\}, \\
\{V_1, V_2, \dots\} &= \{s_{11}, s_{12}, \dots, s_{1m_1}, s_{21}, \dots, s_{2m_2}, s_{31}, \dots\}, \\
\{Z_1, Z_2, \dots\} &= \{\tilde{G}_{111}, \tilde{G}_{122}, \dots, \tilde{G}_{1m_1m_1}, \tilde{G}_{211}, \dots, \tilde{G}_{2m_2m_2}, \tilde{G}_{311}, \dots\}.
\end{aligned}$$

By following the steps of the proof of Lemma 11 or Lemma 13. Verifying $\mathbb{E}\left[Z_1 | U_1 = t, V_1 = s\right] = a(t,t)b(s,s) + \sigma^2$ for $t, s \in [0,1]$, the asymptotic equivalence (A.95), and assumption (B11) implies

$$\widehat{V}(t,s) = a(t,t)b(s,s) + \sigma^2 + \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log N}{Nh^2}} + h^2\right)$$

uniformly in $(t,s) \in [0,1]^2$ as $N \to \infty$.

This fact, together with Corollary 4 reduced to $(t,s,t',s') = (t,s,t,s)$ implies the statement of Proposition 13.

Proof of Theorem 11

By Proposition 12 and Theorem 10, the model components μ, a, b, σ^2 are estimated consistently. Moreover, consider all the following statements conditionally on $\mathbb{Y}^{new}$. Consequently,

$$\widehat{\mathrm{Var}}(\mathbb{Y}^{new}) \overset{def}{=} \left(\widehat{a}(t_m^{new}, t_{m'}^{new})\widehat{b}(s_m^{new}, s_{m'}^{new}) + \widehat{\sigma}^2 \mathbb{1}_{[m=m']}\right)_{m,m'=1}^{M^{new}} \xrightarrow{\mathbb{P}} \mathrm{Var}(\mathbb{Y}^{new}),$$

as $N \to \infty$, in the matrix space $\mathbb{R}^{M^{new} \times M^{new}}$. Due to continuity of the matrix inversion and the fact that $\mathrm{Var}(\mathbb{Y}^{new})$ is positive definite,

$$\left[\widehat{\mathrm{Var}}(\mathbb{Y}^{new})\right]^{-1} \xrightarrow{\mathbb{P}} \left[\mathrm{Var}(\mathbb{Y}^{new})\right]^{-1} \qquad \text{as} \quad N \to \infty.$$

Moreover

$$\widehat{\mathrm{Cov}}(X^{new}(t,s), \mathbb{Y}^{new}) \overset{def}{=} \left(\widehat{a}(t, t_m^{new})\widehat{b}(s, s_m^{new})\right)_{m=1}^{M^{new}} = \mathrm{Cov}(X^{new}(t,s), \mathbb{Y}^{new}) + o_{\mathbb{P}}(1),$$

as $N \to \infty$, in the supremum norm over $(t,s) \in [0,1]^2$. Therefore, together with the consistency of $\widehat{\mu}$ in the supremum norm, we conclude the statement (4.29).

Assuming (A2), we conclude by the similar steps as above that

$$\sup_{(t,s,t',s')\in[0,1]^4} \left| \widehat{\mathrm{Cov}}\left(X^{new}(t,s), X^{new}(t',s')|\mathbb{Y}^{new}\right) - \mathrm{Cov}\left(X^{new}(t,s), X^{new}(t',s')|\mathbb{Y}^{new}\right)\right| = o_{\mathbb{P}}(1), \tag{A.97}$$

as $N \to \infty$.

Fixing $(t,s) \in [0,1]^2$ we have the conditional distribution given $\mathbb{Y}^{new}$

$$\frac{\Pi(X^{new}(t,s)|\mathbb{Y}^{new}) - X^{new}(t,s)}{\mathrm{Var}\left(X^{new}(t,s)|\mathbb{Y}^{new}\right)} \sim N(0,1)$$

where the denominator is positive for all $t, s \in [0,1]$. Therefore

$$\mathbb{P}\left(|\Pi(X^{new}(t,s)|\mathbb{Y}^{new}) - X^{new}(t,s)| \le u_{1-\alpha}\sqrt{\mathrm{Var}\left(X^{new}(t,s)|\mathbb{Y}^{new}\right)} \,\middle|\, \mathbb{Y}^{new}\right) = 1-\alpha.$$

Now, since

$$\frac{\widehat{\Pi}(X^{new}(t,s)|\mathbb{Y}^{new}) - X^{new}(t,s)}{\sqrt{\widehat{\mathrm{Var}}\left(X^{new}(t,s)|\mathbb{Y}^{new}\right)}} \xrightarrow{d} N(0,1), \qquad \text{as} \quad N \to \infty.$$

where d denotes the convergence in distribution and therefore

$$\mathbb{P}\left(\left|\widehat{\Pi}(X^{new}(t,s)|\mathbb{Y}^{new}) - X^{new}(t,s)\right| \le u_{1-\alpha}\sqrt{\widehat{\mathrm{Var}}\left(X^{new}(t,s)|\mathbb{Y}^{new}\right)} \,\middle|\, \mathbb{Y}^{new}\right) \to 1-\alpha.$$

