

Quantifying deviations from separability in space-time functional processes

Holger Dette, Gauthier Dierickx, Tim Kutta

March 26, 2020

Abstract

The estimation of covariance operators of spatio-temporal data is in many applications only computationally feasible under simplifying assumptions, such as separability of the covariance into strictly temporal and spatial factors. Powerful tests for this assumption have been proposed in the literature. However, as real world systems, such as climate data are notoriously inseparable, validating this assumption by statistical tests, seems inherently questionable. In this paper we present an alternative approach: By virtue of separability measures, we quantify how strongly the data's covariance operator diverges from a separable approximation. Confidence intervals localize these measures with statistical guarantees. This method provides users with a flexible tool, to weigh the computational gains of a separable model against the associated increase in bias. As separable approximations we consider the established methods of partial traces and partial products, and develop weak convergence principles for the corresponding estimators. Moreover, we also prove such results for estimators of optimal, separable approximations, which are arguably of most interest in applications. In particular we present for the first time statistical inference for this object, which has been confined to estimation previously.

Besides confidence intervals, our results encompass tests for approximate separability. All methods proposed in this paper are free of nuisance parameters and do neither require computationally expensive resampling procedures nor the estimation of nuisance parameters. A simulation study underlines the advantages of our approach and its applicability is demonstrated by the investigation of German annual temperature data.

Keywords and Phrases: Space-time processes, approximate separability, partial traces, optimal approximation, self-normalization

AMS Subject Classification: 62G10, 62G20

1 Introduction

Spatio-temporal data is ubiquitous in many branches of modern statistics, such as medicine (Worsley et al., 1996; Lindquist, 2008; Skup, 2010), urban pollution (Krall et al., 2015), climate research (Laurini, 2019; Chattopadhyay et al., 2020), spectrograms derived from audio signals or geolocalized data (Rabiner and Schafer, 1978; Bel et al., 2011) and perhaps most prominently in geostatistics (Mitchell et al., 2006; Gneiting et al., 2006). For a review of spatio-temporal models with an extensive list of fields of applications see Kyriakidis and Journel (1999) and the references therein. An important step in the statistical analysis

of this type of data is the estimation of the covariance structure, which is required, for example, for principal component analysis, prediction or kriging. Nowadays, technological advances often yield ultra high-dimensional space-time data and a completely non-parametric estimation of the covariance operator is computationally very expensive (if not impossible). In some applications, such as the European Union earth observation program *Copernicus*, where 4.4Pb data per year are collected, even the storage of the estimated covariance structure is not possible, for instance see <https://www.copernicus.eu/en/access-data>.

A common approach to drastically reduce the number of parameters in the covariance matrix and to obtain computationally feasible procedures for large space-time data sets, is the imposition of structural assumptions on the covariance of the underlying process. One important example in the analysis of space-time data is *separability*, which means that the covariance is modeled as product of a purely spatial and a purely temporal part (see Kyriakidis and Journel, 1999; Gneiting et al., 2006; Genton, 2007; Sherman, 2010; Cressie and Wikle, 2011; Cressie, 2015). The simplifying power of this assumption is tempting in applications, even where separability is not physically self-evident. As a consequence statistical tests for the hypothesis of separability have been developed to validate this assumption where possible and to discard it where not. For example, Matsuda and Yajima (2004); Scaccia and Martin (2005); Fuentes (2006); Crujeiras et al. (2010) proposed tests based on spectral methods for stationarity, axial symmetry and separability. In Mitchell et al. (2005, 2006); Lu and Zimmerman (2005) likelihood ratio tests under the assumption of a normal distribution were investigated. Finite dimensional non-parametric tests for separability were considered in Li et al. (2007). More recent work on testing for separability in a functional, non-parametric framework can be found in the papers of Constantinou et al. (2017); Aston et al. (2017); Constantinou et al. (2018) or Bagchi and Dette (2020).

On the other hand many authors argue that, in applications it is often implausible to assume the complete absence of any space-time interactions in the covariance, which however is essential to separability (see, for example, Kyriakidis and Journel, 1999; Cressie and Huang, 1999; Gneiting, 2002; Huang and Hsu, 2004; Jun and Stein, 2004; Stein, 2005; Lu and Zimmerman, 2005; Li et al., 2007; Tingley and Huybers, 2013). This concern is strengthened by the fact that in most empirical studies found in the literature, the hypothesis of separability is rejected (see Mitchell et al., 2006; Fuentes, 2006; Gneiting et al., 2006; Aston et al., 2017; Constantinou et al., 2018; Bagchi and Dette, 2020). Consequently, testing the hypothesis of exact separability might be questionable from a decision theoretical point of view (as one already knows that it cannot be satisfied). While conceptual critiques haunt "exact separability" in theoretical discussions, we want to highlight a more pragmatic interpretation of this assumption, common in applications: Here separability means that a separable operator provides a reasonable approximation of the true covariance structure (for instance, see Genton, 2007; Aston et al., 2017).

In the present work, we develop statistical methodology, which corresponds to this latter point of view. Instead of testing the hypothesis of "exact separability", we quantify how strongly a covariance operator deviates from separability. This approach is not only more realistic than previous tests, but also offers greater flexibility to users, who can judge in their own case, whether a deviation from separability is too large to justify a separable model, or whether computational gains outweigh the ensuing imprecision.

Our analysis of separability is based on the normed difference between the covariance operator and a separable approximation, which vanishes if and only if exact separability holds. We apply these measures to the empirical covariance operator from the data, to quantify the deviation from separability of the underlying covariance structure. Naturally the outcome depends critically on the particular method of separable approximation. We examine three types of approximations from a general perspective. Two of them have

recently attracted attention in the problem of testing for exact separability, namely partial traces (Constantinou et al., 2017, 2018; Aston et al., 2017) and partial products (Bagchi and Dette, 2020). In order to bridge the aforementioned gap between earlier theory of "exact separability" and an inseparable reality we develop weak invariance principles to encompass models with inseparable covariance operators. Our most important contribution consists in the examination of the third type, the *optimal separable approximation* and the derivation of statistical inference for this vital object. While suboptimal approximations - as treated in previous literature on tests - are reasonable tools to test for exact separability, they are prone to overestimate deviations from separability. On the other hand, optimal approximations do not suffer from this imprecision. Until now optimal, separable approximations have been confined to finite dimensional estimation (Genton, 2007), and inference seemed out of reach due to their implicit definition. However in this paper we present a novel, stochastic linearization, which is employed to prove a weak invariance principle for the empirical, sequential estimator of the distance between the covariance operator and an optimal approximation. Hence our theory provides tools to develop statistical inference for the current estimation practice.

More precisely, the new weak invariance principles for all three approximation types are combined with self-normalization techniques to construct asymptotically pivotal estimates for the measures of deviation from separability. Therewith we construct confidence intervals, such that deviations from separability can be quantified with statistical guarantees. As a further application we develop (asymptotically) valid statistical tests for the null hypothesis that *"the distance between the covariance and a separable approximation is small"*. This hypothesis of *"approximate separability"* sets our approach apart from the body of existing tests, which exclusively focus on the more theoretical concept of exact separability.