It remains to justify the asymptotic coverage of the simultaneous confidence band. By the constriction of the simultaneous confidence bands à la Degras (2011), reviewed in Section 4.6, we have

$$\mathbb{P}\left(\sup_{(t,s)\in[0,1]^2} |\Pi(X^{new}(t,s)|\mathbb{Y}^{new}) - X^{new}(t,s)| \le z_{1-\alpha}\sqrt{\mathrm{Var}\left(X^{new}(t,s)|\mathbb{Y}^{new}\right)} \,\middle|\, \mathbb{Y}^{new}\right)$$

equal to $1-\alpha$. where the quantile $z_{1-\alpha}$ is calculated from the law of $W = \sup_{(t,s)\in[0,1]^2} |Z(t,s)|$ where the true (non-estimated) correlations are used:

$$\mathrm{Cov}(Z(t,s), Z(t',s')) = \mathrm{Corr}\left(X^{new}(t,s), X^{new}(t',s')|\mathbb{Y}^{new}\right)$$

with $t, t', s, s' \in [0,1]$. Recall that we denote the empirical analogue of this law as $\widehat{W}$

already defined in (4.25).

In other words

$$\sup_{(t,s)\in[0,1]^2}\left|\frac{\Pi(X^{new}(t,s)|\mathbb{Y}^{new})-X^{new}(t,s)}{\sqrt{\mathrm{Var}\left(X^{new}(t,s)|\mathbb{Y}^{new}\right)}}\right|\sim W,\qquad \text{conditionally on}\,\mathbb{Y}^{new},$$

and therefore

$$\sup_{(t,s)\in[0,1]^2}\left|\frac{\widehat{\Pi}(X^{new}(t,s)|\mathbb{Y}^{new})-X^{new}(t,s)}{\sqrt{\widehat{\mathrm{Var}}\left(X^{new}(t,s)|\mathbb{Y}^{new}\right)}}\right|\xrightarrow{d} W,\qquad \text{as}\quad N\to\infty,\quad \text{conditionally on}\,\mathbb{Y}^{new}.$$

Now, if $c_n(\cdot,\cdot,\cdot,\cdot)\to c(\cdot,\cdot,\cdot,\cdot)$ uniformly (cf. (A.97)), then $N(0,c_n)\xrightarrow{d}N(0,c)$. Therefore $\widehat{W}\xrightarrow{d}W$ and thus $\widehat{z}_{1-\alpha}\to z_{1-\alpha}$ where $\widehat{z}_{1-\alpha}$ and $z_{1-\alpha}$ are the quantiles calculated from the law of $\widehat{W}$ and W respectively. We conclude the proof by observing

$$\begin{aligned}&\mathbb{P}\left(\sup_{(t,s)\in[0,1]^2}\frac{\left|\widehat{\Pi}(X^{new}(t,s)|\mathbb{Y}^{new})-X^{new}(t,s)\right|}{\sqrt{\widehat{\mathrm{Var}}\left(X^{new}(t,s)|\mathbb{Y}^{new}\right)}}\leq\widehat{z}_{1-\alpha}\,\middle|\,\mathbb{Y}^{new}\right)\\&\quad=\mathbb{P}\left(\sup_{(t,s)\in[0,1]^2}\frac{\left|\widehat{\Pi}(X^{new}(t,s)|\mathbb{Y}^{new})-X^{new}(t,s)\right|}{\sqrt{\widehat{\mathrm{Var}}\left(X^{new}(t,s)|\mathbb{Y}^{new}\right)}}\frac{z_{1-\alpha}}{\widehat{z}_{1-\alpha}}\leq z_{1-\alpha}\,\middle|\,\mathbb{Y}^{new}\right)\to 1-\alpha\end{aligned}$$

as $N\to\infty$.

Appendix B: **surfcov** package

From the software development point of view, the methodologies of Chapter 2 and Chapter 3 are completely self-contained, requiring only standard linear algebra libraries implementing the matrix-matrix multiplication and singular value decomposition. Therefore, these methodologies are made available in the form of a stand-alone `R` package `surfcov`, available on GitHub[2].

Contrarily, the methodology of Chapter 4 is not included in the `surfcov` package, because it requires some external functions. Most notably, we use the local linear smoothers and cross-validation for the smoothers' bandwidth selection as implemented in the `fdapace` package (Chen et al., 2020b).

The `surfcov` package was, however, developed only after most of the simulation studies and data analyses in this thesis had been finished. Hence for the purposes of reproducing the simulation studies and data analyses in this thesis, it does not have to be used. Instead, use another GitHub repository[3] for reproducibility. This repository contains all the codes and instructions on how to reproduce the results reported in the thesis.

The purpose of the `surfcov` package is to distribute the main methodological advancements of this thesis: computationally efficient covariance estimation for random surfaces beyond separability. This thesis offers two such generalizations of separability:

- the separable-plus-banded model

  $$c(t,s,t',s') = a_1(t,t')a_2(s,s') + b(t,s,t',s'),$$

  where $b(t,s,t',s') = 0$ for $\max(|t-t'|,|s-s'|) > \delta$ for some $\delta \in [0,1)$, and

- the truncated separable component decomposition (i.e. the R-separable covariance)

  $$c(t,s,t',s') = \sum_{r=1}^{R} \sigma_r a_r(t,t') b_r(s,s'),$$

---

[2] https://github.com/TMasak/surfcov

[3] https://github.com/TMasak/cerss

where $R \in \mathbb{N}$.

In order to achieve the same level computational efficiency as offered by separability in the separable-plus-banded model, the banded part needs to be either diagonal (corresponding to $\delta = 0$) or stationary. Altogether, we can speak of three covariance decompositions:

I. the separable-plus-stationary model,

II. the separable-plus-diagonal model, and

III. the R-separable covariance.

These three generalizations of separability can be estimated, applied, and inverted with computational costs similar to separability. The same can be said about simulating data from such covariances. However, the main goal of the `surfcov` package is to provide handles that can be used to check whether these generalization can be of interest, when having a specific data set at hand.

The package itself can be installed by loading the `devtools` package in `R` and running the following command: `install_github("TMasak/surfcov")`.

Assume now that the `surfcov` package is installed (and loaded) and that we have a data array `X` of dimensions `N x K1 x K2`, where `N` stands for the sample size, and `K1` and `K2` stand for the grid sizes in the respective domains. The natural way to see, whether one of the three generalizations of separability above can be useful to fit the covariance of this data set, is to run the respective cross-validation algorithms:

I. `spb(X)`

II. `spb(X,stationary=F)`

III. `scd(X)`

All of these return a list containing the estimates (the mean and the respective covariance components), the cross-validation objective values, and the cross-validated choice of the parameter (the discrete bandwidth `d` in the case of the separable-plus-banded model and the degree-of-separability R for the separable component decomposition). The commands above use the default grid values for these parameters. To see how to change these, please consult the respective function's documentations, e.g. by `?scd`. The default is the fit-based cross-validation described in Section 2.2.2 and Section 3.3.1, respectively. To see how different values of the parameter affect the prediction performance, use e.g. `spb(X,predict=T)`, which performs a cross-validated out-of-sample comparison of the prediction performance for different values of the bandwidth.

Note that in the case o the `stationary=F` in the separable-plus-banded model, only the cases `d=0`, `d=1` and `d=2` can be handled computationally efficiently[4]. Still, it is possible to spend additional computational resources to search past `d=2`.

Example 6. *Let* `X` *contain the mortality surfaces analyzed in Section 2.5.2. In the case of this data set,* `spb(X)` *returns* `d=0`*, suggesting that allowing for a stationary banded part does not improve the fit. Running* `spb(X,stationary=F)` *instead returns* `d=1`*, suggesting the presence of (heteroscedastic, since stationarity was not useful before) white noise.*

Finally, validity of the separable-plus-banded model can also be checked by running the bootstrap test described in Section 2.2.3. For example, running

```
test_spb(X,d=1,stationary=F)
```

tests the validity of the separable-plus-diagonal model.

To summarize, the `surfcov` package can be used in the way described above to check suitability of the generalizations of separability described in this thesis for a data set at hand. For the other features of the package, including the efficient inversion algorithm, see the GitHub page of the `surfcov` package[5].

---

[4] The case of `d=2` is not yet implemented efficiently in the `surfcov` package, but it can be done due to the tridiagonal matrix algorithm (Thomas, 1949)

[5] https://github.com/TMasak/surfcov

Bibliography

Aneiros, G., Cao, R., Fraiman, R., Genest, C., and Vieu, P. (2019). Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis*, 170:3–9.

Aston, J. A., Pigoli, D., and Tavakoli, S. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, 45(4):1431–1461.

Aue, A., Norinho, D. D., and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110(509):378–392.

Bagchi, P. and Dette, H. (2020). A test for separability in covariance operators of random surfaces. *Annals of Statistics*, 48(4):2303–2322.

Baíllo, A., Cuevas, A., and Fraiman, R. (2011). Classification methods for functional data. *The Oxford handbook of functional data analysis.*

Bijma, F., De Munck, J. C., and Heethaar, R. M. (2005). The spatiotemporal meg covariance matrix modeled as a sum of Kronecker products. *NeuroImage*, 27(2):402–415.

Billingsley, P. (1999). *Convergence of Probability Measures.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654.

Bosq, D. (2012). *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media.

Brockwell, P. J., Davis, R. A., and Fienberg, S. E. (1991). *Time Series: Theory and Methods: Theory and Methods.* Springer Science & Business Media.

Cai, T. and Yuan, M. (2010). *Nonparametric covariance function estimation for functional and longitudinal data.* University of Pennsylvania and Georgia inistitute of technology. Technical report, url: http://www-stat.wharton.upenn.edu/~tcai/paper/Covariance-Function.pdf.

Cai, T. T., Ren, Z., and Zhou, H. H. (2013). Optimal rates of convergence for estimating toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1):101–143.

Chan, R. H.-F. and Jin, X.-Q. (2007). *An introduction to iterative Toeplitz solvers*, volume 5. SIAM.

Chen, K., Delicado, P., and Müller, H.-G. (2017a). Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):177–196.

Chen, K. and Müller, H.-G. (2012). Modeling repeated functional observations. *Journal of the American Statistical Association*, 107(500):1599–1609.

Chen, K., Zhang, X., Petersen, A., and Müller, H.-G. (2017b). Quantifying infinite-dimensional data: Functional data analysis in action. *Statistics in Biosciences*, 9(2):582–604.

Chen, W., Genton, M. G., and Sun, Y. (2021). Space-time covariance structures and models. *Annual Review of Statistics and Its Application*, 8.

Chen, X., Yang, D., Xu, Y., Xia, Y., Wang, D., and Shen, H. (2020a). Testing and support recovery of correlation structures for matrix-valued observations with an application to stock market data. *arXiv preprint arXiv:2006.16501*.