From a mathematical point of view our approach has several challenges. First - in contrast to tests for exact separability - the investigation of asymptotic properties cannot be conducted under the assumption of separability (the null hypothesis) and weak convergence of the estimators (and their sequential versions) has to be established for any underlying (not necessarily separable) covariance structure. Secondly, even if weak convergence of an estimator can be established in the general case, the limiting distribution depends on several nuisance parameters, which are extremely difficult to estimate. While in some cases the bootstrap might be a solution to this problem, its computational cost is exorbitant for larger data sets. Moreover, for dependent data, as considered in this paper, resampling procedures and methods based on estimation of the nuisance parameters prove labyrinthine, depending on further tuning parameters, such as the block length in a multiplier bootstrap (see Bücher and Kojadinovic, 2016) or a bandwidth in the long-run variance estimator (see Horváth and Kokoszka, 2012, Chapter 16), which are in turn difficult to adjust properly. We sidestep such concerns by virtue of self-normalization, which yields asymptotically pivotal statistics. For this purpose we introduce a sequential process of estimates of the covariance operator and study its weak convergence under an appropriate topology. Thirdly, in contrast to the related literature on non-parametric tests of separability, our weak convergence results are proved not for the space of Hilbert-Schmidt, but for the more restrictive space of trace-class operators, which does not enjoy a Hilbert structure. As a consequence establishing CLTs and invariance principles becomes technically more challenging. The benefit of this approach is, that it allows statistical analysis of all three separability measures (partial traces, partial products and optimal approximations) simultaneously, in a fully functional framework. One consequence of this analysis is the extension of existing work on partial traces, which so far has been confined to finite dimensional spaces and was hence reliant on preliminary projections of the observed functions. In contrast our methodology does not require any preprocessing and is thus easy to implement.

The remaining part of this paper is organized as follows. In Section 2 we recall some basic facts about Hilbert spaces and their tensor products and afterwards introduce measures for deviations from separability. Next we specify our framework in Section 3 and then state our main theoretical results. In particular we establish the weak convergence of a sequential version of the empirical covariance operator in the vector space of all trace-class-valued continuous functions. These results are used to prove the weak convergence of appropriately standardized estimates of the measures of deviation from separability with a distribution free limit. As applications we derive asymptotically valid confidence intervals for these measures and valid statistical tests for the hypothesis of *approximate separability*. Section 4 is devoted to the finite sample properties of the proposed methods, both in simulations and by virtue of a data example. Finally, all proofs and technical lemmas are gathered in the Appendix.

2 Deviations from separability in Hilbert spaces

In this section we develop the mathematical tools to analyze the separability of linear operators. In Section 2.1 we provide a minimal background on spectral theory, as well as tensor product Hilbert spaces. For more details we refer the interested reader to the monographs Weidmann (1980) and Gohberg et al. (1993, 2003). In Section 2.2, we introduce different measures to quantify separability. Separability is assessed by the norm of the difference between a covariance operator and its separable approximation.

2.1 Operators on Hilbert spaces

For $i = 1, 2$ let $(\mathcal{H}_i, \langle \cdot, \cdot \rangle_{\mathcal{H}_i})$ denote generic, real, separable Hilbert spaces, equipped with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_i}$. The space of bounded, linear operators mapping from \mathcal{H}_1 , to \mathcal{H}_2 , i.e. such operators T , which fulfill

$$\|T\|_{\mathcal{L}} := \sup_{\|x\|_{\mathcal{H}_1} \leq 1} \|T(x)\|_{\mathcal{H}_2} < \infty,$$

will be denoted throughout our discussion by $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$. Notice that all operators in this space are continuous.

An important subspace of $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ consists of all compact operators, i.e., all operators which map the unit ball of \mathcal{H}_1 to a compact set in \mathcal{H}_2 . Any such operator $T : \mathcal{H}_1 \rightarrow \mathcal{H}_2$ can be diagonalized as follows

$$T[x] = \sum_{i=1}^{\infty} \sigma_i \langle e_i, x \rangle_{\mathcal{H}_1} f_i, \quad x \in \mathcal{H}_1,$$

where $\{e_i\}_{i \in \mathbb{N}}$ is an orthonormal basis of \mathcal{H}_1 and $\{f_i\}_{i \in \mathbb{N}}$ of \mathcal{H}_2 . The *singular values* of T , $\{\sigma_i\}_{i \in \mathbb{N}}$, form a decreasing sequence of non-negative numbers, converging to 0. The decay rate of the singular values, which may be seen as a measure of T 's regularity, is captured by the *p-Schatten-norm*

$$\|T\|_p := \left(\sum_{i=1}^{\infty} \sigma_i^p \right)^{1/p}.$$

The corresponding Schatten class consists of all operators with finite norm

$$\mathcal{S}_p(\mathcal{H}_1, \mathcal{H}_2) := \{T \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2) : \|T\|_p < \infty\}.$$

It is well known that for any compact operator T , the inequality $\|T\|_p \leq \|T\|_q$ holds, for all $p > q \geq 1$. In particular this entails $\mathcal{S}_p(\mathcal{H}_1, \mathcal{H}_2) \supset \mathcal{S}_q(\mathcal{H}_1, \mathcal{H}_2)$.

In the following we will be particularly interested in the classes $\mathcal{S}_1(\mathcal{H}_1, \mathcal{H}_2)$, the so called *trace-class operators* and $\mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_2)$, the *Hilbert–Schmidt operators*. The name "trace-class" is derived from the concept of "trace", as known from square-matrices. It can be extended to an operator $T \in \mathcal{S}_1(\mathcal{H}, \mathcal{H})$ acting on a separable Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ as

$$\text{Tr}[T] := \sum_{i=1}^{\infty} \langle T[e_i], e_i \rangle_{\mathcal{H}},$$

where $\{e_i\}_{i \in \mathbb{N}}$, is an arbitrary orthonormal basis of \mathcal{H} . If the operator T is positive and symmetric, it holds that $\text{Tr}(T) = \|T\|_1$. More generally for an operator $T : \mathcal{H}_1 \rightarrow \mathcal{H}_2$ the trace may be used to express the trace-class and Hilbert–Schmidt norms as follows

$$\|T\|_1 = \text{Tr}[\sqrt{TT^*}] \quad \text{and} \quad \|T\|_2 = \sqrt{\text{Tr}[TT^*]}.$$

Here T^* denotes the adjoint operator of T and $\sqrt{TT^*}$ the canonical square root operator as defined for instance in Horváth and Kokoszka (2012). We now turn to tensor products of Hilbert spaces. For the Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$, we denote by $\mathcal{H}_1 \odot \mathcal{H}_2$ their algebraic tensor product, on which a unique bilinear form, denoted by $\langle \cdot, \cdot \rangle$, can be defined such that

$$\langle h_1 \otimes h_2, h'_1 \otimes h'_2 \rangle := \langle h_1, h'_1 \rangle_{\mathcal{H}_1} \langle h_2, h'_2 \rangle_{\mathcal{H}_2}$$

for any $h_1, h'_1 \in \mathcal{H}_1$ and $h_2, h'_2 \in \mathcal{H}_2$. Completing $\mathcal{H}_1 \odot \mathcal{H}_2$ with respect to the induced norm yields again a Hilbert space, referred to as the *tensor product Hilbert space*, which is denoted by $\mathcal{H}_1 \otimes \mathcal{H}_2$. For details on tensor product spaces we refer to Weidmann (1980).