Chen, Y., Carroll, C., Dai, X., Fan, J., Hadjipantelis, P. Z., Han, K., Ji, H., Müller, H.-G., and Wang, J.-L. (2020b). *fdapace: Functional Data Analysis and Empirical Dynamics*. R package version 0.5. url: https://cran.r-project.org/web/packages/fdapace/.

Chiou, J.-M., Chen, Y.-T., and Yang, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, pages 1571–1596.

Constantinou, P., Kokoszka, P., and Reimherr, M. (2017). Testing separability of space-time functional processes. *Biometrika*, 104(2):425–437.

Cont, R. and Da Fonseca, J. (2002). Dynamics of implied volatility surfaces. *Quantitative finance*, 2:45–60.

Conway, J. B. (2019). *A course in functional analysis*, volume 96. Springer.

Cox, J. C., Ross, S. A., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of financial Economics*, 7(3):229–263.

Crane, E., Childers, D., Gerstner, G., and Rothman, E. (2011). Functional data analysis for biomechanics. *Theoretical biomechanics*, pages 77–92.

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122.

Davis, P. J. (2013). *Circulant matrices*. American Mathematical Society.

Degras, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, pages 1735–1765.

Delaigle, A., Hall, P., and Bathia, N. (2012). Componentwise classification and clustering of functional data. *Biometrika*, 99(2):299–313.

Descary, M.-H. (2017). *Functional data analysis by matrix completion.* PhD thesis, Ecole Polytechnique Fédérale de Lausanne.

Descary, M.-H. and Panaretos, V. M. (2019). Functional data analysis by matrix completion. *The Annals of Statistics*, 47(1):1–38.

Dette, H., Dierickx, G., and Kutta, T. (2020). Quantifying deviations from separability in space-time functional processes. *arXiv preprint arXiv:2003.12126*.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of statistical computation and simulation*, 64(2):105–123.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66. Chapman and Hall & CRC.

Fengler, M. R. (2009). Arbitrage-free smoothing of the implied volatility surface. *Quantitative Finance*, 9(4):417–428.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice.* Springer Science & Business Media.

Genton, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics: The official journal of the International Environmetrics Society*, 18(7):681–695.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600.

Gneiting, T., Genton, M. G., and Guttorp, P. (2006). *Geostatistical Space-Time Models, Stationarity, Separability, and Full Symmetry*, pages 151–175. Chapman and Hall & CRC.

Gohberg, I. and Krein, M. G. (1978). *Introduction to the theory of linear nonselfadjoint operators*, volume 18. American Mathematical Society.

Greenewald, K., Zhou, S., and Hero III, A. (2019). Tensor graphical lasso (teralasso). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(5):901–931.

Grenander, U. (1981). *Abstract inference.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2011). Longitudinal functional principal component analysis. In *Recent Advances in Functional Data Analysis and Related Topics*, pages 149–154. Springer.

Gupta, A. K. and Nagar, D. K. (2018). *Matrix variate distributions*, volume 104. Chapman and Hall & CRC.

Haase, M. (2014). *Functional analysis: an elementary introduction*, volume 156. American Mathematical Society.

Hall, P., Fisher, N. I., and Hoffmann, B. (1994). On the nonparametric estimation of covariance functions. *The Annals of Statistics*, pages 2115–2134.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126.

Hall, P. and Patil, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields*, 99(3):399–424.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748.

Hasenstab, K., Scheffler, A., Telesca, D., Sugar, C. A., Jeste, S., DiStefano, C., and Şentürk, D. (2017). A multi-dimensional functional principal components analysis of EEG data. *Biometrics*, 73(3):999–1009.

Haslett, J. and Raftery, A. E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 38(1):1–21.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of statistical software*, 27(5):1–32.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators.* John Wiley & Sons.

Huang, H. and Sun, Y. (2019). Visualization and assessment of spatio-temporal covariance properties. *Spatial Statistics*, 34:100272.

Hull, J. (2006). *Options, Futures, and Other Derivatives.* Pearson International edition. Pearson/Prentice Hall.

Jiang, C.-R., Aston, J. A., and Wang, J.-L. (2009). Smoothing dynamic positron emission tomography time courses using functional principal components. *NeuroImage*, 47(1):184–193.

Jirak, M. (2016). Optimal eigen expansions and uniform bounds. *Probability Theory and Related Fields*, 166(3-4):753–799.

Jolliffe, I. (1986). *Principal Component Analysis.* Springer Verlag.

Kearney, F., Cummins, M., and Murphy, F. (2018). Forecasting implied volatility in foreign exchange markets: A functional time series approach. *The European Journal of Finance*, 24(1):1–18.

Kidziński, Ł. and Hastie, T. (2018). Longitudinal data analysis using matrix completion. *arXiv preprint arXiv:1809.08771.*

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

Kraus, D. and Stefanucci, M. (2019). Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika*, 106(1):161–180.

Langrené, N. and Warin, X. (2019). Fast and stable multivariate kernel density estimation by fast sum updating. *Journal of Computational and Graphical Statistics*, 28(3):596–608.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321–3351.

Lopez, G., Eisenberg, D. P., Gregory, M. D., Ianni, A. M., Grogans, S. E., Masdeu, J. C., Kim, J., Groden, C., Sidransky, E., and Berman, K. F. (2020). Longitudinal positron emission tomography of dopamine synthesis in subjects with gba1 mutations. *Annals of neurology*, 87(4):652–657.