The tensor product structure can be extended to the bounded, linear maps acting on $\mathcal{H}_1 \otimes \mathcal{H}_2$. Suppose that two operators $A \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_1)$ and $B \in \mathcal{L}(\mathcal{H}_2, \mathcal{H}_2)$ are given. Then we may define the map

$$A \otimes B : \mathcal{H}_1 \otimes \mathcal{H}_2 \rightarrow \mathcal{H}_1 \otimes \mathcal{H}_2$$

for all $h_1 \otimes h_2 \in \mathcal{H}_1 \otimes \mathcal{H}_2$ with $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ as

$$A \otimes B[h_1 \otimes h_2] := A[h_1] \otimes B[h_2].$$

Requiring $A \otimes B$ to be linear extends the map uniquely to the whole space $\mathcal{H}_1 \otimes \mathcal{H}_2$. Its norm $\|A \otimes B\|_{\mathcal{L}}$ is easily seen to equal $\|A\|_{\mathcal{L}} \|B\|_{\mathcal{L}}$. Operators in $\mathcal{L}(\mathcal{H}_1 \otimes \mathcal{H}_2, \mathcal{H}_1 \otimes \mathcal{H}_2)$, which can be decomposed as a tensor product of operators in $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_1)$ and $\mathcal{L}(\mathcal{H}_2, \mathcal{H}_2)$ are called *separable*.

2.2 Measures of separability

Separability is an attractive property in applications of high dimensional and functional data, as it cuts memory space and boosts computational speed. While plain separability hardly exist in practice, there are cases where an operator is on the verge of being separable. To make this notion quantitative, different measures of separability have been proposed, most of which inspect the normed difference of an operator and a separable approximation. Here we consider three ways to construct such approximations, which have gained particular attention in the literature, namely approximations by partial traces (Constantinou et al., 2017; Aston et al., 2017), by partial products (Bagchi and Dette, 2020) and by optimal approximations (Van Loan and Pitsianis, 1993; Genton, 2007).

2.2.1 Partial traces

The concept of *partial traces* plays an important role in quantum mechanics, where it is used to decompose so-called density matrices in their marginals while preserving the "consistency of expectation" of observables of the subsystem. For linear algebraic properties of partial traces, see for instance Bhatia (2003) and the references therein for the physical motivation. Its extension to separable Hilbert spaces can be constructed as follows. Let $\mathcal{H} := \mathcal{H}_1 \otimes \mathcal{H}_2$ be the tensor product Hilbert space of the separable Hilbert spaces \mathcal{H}_1 and \mathcal{H}_2 . We then define the partial traces

$$\mathrm{Tr}_1 : \mathcal{S}_1(\mathcal{H}, \mathcal{H}) \rightarrow \mathcal{S}_1(\mathcal{H}_2, \mathcal{H}_2) \quad \text{and} \quad \mathrm{Tr}_2 : \mathcal{S}_1(\mathcal{H}, \mathcal{H}) \rightarrow \mathcal{S}_1(\mathcal{H}_1, \mathcal{H}_1) \quad (2.1)$$

for any trace-class operators $A \in \mathcal{S}_1(\mathcal{H}_1, \mathcal{H}_1)$ and $B \in \mathcal{S}_1(\mathcal{H}_2, \mathcal{H}_2)$ as

$$\mathrm{Tr}_1[A \otimes B] = \mathrm{Tr}[A]B \quad \text{and} \quad \mathrm{Tr}_2[A \otimes B] = A\mathrm{Tr}[B].$$

Requiring the partial traces $\mathrm{Tr}_1, \mathrm{Tr}_2$, to be linear, yields unique extensions to the whole space. We make a brief remark on the partial traces, which highlights their usefulness in the study of separability (see Proposition C.1 in Aston et al. (2017), as well as Constantinou et al. (2018)).

Remark 2.1. *Suppose $C \in \mathcal{S}_1(\mathcal{H}, \mathcal{H})$ is a trace-class operator. Then it holds that:*

1. *If C is separable*

$$\mathrm{Tr}[C]C = \mathrm{Tr}_2[C] \otimes \mathrm{Tr}_1[C].$$

2. *The partial traces are Lipschitz-continuous with respect to the $\|\cdot\|_1$ -norm.*

Note that in the finite dimensional setting partial traces are also continuous w.r.t. the $\|\cdot\|_2$ -norm, since all norms are equivalent. This contrasts with the infinite dimensional setting. Indeed, consider a (separable) operator $C \otimes D$, where $C \in \mathcal{S}_1(\mathcal{H}_1, \mathcal{H}_1)$ and $D \in \mathcal{S}_1(\mathcal{H}_2, \mathcal{H}_2)$ are positive definite. It is easy to show that both equalities $\|C \otimes D\|_2 = \|C\|_2 \|D\|_2$ and $\|\mathrm{Tr}_1[C \otimes D]\|_2 = \mathrm{Tr}[C] \|D\|_2 = \|C\|_1 \|D\|_2$ hold. However, the trace-norm is in infinite dimensions not equivalent to the Hilbert–Schmidt norm. Hence the partial traces are not continuous with respect to the Hilbert–Schmidt norm. This seemingly minute technicality has important implications for the weak convergence results in Section 3.2 below, since the more restrictive space of trace-class operators is not a Hilbert space anymore. By virtue of the partial traces, we define the marginals of an operator C as

$$C^{(1)} := \mathrm{Tr}_2[C]/\mathrm{Tr}[C] \quad \text{and} \quad C^{(2)} := \mathrm{Tr}_1[C], \quad (2.2)$$

where we have to assume that $\mathrm{Tr}[C] \neq 0$. For a covariance operator this means that the corresponding data is not deterministic. Notice that if C is separable, Remark 2.1 guarantees the equality $C = C^{(1)} \otimes C^{(2)}$. If C is not separable, the marginal product

$$C^{\mathrm{tr}} := C^{(1)} \otimes C^{(2)} \quad (2.3)$$

may still be regarded as a separable approximation of C . Hence one can assess separability by considering the measure

$$M^{\mathrm{tr}}[C] := \|C - C^{\mathrm{tr}}\|_2^2. \quad (2.4)$$

Large values of $M^{\mathrm{tr}}[C]$ indicate that C^{tr} is a poor approximation of C and hence suggest only a small degree of separability.

2.2.2 Partial products

The definition of *partial products* was introduced (without this particular name) in the work of Bagchi and Dette (2020). Partial products provide an alternative to the partial traces in the study of separability. They are defined on the larger space of all Hilbert–Schmidt operators $\mathcal{S}_2(\mathcal{H}, \mathcal{H})$ instead of the trace-class operators $\mathcal{S}_1(\mathcal{H}, \mathcal{H})$. The partial products are bilinear maps

$$P_1 : \mathcal{S}_2(\mathcal{H}, \mathcal{H}) \times \mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_1) \rightarrow \mathcal{S}_2(\mathcal{H}_2, \mathcal{H}_2) \quad \text{and} \quad P_2 : \mathcal{S}_2(\mathcal{H}, \mathcal{H}) \times \mathcal{S}_2(\mathcal{H}_2, \mathcal{H}_2) \rightarrow \mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_1)$$

defined for a separable element $A \otimes B \in \mathcal{S}_2(\mathcal{H}, \mathcal{H})$ and $C_1 \in \mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_1)$, $C_2 \in \mathcal{S}_2(\mathcal{H}_2, \mathcal{H}_2)$ as

$$P_1(A \otimes B, C_1) = \langle A, C_1 \rangle_{\mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_1)} B \quad \text{and} \quad P_2(A \otimes B, C_2) = \langle B, C_2 \rangle_{\mathcal{S}_2(\mathcal{H}_2, \mathcal{H}_2)} A. \quad (2.5)$$

If A and B are trace-class, we can express the partial products in terms of the partial traces, e.g., when $B \in \mathcal{S}_1(\mathcal{H}_2, \mathcal{H}_2)$, then denoting the adjoint of an operator A by A^* , the identity

$$P_1(A \otimes B, C_1) = \text{Tr}_1[(A^* C_1) \otimes B]$$

holds. It was shown in Bagchi and Dette (2020) that the operators P_1 and P_2 are well-defined, bi-linear and continuous with

$$\begin{aligned} \langle B, P_1(C, C_1) \rangle_{\mathcal{S}_2(\mathcal{H}_2, \mathcal{H}_2)} &= \langle C, C_1 \otimes B \rangle_{\mathcal{S}_2(\mathcal{H}, \mathcal{H})} \\ \langle A, P_2(C, C_2) \rangle_{\mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_1)} &= \langle C, A \otimes C_2 \rangle_{\mathcal{S}_2(\mathcal{H}, \mathcal{H})}. \end{aligned}$$

In principle the partial products allow the decomposition of a separable operator C as follows

$$C = \frac{P_2(C, \Delta_2) \otimes P_1(C, \Delta_1)}{\langle C, \Delta_1 \otimes \Delta_2 \rangle_{\mathcal{S}_2(\mathcal{H}, \mathcal{H})}},$$

for any $\Delta_i \in \mathcal{S}_2(\mathcal{H}_i, \mathcal{H}_i)$ ($i = 1, 2$), as long as the denominator is not degenerate. When C is not separable the right side may be regarded as a separable approximation of C . Its Hilbert–Schmidt distance to C is minimized for a fixed Δ_2 by $\Delta_1 = P_1(C, P_2(C, \Delta_2)) / \|P_2(C, \Delta_2)\|_2^2$. Thus Bagchi and Dette (2020) suggest the following measure of separability

$$M^{\text{prod}}[C] := M^{\text{prod}}[C, \Delta_2] := \|C - C^{\text{prod}}\|_2^2, \quad (2.6)$$

where

$$C^{\text{prod}} := \frac{P_2(C, \Delta_2) \otimes P_1(C, P_2(C, \Delta_2))}{\|P_2(C, \Delta_2)\|_2^2}. \quad (2.7)$$

Obviously the quantification of C 's separability still depends on the specific choice of Δ_2 , which is an infinite dimensional parameter. A natural, ad hoc solution could be to choose Δ_2 as a finite rank approximation of the identity. Then $P_2(C, \Delta_2) \approx \text{Tr}_2[C]$, and the corresponding separability measure would be similar to that in (2.4).

2.2.3 Optimal approximations

The evaluation of the above separability measures has important implications for users: A small measure suggests, that C is well approximated by a separable operator, be it constructed via partial traces or partial products. However the interpretation for larger values is unsatisfactory. Is the covariance operator

C indeed far away from separability, or is it just compared to an unsuitable approximation? To address this question it is natural to consider the *optimal, separable approximations* of C , that is an operator

$$C^{\text{opt}} \in \operatorname{argmin} \left\{ \|C - \tilde{C}\|_2^2 : \tilde{C} = A \otimes B, A \in \mathcal{S}_2(\mathcal{H}_1, \mathcal{H}_1), B \in \mathcal{S}_2(\mathcal{H}_2, \mathcal{H}_2) \right\}.$$

In finite dimensions the existence of such approximations has been established already in the last century, see e.g. Van Loan and Pitsianis (1993). These results were employed to study space-time data by Genton (2007). The existence proof of optimal, separable approximations can be easily extended to arbitrary, separable Hilbert spaces. Notice that we may identify $\mathcal{S}_2(\mathcal{H}, \mathcal{H}) \cong \mathcal{H} \otimes \mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2$ and $\mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_1, \mathcal{H}_2 \otimes \mathcal{H}_2) \cong \mathcal{H}_1 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_2$. There exists a canonical, isometric isomorphism of Hilbert spaces

$$\Pi : \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2 \rightarrow \mathcal{H}_1 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_2,$$

which is uniquely determined by the images of simple tensors, namely

$$h_1 \otimes h_2 \otimes h'_1 \otimes h'_2 \mapsto h_1 \otimes h'_1 \otimes h_2 \otimes h'_2,$$

where $h_1, h'_1 \in \mathcal{H}_1$ and $h_2, h'_2 \in \mathcal{H}_2$. For any separable map $A \otimes B \in \mathcal{S}_2(\mathcal{H}, \mathcal{H})$, its image $\Pi[A \otimes B]$ is a rank one operator in $\mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_1, \mathcal{H}_2 \otimes \mathcal{H}_2)$. Conversely the preimage of any rank one operator in $\mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_1, \mathcal{H}_2 \otimes \mathcal{H}_2)$ under Π is separable. As Π is in particular norm preserving, to find an optimal, separable approximation C^{opt} of C is equivalent to determine the optimal rank-one approximation of $\Pi[C]$. We accomplish this by calculating the largest singular value γ_1 of $\Pi[C]$ and corresponding vectors $e_1 \in \mathcal{H}_1 \otimes \mathcal{H}_1$ and $f_1 \in \mathcal{H}_2 \otimes \mathcal{H}_2$. Then we define the separable approximation

$$C^{\text{opt}} := \Pi^{-1}[\gamma_1 \cdot e_1 \otimes f_1], \quad (2.8)$$

which equals C in case of separability. The corresponding distance measure takes a particularly simple form

$$M^{\text{opt}}[C] := \|C - C^{\text{opt}}\|_2^2 = \|\Pi[C] - \Pi[C^{\text{opt}}]\|_2^2 = \|C\|_2^2 - \gamma_1^2. \quad (2.9)$$

This measure is more easily interpretable than those constructed by partial traces and partial products, because it informs the user precisely about the possibilities and limits of separable approximations. Notice that in general multiple, optimal approximations may exist. However if $\gamma_1 > \gamma_2$ holds, which as we will see is a very mild condition, the optimal approximation is already unique.

Remark 2.2. We briefly comment on the favorable computational aspects of optimal approximations, in the context of covariance operators. In the subsequent statistical applications we estimate the separability of an unknown covariance operator C , by considering the measure $M^{\text{opt}}[\hat{C}_n]$, where \hat{C}_n is the empirical covariance operator, defined in (3.2) below. Calculating $M^{\text{opt}}[\hat{C}_n]$, i.e., the Frobenius norm and the largest singular value of $\Pi[\hat{C}_n]$ can be efficiently implemented (Golub and Van Loan, 2013; Cardot and Degras, 2018). Moreover it is possible to numerically determine these entities without saving the whole operator $\Pi[\hat{C}_n]$, but only the observed data (Tzeng, 2013). Such storage concerns are material in some applications, as indicated in the introduction. We also note that often one of the spaces, say \mathcal{H}_1 , is low dimensional (for details see Remark 3.1). Under such circumstances $\hat{\gamma}_1$ is the largest eigenvalue of the low-dimensional matrix $\Pi[\hat{C}_n]^* \Pi[\hat{C}_n]$, which can be easily calculated.

Example 2.3. We conclude this section by considering a small example to illustrate similarities and differences of the three separability measures in the finite dimensional case. Notice therefore that $\mathbb{R}^4 \cong \mathbb{R}^2 \otimes \mathbb{R}^2$. We consider the matrix

$$C(q) := \begin{pmatrix} 2 & 0 & 1 & q \\ 0 & 2 & q & 1 \\ 1 & q & 2 & q \\ q & 1 & q & 2 \end{pmatrix} \in \mathcal{S}_2(\mathbb{R}^2 \otimes \mathbb{R}^2, \mathbb{R}^2 \otimes \mathbb{R}^2)$$

where $q \in [0, 1]$ is a separability parameter.