Lynch, B. and Chen, K. (2018). A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika*, 105(4):815–831.

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(14):1503–1509.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of economics and management science*, pages 141–183.

Miller, F., Neill, J., and Wang, H. (2008). Nonparametric clustering of functional data. *Statistics and its interface*, 1(1):47–62.

Mueller, H.-G. (2016). Peter hall, functional data analysis and random objects. *The Annals of Statistics*, 44(5):1867–1887.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, 33(2):774–805.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective.* MIT press.

Palitta, D. and Simoncini, V. (2020). Optimality properties of galerkin and petrov–galerkin methods for linear matrix equations. *Vietnam Journal of Mathematics*, pages 1–17.

Park, S., Shedden, K., and Zhou, S. (2017). Non-separable covariance models for spatio-temporal data, with applications to neural encoding analysis. *arXiv preprint arXiv:1705.05265*.

Park, S. Y. and Staicu, A.-M. (2015). Longitudinal functional data analysis. *STAT*, 4(1):212–226.

Pigoli, D., Hadjipantelis, P. Z., Coleman, J. S., and Aston, J. A. (2018). The statistical analysis of acoustic phonetic data: exploring differences between spoken romance languages. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1103–1145.

Pomann, G.-M., Staicu, A.-M., and Ghosh, S. (2016). A two sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(3):395.

Prabhakar, S. K. and Rajaguru, H. (2020). Alcoholic EEG signal classification with correlation dimension based distance metrics approach and modified adaboost classification. *Heliyon*, 6(12):e05689.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. url: https://www.R-project.org/.

Ramsay, J. and Silverman, B. (2007). *Applied Functional Data Analysis: Methods and Case Studies.* Springer Series in Statistics. Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis.* Springer, New York.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing L1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Raykar, V. C., Duraiswami, R., and Zhao, L. H. (2010). Fast computation of kernel estimators. *Journal of Computational and Graphical Statistics*, 19(1):205–220.

Rougier, J. (2017). A representation theorem for stochastic processes with separable covariance functions, and its implications for emulation. *arXiv preprint arXiv:1702.05599.*

Rubín, T. and Panaretos, V. M. (2020). Sparsely observed functional time series: Estimation and prediction. *Electronic Journal of Statistics*, 14(1):1137–1210.

Schumacher, B. and Westmoreland, M. (2010). *Quantum processes systems, and information.* Cambridge University Press.

Shewchuk, J. R. (1994). *An introduction to the conjugate gradient method without the agonizing pain.* Carnegie-Mellon University. Department of Computer Science. Technical report, url: https://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf.

Silverman, B. W. (1982). Algorithm AS 176: Kernel density estimation using the fast fourier transform. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(1):93–99.

Simoncini, V. (2016). Computational methods for linear matrix equations. *SIAM Review*, 58(3):377–441.

Sørensen, H., Goldsmith, J., and Sangalli, L. M. (2013). An introduction with medical applications to functional data analysis. *Statistics in medicine*, 32(30):5222–5240.

Stoehr, C., Aston, J. A., and Kirch, C. (2020). Detecting changes in the covariance structure of functional time series with application to fMRI data. *Econometrics and Statistics.*

Tavakoli, S. (2016). covsep: Tests for determining if the covariance structure of 2-dimensional data is separable. *R package version 1.0.0.*

Tessmer, O. L., Jiao, Y., Cruz, J. A., Kramer, D. M., and Chen, J. (2013). Functional approach to high-throughput plant growth analysis. *BMC systems biology*, 7(6):1–13.

Thomas, L. (1949). *Elliptic problems in linear differential equations over a network.* Watson Scientific Computing Laboratory Report, Columbia University, NY.

Tsiligkaridis, T. and Hero, A. O. (2013). Covariance estimation in high dimensions via Kronecker product expansions. *IEEE Transactions on Signal Processing*, 61(21):5347–5360.

Unser, M. and Tafti, P. D. (2014). *An introduction to sparse stochastic processes.* Cambridge University Press.

Van Loan, C. F. (2000). The ubiquitous Kronecker product. *Journal of computational and applied mathematics*, 123(1-2):85–100.

Van Loan, C. F. and Golub, G. H. (1983). *Matrix computations.* Johns Hopkins University Press.

Van Loan, C. F. and Pitsianis, N. (1993). Approximation with Kronecker products. In *Linear algebra for large scale and real-time applications*, pages 293–314. Springer.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing.* Chapman and Hall & CRC.

Wang, J., Wong, R. K., and Zhang, X. (2020). Low-rank covariance function estimation for multidimensional functional data. *Journal of the American Statistical Association*, pages 1–14.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

Weidmann, J. (2012). *Linear operators in Hilbert spaces*, volume 68. Springer Science & Business Media.

Wilde, M. M. (2013). *Quantum information theory.* Cambridge University Press.

Wilmoth, 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. url: http://mortality.org [version 31/05/2007]*, 9:10–11.

Wong, R. K. and Zhang, X. (2019). Nonparametric operator-regularized covariance function estimation for functional data. *Computational statistics & data analysis*, 131:131–144.

Wu, L. and Stathopoulos, A. (2014). Primme svds: A preconditioned svd solver for computing accurately singular triplets of large matrices based on the primme eigensolver. *arXiv preprint arXiv:1408.5535.*

Xie, L., Ding, J., and Ding, F. (2009). Gradient based iterative solutions for general linear matrix equations. *Computers & Mathematics with Applications*, 58(7):1441–1448.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, pages 2873–2903.

Yarger, D., Stoev, S., and Hsing, T. (2020). A functional-data approach to the argo data. *arXiv preprint arXiv:2006.05020.*

Young, D. M. (2014). *Iterative solution of large linear systems.* Elsevier.

Young, N. (1988). *An introduction to Hilbert space.* Cambridge university press.

Zhang, H. and Li, Y. (2020). Unified principal component analysis for sparse and dense functional data under spatial dependency. *arXiv preprint arXiv:2006.13489.*

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321.

Zhang, X. and Wang, J.-L. (2018). Optimal weighting schemes for longitudinal and functional data. *Statistics & Probability Letters*, 138:165–170.

List of Abbreviations

ADI: alternating direction implicit 56

ALS: alternating least squares 82

BLUP: best linear unbiased predictor 97

CG: conjugate gradient 58

CLT: central limit theorem 16

CMT: continuous mapping theorem 17

CV: cross-validation 49

DFT: discrete Fourier transform 52

DoS: degree-of-separability 80

EEG: electroencephalography 113

FDA: functional data analysis 1

FFT: fast Fourier transform 52

MLE: maximum likelihood estimator 26

NKP: singular value decomposition 29

ONB: orthonormal basis 6

PCA: principal component analysis 14

PCG: preconditioned conjugate gradient 58

PSD: positive semi-definite, positive semi-definiteness 7

SCD: separable component decomposition 80

SVD: singular value decomposition 6

fLDA: separable component decomposition 114

List of Figures

1.1 Structure of $\mathbf{I}_K \otimes_K \mathbf{\Psi}$ and $\mathbf{\Sigma} \otimes_K \mathbf{I}_K$. 28

2.1 Partial tracing visualized. 43

2.2 Shifted partial tracing visualized. 44

2.3 The Legendre and Wiener covariance choices for simulations. 71

2.4 The signed and Epanechnikov choices for simulations. 71

2.5 Estimation errors of separable-plus-banded model estimation – the Legendre-signed case. 72

2.6 Estimation errors of separable-plus-banded model estimation – other cases. 73

2.7 Estimation errors and number of iterations needed to solve the inverse problem depending on the grid size. 75

2.8 Raw mortality rate surfaces for the Czech Republic and Switzerland. . . 76

2.9 Cross-validation objective and log-heatmap of the heteroscedastic white noise. 77

2.10 Leading product eigenfunctions. 77

3.1 Eigenvalues and separable component scores of the Gneiting covariance. 93

3.2 Relative estimation errors of R-separable estimators. 106

3.3 Scree plots. 107

3.4 Prediction runtimes. 108

3.5 Estimation and prediction errors of for R-separable estimators. 109

3.6 Bias-free errors of R-separable estimators. 111

3.7 Number of iterations required for prediction with increasing grid size, and number of iterations versus the leading separable score. 113

4.1 Workflow of the proposed procedure to estimate the separable model from from sparse measurements. 123

4.2 Contribution of raw covariances to estimation. 126

4.3 Covariances used as building blocks in the simulation study. 139

4.4 Relative estimation errors depending on percentages of the surfaces observed. 140

4.5 Estimation runtimes, and estimation errors for multi-step approach. . . 142

4.6 Implied volatilities observed for Dell and Qualcomm. 144

4.7 Marginal kernels and their eigendecompositions, estimated from observed implied volatilities. 145

4.8 Prediction of the latent implied volatility surface for Dell. 146

4.9 Boxplots of root mean square errors relative to the pre-smoothing benchmark. 148

List of Statements

C

Corollary 1: Definiteness of a separable operator. 11

Corollary 2: Multivariate asymptotic distr. for the separable-plus-banded model. . . . 64

Corollary 3: Discrete representation of partial inner product. 84

Corollary 4: Convergence rate for sparsely observed separable model. 136
  Proof 203

Corollary 5: Consequence of Theorem 12 196

Corollary 6: Convergence rate for the spatial kernel smoother. 202

D

Definition 1: Separability. 10

Definition 2: Partial tracing. 22

Definition 3: Partial inner product. 24

Definition 4: Matrix-variate normal distribution. 26

Definition 5: Shifted trace. 34

Definition 6: Shifted partial trace. 34

Definition 7: Shifting operator. 36

Definition 8: General shifted trace. 37

Definition 9: General shifted partial trace. 37

Definition 10: Discrete shifted (partial) trace. 40

Definition 11: Toeplitz averaging. 46

Definition 12: Discrete Toeplitz averaging. 47

Definition 13: Partial inner product. 83

E

Example 1: Shifted partial tracing visualized. 42

Example 2: Power iteration to find principal components. 82

Example 3: Generalized power iteration and matrix-vector multiplication. 84

Example 4: Best separable approximation. 88

Example 5: Irish wind data – degree-of-separability smaller than rank. 92

L

Lemma 1: Eigenstructure of a separable operator. 10

Lemma 2: Identifiability of the separable-plus-banded model. 33

Lemma 3: Banded operators are in the kernel of shifted partial tracing. 35

Lemma 4: Shifted (partial) traces of continuous kernels and their discretizations. .. 40