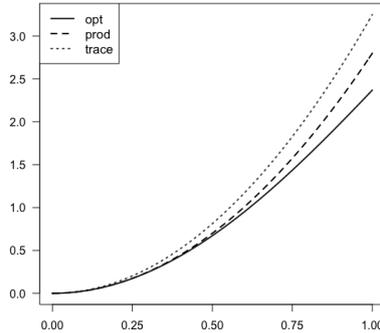


Figure 1: Comparison of the different measures $M^{\text{tr}}, M^{\text{prod}}, M^{\text{opt}}$ in a matrix example. For the partial products we have chosen Δ_2 as the two dimensional identity matrix.

The exact values of the measures (2.4), (2.6) and (2.9) are calculated analytically in Appendix A.1 and given by

$$\begin{aligned} M^{\text{tr}}[C(q)] &= \|C(q) - C(q)^{\text{tr}}\|_2^2 = 13/4 \cdot q^2 \\ M^{\text{prod}}[C(q)] &= \|C(q) - C(q)^{\text{prod}}\|_2^2 = 70/25 \cdot q^2 \\ M^{\text{opt}}[C(q)] &= \|C(q)\|_2^2 - \gamma_1^2 = 10 + 3q^2 - \sqrt{9q^4 + 4q^2 + 100} \end{aligned}$$

For $q = 0$, the matrix is separable and for larger q the matrix gets increasingly inseparable. In Figure 1 we observe a monotonic increase of all three measures in q . Even for this simple example the quality of the approximations $C^{\text{tr}}, C^{\text{prod}}, C^{\text{opt}}$ varies considerably. In particular $M^{\text{opt}}[C(1)]$ is only about 3/4 of $M^{\text{tr}}[C(1)]$ and for larger matrices this effect is generally much enhanced (see Section 4).

Note that the inequality $M^{\text{prod}}[C(q)] \leq M^{\text{tr}}[C(q)]$ holds a priori for all q , since C^{tr} is the product of the marginals $C^{(1)} \otimes C^{(2)}$ as defined in (2.2), while M^{prod} is the product of $C^{(1)}$ with an optimally adapted second factor.

3 Statistical Inference

In this section we proceed to the examination of separability in the covariance structure of space-time data. First we describe the statistical model considered in this paper and subsequently introduce the empirical

covariance operator \hat{C}_n , as the natural, non-parametric estimator of the data's covariance operator C (see equation (3.3) for a precise definition). Secondly, we derive an invariance principle for the operator \hat{C}_n , which is of interest in its own right. Since all measures $M^x[C]$ ($x \in \{\text{tr}, \text{prod}, \text{opt}\}$) introduced in Section 2 are functionals of C , corresponding estimates are obtained replacing C by \hat{C}_n . Stochastic linearizations of the estimated measures $M^x[\hat{C}_n]$ can be employed to prove their weak convergence to normal distributions. However, the asymptotic variances depend in an intricate way on the unknown covariance operator C and higher order moments of the underlying model. To address this problem we use the concept of self-normalization to introduce standardized estimates of $M^x[C]$, that are asymptotically pivotal. These results have profound statistical consequences, which are explored in Section 3.3 and 3.4.

3.1 Space-Time Data

Let μ_1, μ_2 denote finite measures on the interval $[0, 1]$ and define $H_1 := L^2_{\mu_1}[0, 1]$ and $H_2 := L^2_{\mu_2}[0, 1]$ as Hilbert spaces of square integrable functions with respect to the measure μ_1 and μ_2 , respectively. The natural inner product on the space H_ℓ is defined as

$$\langle f_1, f_2 \rangle_\ell := \int_0^1 f_1(t)f_2(t)\mu_\ell(dt) \quad f_1, f_2 \in H_\ell, \quad \ell = 1, 2.$$

For the ease of reading we will subsequently dispense with the indices of the inner products and norms, whenever there is no ambiguity. In the above situation, the tensor product Hilbert space $H := H_1 \otimes H_2$ of spatio-temporal functions consists of all square integrable functions with respect to the product measure $\mu_1 \otimes \mu_2$, i.e.

$$H = \left\{ h : [0, 1]^2 \rightarrow \mathbb{R} \mid \int_{[0, 1]^2} h^2(s, t)(\mu_1 \otimes \mu_2)(ds, dt) < \infty \right\}. \quad (3.1)$$

The corresponding inner product and norm on H are given by

$$\langle h_1, h_2 \rangle = \int_{[0, 1]^2} h_1(s, t)h_2(s, t)(\mu_1 \otimes \mu_2)(ds, dt), \quad h_1, h_2 \in H$$

and $\|h\| = (\langle h, h \rangle)^{1/2}$, respectively.

Remark 3.1. In several applications of spatio-temporal statistics such as meteorology there exists a considerable asymmetry in the difficulty of acquiring spatial and temporal data. Whereas raising large numbers of densely timed observations at a certain location, say a weather station, is quite reasonable, establishing additional measuring stations across the country is often prohibitively expensive. In this case one can use statistical models, which treat temporal data as time-continuous functions (here continuity is achieved by interpolations between the densely placed timepoints) while the spatial locations are modeled by a discrete parameter.

This situation corresponds to the choice of the Lebesgue measure for the measure μ_2 and a discrete measure supported on S points, say $\{p_1, \dots, p_S\}$ for the measure μ_1 in (3.1). The inner product on the set H_1 is then given by

$$\langle g_1, g_2 \rangle = \sum_{s=1}^S g_1(p_s)g_2(p_s) \quad g_1, g_2 \in H_1,$$

and the tensor product Hilbert space H in (3.1) is given by

$$H = \left\{ h : \{p_1, \dots, p_S\} \times [0, 1] \rightarrow \mathbb{R} \mid \sum_{s=1}^S \int_0^1 h^2(p_s, t) dt < \infty \right\}.$$

Suppose that we observe a stretch X_1, X_2, \dots, X_n of random functions from a stationary, H -valued time series $(X_i)_{i \in \mathbb{Z}}$, satisfying $\mathbb{E}\|X_1\|^2 < \infty$. For the sake of a simple presentation we assume that the random functions are centered, i.e.

$$\mathbb{E}X_i(s, t) = 0, \quad \forall s, t \in [0, 1]^2.$$

In practical applications this assumption is often unrealistic and it can be removed by empirically centering all the collected data without changing any of the following results (see Remark 3.13 below for more details). The second order structure of the random functions X_i is captured by its covariance operator $C := \mathbb{E}X_1 \otimes X_1 \in \mathcal{S}_1(H, H)$ which exists under the stated assumptions and is defined pointwise by

$$C(h) := \mathbb{E}X_1 \langle X_1, h \rangle \quad h \in H.$$

Based on the sample X_1, \dots, X_n the natural, non-parametric estimator of the covariance operator C is given by

$$\hat{C}_n := \frac{1}{n} \sum_{i=1}^n \{X_i \otimes X_i\}. \quad (3.2)$$

By virtue of \hat{C}_n it is possible to estimate the separability of the true covariance operator C , as quantified by any of the measures M^x , for $x \in \{\text{tr}, \text{prod}, \text{opt}\}$, simply by the plug-in $M^x[\hat{C}_n]$. In the following section we develop the analytic tools to examine the asymptotic behavior of these statistics.