Lemma 5: Gaussian random elements. 61

Lemma 6: Basic properties of partial inner product. 85

Lemma 7: A simple second moment inequality. 165

Lemma 8: Perturbation bounds for separable component decomposition. 184

Lemma 9: Perturbation bounds for R-separable approximations. 185

Lemma 10: Hilbert-Schmidt vs. Frobenius norms. 187

Lemma 11: Rates for smoothers' building blocks. 197

Lemma 11: Rates for smoothers' building blocks II. 198

Lemma 13: Convergence rates for a single smoothing step. 199

P

Proposition 1: Basic properties of partial tracing. 22

Proposition 2: Integral representation of partial tracing. 23

Proposition 3: Basic properties of the partial inner product. 24

Proposition 4: Integral representation of the partial inner product. 25

Proposition 5: Basic properties of the shifted partial trace. 35

Proposition 6: Characterization of shifted partial trace. 37

Proposition 7: General definition of the shifted partial trace works. 38

Proposition 8: Uniform rates for the estimators of the separable-plus-banded model. 66
  Proof 175

Proposition 9: Asymptotic distribution with adaptive bandwidth. 67
  Proof 181

Proposition 10: Integral representation of partial inner product. 84

Proposition 11: Convergence of generalized power iteration. 86

Proposition 12: Convergence rates for the mean estimated from sparsely observed random surfaces. 136
  Proof 198

Proposition 13: Convergence rates for the noise level. 137
  Proof 204

Proposition 14: Convergence rate for a separable-plus-banded proxy. 177

Proposition 14: Consistency of the empirical CV objective. 178

R

Remark 1: Abstract outer product and Kronecker product terminology. 9

Remark 2: Notation for marginalization operators. 29

Remark 3: Symmetrizing results of the shifted partial trace. 34

Remark 4: Computational constraints placed. 45

Remark 5: Shift is preferable when noise is present. 66

Remark 6: Perks of working with general Hilbert spaces. 80

Remark 7: Partial inner product in multiple dimensions. 85

Remark 8: Relation to previous work. 89

Remark 9: Computational advantages of generalized power iteration. 91

Remark 10: Assumptions for rates of separable component estimators. 100

Remark 11: The inversion algorithm works well. 112

T

Theorem 1: Central limit theorem for the empirical covariance. 17

Theorem 2: Linear transformations of the matrix-variate normal distribution. 26

Theorem 3: Asymptotic distribution for the separable-plus-banded model. 60
Proof 61

Theorem 4: Rates of convergence for discrete observations. 64
Proof 165

Theorem 5: Rates of convergence with adaptive parameter. 67
Proof 177

Theorem 6: Rates of convergence with adaptive bandwidth for discrete observations. 68
Proof 181

Theorem 7: Convergence rates for separable component decomposition and fully observed surfaces. 98
Proof 184

Theorem 8: Convergence rates for discrete observations. 102
Proof 187

Theorem 9: Convergence rates for discrete observations. 103
Proof 193

Theorem 10: Convergence rates for the separable constituents estimated from sparse observations. 136
Proof 203

Theorem 11: Asymptotic coverage of the prediction bands. 137
Proof 204

Theorem 12: Strong mixing conditions for kernel averages. 195

List of Tables

3.1 Power iteration method for finding the leading separable components on a general Hilbert space. 89

3.2 Time and memory complexities of computing the empirical covariance estimator and a separable estimator when N surfaces are observed discretely on a $K \times K$ grid. 91

3.3 Constructing the R-separable estimator from discrete measurements. . . 92

3.4 Relative prediction error in the parametric covariance setup for the R-separable estimators ($R = 1, 2, 3$) and the empirical covariance. Minima for a fixed N (given column) are depicted in bold. 107

3.5 Out-of-sample cross-validated classification accuracy (i.e. the ratio between correctly classified subjects and all subjects) for the fLDA classifier with the separable ($R = 1$) or R-separable ($R = 2$) estimator of the covariance, and for three different data sets given by different conditions. 115

4.2 Algorithm for estimation of the separable model from sparsely observed (w.l.o.g. zero-mean) surfaces. 130

4.3 Complexities for covariance estimation of a random surface observed on a $K \times K$ grid. 131

4.4 Relative estimation errors of 4D smoothing with cross-validated bandwidths and bandwidths suggested by the separable model for the four scenarios considered in Section 4.8 with $p = 2$ percentages of the surfaces observed. 142

Tomas Masak

Curriculum Vitae

✉ *tom.masak@gmail.com*

Education

2022 (expected) **Ph.D. in Statistics**, *École polytechnique fédérale de Lausanne (EPFL)*.
Thesis: *Covariance Estimation for Random Surfaces Beyond Separability*,
Advisor: Prof. **Victor M. Panaretos**

2020 **RNDr. in Mathematical Statistics**, *Charles University*.

2017 **Mgr. (M.Sc.) in Probability, Mathematical Statistics and Econometrics**, *Charles University*.
Thesis: *Big Data – Extraction of Key Information Combining Methods of Mathematical Statistics and Machine Learning* Advisor: Prof. **Jaromír Antoch**

2014 **Bc. (B.Sc.) in General Mathematics**, *Charles University*.
Thesis: *Fault Tree Analysis*, Advisor: Prof. **Jaromír Antoch**

Additional Experience

Fall 2017 **Research & teaching assistant**, *Technische Universität München (TUM)*.
Supervisors: Prof. **Felix Krahmer** & Prof. **Claudia Klüppelberg**

Fall 2016 **Erasmus exchange program**, *TUM*.