3.2 Weak convergence in Banach spaces

We now devise the analytic tools for the investigation of the statistics $M^x[\hat{C}_n]$. First we prove an invariance principle for a sequential version of the estimator \hat{C}_n , which is key for statistical self-normalization. We prove weak convergence on the space of trace-class-valued, continuous functions, i.e., sequential operators $L(\lambda)$, which live for some fixed λ in $\mathcal{S}_1(H, H)$.

This approach contrasts with the body of existing literature, where weak convergence on the space of Hilbert–Schmidt operators $\mathcal{S}_2(H, H)$ is considered, see for instance the references in Mas (2006) and more recently, Constantinou et al. (2018); Bagchi and Dette (2020). While central limit theorems as well as invariance principles for this Hilbert space can be derived (Hörmann and Kokoszka, 2010; Berkes et al., 2013), this approach is ineffective for the study of M^{tr} , which has recently attracted much attention. The reason is, that M^{tr} is not well defined for all Hilbert–Schmidt operators (for details see Remark 3.6). This problem can be sidestepped by projecting on a finite dimensional space, where all matrix norms are equivalent (see, for instance Aston et al., 2017, for this approach). Of course this requires the determination of an adequate projection space. For a fully functional approach we have to consider the covariance operator as a random element in $\mathcal{S}_1(H, H)$, which equipped with $\|\cdot\|_1$ is a (separable) Banach space (see for instance Dunford and Schwartz, 1958). As pointed out by Mas (2006) this space might even be considered more natural for the study of weak convergence of empirical covariance operators.

In order to specify the dependence structure of our data, we introduce the notion of ϕ -mixing, as described, for example in Samur (1984). To be precise for a strictly stationary sequence $(Y_j)_{j \in \mathbb{Z}}$ of random variables we define

$$\mathcal{F}_{h,k} := \sigma(Y_h, \dots, Y_k),$$

as the σ -algebra, generated by Y_h, \dots, Y_k ($k \geq h$) and consider the ϕ -mixing dependence coefficients

$$\phi(k) := \sup_{h \in \mathbb{Z}} \sup \{ \mathbb{P}(F|E) - \mathbb{P}(F) : E \in \mathcal{F}_{1,h}, F \in \mathcal{F}_{h+k,\infty}, \mathbb{P}(E) > 0 \}.$$

The sequence $(Y_j)_{j \in \mathbb{Z}}$ is called ϕ -mixing, if $\lim_{k \rightarrow \infty} \phi(k) = 0$. In order to prove an invariance principle for a sequence of random elements in the space H defined in (3.1) we make the following assumptions.

Assumption 3.2.

- (1) $(X_i)_{i \in \mathbb{Z}}$ is a strictly stationary time series of centered, random functions in the space H defined in (3.1).
- (2) The sequence $(X_i)_{i \in \mathbb{Z}}$ is ϕ -mixing, with $\phi(1) < 1$ and such that $\sum_{k=1}^{\infty} \sqrt{\phi(k)} < \infty$.
- (3) There exists a basis $(e_q)_{q \in \mathbb{N}}$ of H , such that $\sum_{q=1}^{\infty} \sqrt[4]{\mathbb{E}|\langle X_1, e_q \rangle|^4} < \infty$.
- (4) The optimal separable approximation C^{opt} of C is unique.

Notice that the third point in particular implies the existence of fourth moments of X_1 , which follows by a simple calculation. The fourth point is much less restrictive than it might appear. In fact it is satisfied by a vast majority of geostatistical models considered in the literature. It is equivalent to assuming the strict inequality $\gamma_1 > \gamma_2$ for the two largest singular values of the restacked operator $\Pi[C]$ (see Section 2.2.3). A simple, sufficient condition, that already entails it, is the basic approximation requirement

$$\|C - C^{\text{opt}}\|_2^2 < \frac{\|C\|_2^2}{2}.$$

We now introduce a sequential estimator of the covariance operator C

$$\hat{C}_n(\lambda) := \frac{1}{n} \sum_{i=1}^{\lfloor n\lambda \rfloor} (X_i \otimes X_i) + \frac{n\lambda + 1 - \lfloor n\lambda \rfloor}{n} (X_{\lfloor n\lambda \rfloor} \otimes X_{\lfloor n\lambda \rfloor}), \quad (3.3)$$

where $\lceil x \rceil = \min\{z \in \mathbb{Z} : z \geq x\}$ and $\lambda \in [0, 1]$. Note that λ determines the proportion of data used for estimation and that the estimate in (3.2) is obtained as $\hat{C}_n = \hat{C}_n(1)$. The linear interpolation term on the right side of (3.3) ensures that $\hat{C}_n(\lambda)$ is continuous with respect to λ , which is important for technical reasons (details are given in Appendix A.2). A simple calculation shows that \hat{C}_n is a trace-class operator and by the discussion in Section 2.1 a fortiori Hilbert–Schmidt. In Theorem 3.3 below, we provide an invariance principle, which lies at the heart of the subsequent self-normalization approach.

To state it in detail, we first recall the concept of a Gaussian measure and its associated Brownian motion on a Banach space. Let $(F, \|\cdot\|_F)$ be a (separable) Banach space. Then, a measure μ_G on the Borel sets of F is said to be centred Gaussian, if and only if $\mu_G \circ g^{-1}$ is a centred Gaussian measure on \mathbb{R} for all $g \in F^*$, the topological dual of F . Now, the vector space of F -valued, continuous functions denoted by

$$\mathcal{C}([0, 1], F) := \{f : [0, 1] \rightarrow F \mid f \text{ continuous}\},$$

and equipped with the maximum norm

$$\|f\|_\infty := \sup_{\lambda \in [0,1]} \|f(\lambda)\|_F, \quad (3.4)$$

is again a (separable) Banach space. Let μ_G be a centred Gaussian measure on F and let G be the associated F -valued random variable. The *Brownian motion* \mathbb{B}_G on F corresponding to G is defined as the random function taking values in $\mathcal{C}([0,1], F)$ with the following properties:

- (a) $\mathbb{B}_G(\lambda)$ is distributed as $\sqrt{\lambda}G$ for any $\lambda \in [0,1]$.
- (b) The increments of \mathbb{B}_G are stationary, centred and independent.

An existence proof of the Brownian motion induced by a Gaussian measure on a (separable) Banach space can be found in Gross (1970).

Theorem 3.3. *Let $(X_i)_{i \in \mathbb{Z}}$ be a sequence of random functions in H satisfying Assumption 3.2. Then there exists a Gaussian random operator $G \in \mathcal{S}_1(H, H)$ such that*

$$\{\sqrt{n}(\hat{C}_n(\lambda) - C\lambda)\}_{\lambda \in [0,1]} \xrightarrow{d} \{\mathbb{B}_G(\lambda)\}_{\lambda \in [0,1]}$$

in $\mathcal{C}([0,1], \mathcal{S}_1(H, H))$. Here \mathbb{B}_G is the Brownian motion on $\mathcal{S}_1(H, H)$ corresponding to G .

The proof of Theorem 3.3 is deferred to the Appendix. In the subsequent theorem we examine the Fréchet differentiability of the difference of a trace-class operator and its separable approximation.