Spring 2016 **Visiting scholar**, *Eidgenössische Technische Hochschule Zürich (ETHZ)*.
Supervisor: Prof. **Sara van de Geer**

Fall 2015 **Modelling consultant**, *CSOB (KBC group)*.

Journal Publications

2021 (submitted) Inference and Computation for Sparsely Observed Random Surfaces (with V.M. Panaretos & T. Rubín)

2020 (submitted) Separable Expansions for Covariance Estimation via the Partial Inner Product (with V.M. Panaretos & S. Sarkar)

2020 (submitted) Random Surface Covariance Estimation by Shifted Partial Tracing (with V.M. Panaretos)

2017 Iteratively Reweighted Least Squares Algorithm for Sparse Principal Component Analysis with Application to Voting Records. *Statistika: Statistics and Economy Journal 97*, 88-106

Honors & Awards

2017 **Best Master's thesis**, *Charles University, Department of Probability and Statistics.*

2017 **1st place**, *Student research competition "SVOC" in mathematics and computer science.*

2017 **Students' faculty grant**, *Charles University, Faculty of Mathematics and Physics.*

2016 **Best poster presentation award**, *ROBUST 2016.*

2016 **Mobility fund fellowship**, *Charles University.*

2015-2016 **Scholarship for excellent study results**, *Charles University, Faculty of Mathematics and Physics*, (category A – Top 5%).

Presentations & Posters

2021 Bernoulli-IMS 10th World Congress in Probability and Statistics, *Seoul (online)*, contributed talk

2021 Department of Probability and Mathematical Statistics, *Prague*, seminar talk

2020 **CMStatistics**, *London*, **invited talk**

2020 Low Rank Models 2020, *Switzerland*, poster

2018 ROBUST 2018, *Czech Republic*, poster presentation

2018 Conference Universitaire de Swiss Occidentale, *Geneva*, contributed talk

2018 **Workshop on Sparsity in Applied Mathematics and Statistics**, *Brussels*, **invited talk**

2015 Modelling Smart Grids Workshop at ENBIS 2015, *Prague*, contributed talk

Supervised Projects

2021 Forecasting Functional Time Series, *semester project (Master)*, with Prof. V.M. Panaretos & L. Santoro

2021 Functional Principal Component Analysis with Application to COVID-19 Dynamics, *semester project (Bachelor)*, with Prof. V.M. Panaretos & L. Santoro

2020 Introduction to Non-parametric Density Estimation, *semester project (Bachelor)*, with Prof. V.M. Panaretos (informal)

2019 Modelling of Selective Nerve Stimulation for Prostheses with Sensory Feedback Using Machine Learning Methods, *semester project (Master)*, with Prof. F. Eisenbrandt & Dr. I. Malinovic

2019 Matrix-variate Normal Distribution, *semester project (Bachelor)*, with Prof. V.M. Panaretos

Research Interests

FDA	Functional data analysis on multi-dimensional domains, separability of covariance operators, non-parametric modelling, inverse problems, classification
PCA	NP-hard versions of principal component analysis: sparse PCA, matrix completion and robust PCA
IRLS	Iteratively reweighted least squares algorithms for non-convex optimization

Teaching Record

Fall 2021	**Principal teaching assistant (TA)**, Linear Models, *EPFL* (Prof. V.M. Panaretos).
Spring 2021	**Principal TA**, Statistique pour Mathematiciens, *EPFL* (Dr. M. Suveges).
Fall 2020	**Principal TA**, Linear Models, *EPFL* (Prof. V.M. Panaretos).
Spring 2020	**TA**, Statistique pour Mathematiciens, *EPFL* (Prof. V.M.Panaretos).
Fall 2019	**Principal TA**, Linear Models, *EPFL* (Prof. V.M. Panaretos).
Spring 2019	**Principal TA**, Time Series, *EPFL* (Prof. A.C. Davison).
Fall 2018	**Principal TA**, Statistics for Data Science, *EPFL* (Prof. V.M. Panaretos).
Spring 2018	**TA**, Time Series, *EPFL* (Dr. E. Thibaud).
Fall 2017	**Teaching fellow (TF)**, Time Series Analysis, *TUM* (Prof. C. Klüppelberg). 4 hours/week of independent teaching
Fall 2015	**TF**, Probability and Math. Statistics, *Charles University* (Prof. D. Hlubinka).
Fall 2015	**TF**, Statistics, *Charles University, IES* (Dr. M. Cervinka).
Spring 2015	**TA**, Mathematics 1A, Czech Technical University (Prof. P. Kucera).
Lecturing	Respectively 4, 2, and 2 hours, replacing Prof. A.C. Davison or Prof. V.M. Panaretos for the courses Time Series, Linear Models, and Statistics for Data Science, *EPFL*

Refereeing

Journal of the American Statistical Association

Biometrika

Computtional Statistics

IEEE Journal of Selected Topics in Signal Processing

Skills

Languages	English (advanced), German (intermediate), French (beginner), Czech (mother tongue)
Programming	R (advanced), Matlab (advanced), C++ (previously intermediate, now rusty), HTML/CSS (beginner), PHP (beginner), SQL (beginner)
Data Analysis	Functional data, time series, regression (GLM, GAM, etc.), regularization/shrinkage, missing data, statistical learning, experimental design