Theorem 3.4. *Let $I \subset (0,1]$ be a closed interval and suppose that the covariance operator C satisfies Assumption 3.2, (4). Then the maps*

- i) $\mathbf{F}^{\text{tr}} : \begin{cases} \mathcal{C}(I, \mathcal{S}_1(H, H)) \rightarrow \mathcal{C}(I, \mathcal{S}_1(H, H)) \\ L(\cdot) \mapsto L(\cdot) - L^{\text{tr}}(\cdot), \end{cases}$
- ii) $\mathbf{F}^{\text{prod}} : \begin{cases} \mathcal{C}(I, \mathcal{S}_2(H, H)) \rightarrow \mathcal{C}(I, \mathcal{S}_2(H, H)) \\ L(\cdot) \mapsto L(\cdot) - L^{\text{prod}}(\cdot), \end{cases}$
- iii) $\mathbf{F}^{\text{opt}} : \begin{cases} \mathcal{C}(I, \mathcal{S}_2(H, H)) \rightarrow \mathcal{C}(I, \mathcal{S}_2(H, H)) \\ L(\cdot) \mapsto L(\cdot) - L^{\text{opt}}(\cdot), \end{cases}$

are Fréchet differentiable in the function $\lambda \mapsto \lambda C$, where for fixed $\lambda \in I$ the operators $L^{\text{tr}}(\lambda)$, $L^{\text{prod}}(\lambda)$ and $L^{\text{opt}}(\lambda)$ are defined in (2.3), (2.7) and (2.8), respectively. The derivatives are given in Section A.3 in the Appendix.

Theorem 3.4 implies that any of the three approximation-maps presented in Section 2.2 is Fréchet-differentiable. Since the Hilbert–Schmidt norm is also differentiable, the functional- Δ -method yields the following, weak convergence result for the estimated separability measures.

Corollary 3.5. *Under the assumptions of Theorem 3.3, it holds that for any $x \in \{\text{tr}, \text{prod}, \text{opt}\}$*

$$\left\{ \sqrt{n}(M^x[\hat{C}_n(\lambda)] - M^x[\lambda C]) \right\}_{\lambda \in I} \xrightarrow{d} \{\mathbb{B}(\lambda)\lambda\sigma^x\}_{\lambda \in I},$$

where \mathbb{B} is a standard Brownian motion on the interval $[0,1]$. The definition of the standard deviation σ^x is given in (A.11).

Remark 3.6. As mentioned in the preamble of this section, the statistic $\hat{C}_n - \hat{C}_n^{\text{tr}}$ is popular in the study of separability. It has been investigated by Constantinou et al. (2018) in the model described in Remark 3.1. In their Theorem 2 it is indicated that the limit of the operators $\sqrt{n}(\hat{C}_n - \hat{C}_n^{\text{tr}})$ is Gaussian in the space $\mathcal{S}_2(H, H)$. Using their theorem, one could - in principle - derive the weak convergence of the corresponding measure $M^{\text{tr}}[\hat{C}_n]$ via a continuous mapping argument, but the proof of this result is not complete. More precisely, the partial traces are only continuous with respect to the Hilbert–Schmidt topology in finite dimensions, see the discussion following Remark 2.1. However, in the functional setting continuity in an infinite dimensional space is required and crucial for an application of the functional- Δ -method. Thus, strictly speaking Theorem 2 in Constantinou et al. (2018) is only correct for finite dimensional spaces. However, Theorem 3.3 and 3.4 and their proofs in Section A.2 and A.3 provide a solution to this problem. They establish weak convergence of $\sqrt{n}(\hat{C}_n - \hat{C}_n^{\text{tr}})$ in $\mathcal{S}_1(H, H)$ as well as Fréchet differentiability of the map $L \mapsto L - L^{\text{tr}}$ in $\mathcal{S}_1(H, H)$.

Remark 3.7. There exists a large amount of literature on CLTs (often as a consequence of strong approximations) for Banach space valued random variables under mixing conditions (see, for example, Kuelbs and Philip, 1980; Dehling and Philip, 1982; Dehling, 1983)). Strong approximations imply weak invariance principles, i.e., weak convergence of the process $\{\sqrt{n}(\hat{C}_n(\lambda) - C\lambda)\}_{\lambda \in [0,1]}$. This holds in particular if $(X_j)_{j \in \mathbb{Z}}$ is an absolute regular sequence with mixing coefficients $\beta(k)$ satisfying similar conditions as stated in Assumption 3.2.

Another popular dependency concept is that of L^p - m -approximability, which extends m -dependence. It is frequently used for Hilbert space valued time series (Hörmann and Kokoszka, 2010; Berkes et al., 2013) and is therefore potentially suited to the study of the separability measures M^{prod} or M^{opt} , which are defined on the Hilbert space $\mathcal{S}_2(H, H)$. However the measure M^{tr} requires a Banach space setting, which motivates the consideration of ϕ -mixing in Assumption 3.2. We expect that a CLT and perhaps an invariance principle for $\{\sqrt{n}(\hat{C}_n(\lambda) - C\lambda)\}_{\lambda \in [0,1]}$ can be derived in the case of L^p - m -approximability on $\mathcal{S}_1(H, H)$.

3.3 Pivotal measures of the deviation from separability

In this section we assess the separability of the data’s covariance structure, by plug-in estimates of the separability measures $M^x[C]$ ($x \in \{\text{tr}, \text{prod}, \text{opt}\}$) introduced in Section 2.2. Note that Corollary 3.5 implies that the statistic $M_n^x[\hat{C}_n] - M^x[C]$ satisfies a central limit theorem

$$\sqrt{n}(M_n^x[\hat{C}_n] - M^x[C]) \xrightarrow{d} \mathcal{N}(0, (\sigma^x)^2), \quad (3.5)$$

where the variance $(\sigma^x)^2$ is defined in equation (A.11) of the Appendix. The normal distribution in (3.5) is non-degenerate if $\sigma^x > 0$, which in particular requires that C is inseparable. Otherwise, an application of Theorem 3.3 and the continuous mapping theorem, show that

$$nM_n^x[\hat{C}_n] \xrightarrow{d} \|G\|_2^2,$$

where G is Gaussian random variable in $\mathcal{S}_1(H, H)$. Note that $\|G\|_2^2 \stackrel{d}{=} \sum_{j=1}^{\infty} \eta_j Z_j^2$, where Z_1, Z_2, \dots , are independent, standard normal random variables and η_1, η_2, \dots , eigenvalues of a certain covariance operator Q . However, the variance $(\sigma^x)^2$ and the values η_1, η_2, \dots , depend in an intricate way on the dependence structure of the time series $(X_j)_{j \in \mathbb{Z}}$ and higher order moments of X_j . As these quantities are intensely

difficult to estimate, the development of statistical inference for measures of separability, which is directly based on the asymptotic normality of $M^x[\hat{C}_n]$ is an extremely hard task.

To avoid the problem of estimating nuisance parameters we develop a self-normalized statistic, which has an easily accessible, asymptotic pivotal distribution. In particular its quantiles can be readily simulated. For this purpose we define for $x \in \{\text{tr}, \text{prod}, \text{opt}\}$ the self-normalization term

$$\hat{V}_n^x := \left\{ \int_I (M^x[\hat{C}_n(\lambda)] - \lambda^2 M^x[\hat{C}_n])^2 d\nu(\lambda) \right\}^{1/2}, \quad (3.6)$$

where ν is a probability measure on a closed interval $I \subset (0, 1]$. This particular type of self-normalization was also considered by Dette et al. (2020) in the context of change point detection in the mean function of a functional time series. The following result shows that the self-normalized statistic converges to a pivot.

Theorem 3.8. *Let $x \in \{\text{tr}, \text{prod}, \text{opt}\}$. Suppose that Assumption 3.2 holds and that the variance $(\sigma^x)^2$ defined in (A.11) is positive. Then the following, weak convergence holds*

$$\frac{M^x[\hat{C}_n] - M^x[C]}{\hat{V}_n^x} \xrightarrow{d} W := \frac{\mathbb{B}(1)}{\left\{ \int_I \lambda^2 (\mathbb{B}(\lambda) - \lambda \mathbb{B}(1))^2 d\nu(\lambda) \right\}^{1/2}}, \quad (3.7)$$

where \mathbb{B} denotes a Brownian motion on the interval $[0, 1]$.

The quantiles of the limiting distribution in Theorem 3.8 can easily be simulated. Exemplarily, we show in Table 1 the quantiles of the distribution of W if the measure ν is a uniform distribution on the sets $\{1/20, \dots, 19/20\}$ and $\{1/30, \dots, 29/30\}$. We also note that the distribution is symmetric, which implies that $q_\alpha = -q_{1-\alpha}$.

K	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
20	16.479	9.895	7.097
30	16.248	9.925	7.149

Table 1: $(1 - \alpha)$ -quantiles of the distribution of the random variable W defined in (3.7), where ν is the uniform measure on the discrete set $\{l/K : l = 1, \dots, K - 1\}$, for $K = 20, 30$.

Remark 3.9.

- (1) Theorem 3.8 allows statistical inference for the important separability measure $M^{\text{opt}}[C]$, which due to its implicit definition has not been conducted previously. In particular we can derive statistical tests and confidence intervals for the approximations presented in Genton (2007).
- (2) Other normalizing factors could be used as well. For example, defining

$$\tilde{V}_n^x := \int_I |M^x[\hat{C}_n(\lambda)] - \lambda^2 M^x[\hat{C}_n]| d\nu(\lambda)$$

we obtain (if $\sigma^x > 0$) the weak convergence

$$\frac{M^x[\hat{C}_n] - M^x[C]}{\tilde{V}_n^x} \xrightarrow{d} \frac{\mathbb{B}(1)}{\int_I \lambda |\mathbb{B}(\lambda) - \lambda \mathbb{B}(1)| d\nu(\lambda)}.$$

The separability measures considered so far quantify how good an operator C can be approximated by a separable version in absolute terms. However these entities are only interpretable if the user has some preconceived idea of how large $\|C\|_2^2$ is. If this knowledge is not readily available, the most natural approach seems to be a comparison of $M^x[\hat{C}_n]$ with the estimated Hilbert–Schmidt norm, i.e., $\|\hat{C}_n\|_2^2$. Accordingly we consider the relative separability measures

$$M_{\text{rel}}^x[C] := \frac{M^x[C]}{\|C\|_2^2},$$

its corresponding estimates $M_{\text{rel}}^x[\hat{C}_n]$ and the normalization factors

$$\hat{V}_{\text{rel},n}^x := \left\{ \int_I (\lambda^2 M_{\text{rel}}^x[\hat{C}_n(\lambda)] - \lambda^2 M_{\text{rel}}^x[\hat{C}_n])^2 d\nu(\lambda) \right\}^{1/2}.$$

The measure $M_{\text{rel}}^{\text{opt}}$ has been proposed by Genton (2007) for space-time matrices, who called it "separability approximation error index". It is easy to see that $M_{\text{rel}}^{\text{opt}}[C] \in [0, 1)$. We can now prove an analogue to Theorem 3.8 for the relative separability measures.

Theorem 3.10. *Let $x \in \{\text{tr}, \text{prod}, \text{opt}\}$. Under the Assumptions of Theorem 3.8 the following weak convergence holds*

$$\frac{M_{\text{rel}}^x[\hat{C}_n] - M_{\text{rel}}^x[C]}{\hat{V}_{\text{rel},n}^x} \xrightarrow{d} W, \quad (3.8)$$

where the random variable W is defined in (3.7).

3.4 Statistical consequences

In this section we discuss some statistical applications of the above, weak convergence results. We begin with the construction of confidence regions for the absolute and relative deviation measures. Throughout this paper we denote by q_α the α -quantile of the distribution of the random variable W defined in (3.7).

Theorem 3.11. *Let $x \in \{\text{tr}, \text{prod}, \text{opt}\}$. If Assumption 3.2 is satisfied and $\sigma^x > 0$, which is defined in (A.11), then the intervals*

$$\hat{I}_n^x = \left[M^x[\hat{C}_n] + q_{\alpha/2} \hat{V}_n^x, M^x[\hat{C}_n] + q_{1-\alpha/2} \hat{V}_n^x \right] \quad (3.9)$$

$$\hat{I}_{\text{rel},n}^x = \left[M_{\text{rel}}^x[\hat{C}_n] + q_{\alpha/2} \hat{V}_{\text{rel},n}^x, M_{\text{rel}}^x[\hat{C}_n] + q_{1-\alpha/2} \hat{V}_{\text{rel},n}^x \right] \quad (3.10)$$

are asymptotic $(1 - \alpha)$ -confidence intervals for the measures $M^x[C], M_{\text{rel}}^x[C]$, respectively.

By the duality of confidence intervals and tests the result of Theorem 3.11 can be employed to construct separability tests. In contrast to the currently available literature (see Fuentes, 2006; Genton, 2007; Constantinou et al., 2017; Aston et al., 2017; Constantinou et al., 2018; Bagchi and Dette, 2020, among others) we do not examine the problem of testing for exact separability, because as discussed in our introduction, separability is rarely met in applications. Instead we want to determine whether a separable

approximation C^x deviates from the true operator C substantially or not. Therefore we propose to test the hypotheses of a *relevant (relative) deviation from separability*, that is

$$H_0 : M^x[C] = \|C - C^x\|_2^2 \leq \Delta \quad \text{vs.} \quad H_1 : M^x[C] > \Delta \quad (3.11)$$

$$H_0 : M_{\text{rel}}^x[C] = \frac{\|C - C^x\|_2^2}{\|C\|_2^2} \leq \Delta \quad \text{vs.} \quad H_1 : M_{\text{rel}}^x[C] > \Delta_{\text{rel}} \quad (3.12)$$

where $\Delta, \Delta_{\text{rel}} > 0$ are given thresholds of relevance expressing the greatest deviation from separability, which is still considered as irrelevant in the application at hand. We propose to reject the null hypothesis in (3.11), whenever

$$q_\alpha \hat{V}_n^x + M^x[\hat{C}_n] > \Delta, \quad (3.13)$$

where M^x is defined in Section 2.1, the normalizing factor \hat{V}_n^x in (3.6) and q_α is the α -quantile of the distribution of the random variable W defined in (3.7). Similarly, the null hypothesis in (3.12) is rejected, whenever

$$q_\alpha \hat{V}_{\text{rel},n}^x + M_{\text{rel}}^x[\hat{C}_n] > \Delta. \quad (3.14)$$

These tests are motivated by the duality between confidence intervals and tests as, for example, described in Aitchison (1964). To be precise, note that it follows by similar arguments as given in the proof of Theorem 3.11 that the interval

$$\tilde{I}_n^x = [q_\alpha \hat{V}_n^x + M^x[\hat{C}_n], \infty)$$

defines a $(1 - \alpha)$ one-sided confidence interval for the measure $M^x[C]$. Now, by the duality between confidence interval and tests, an asymptotic level α -test for the hypotheses in (3.11) is obtained by rejecting the null hypothesis, whenever

$$[0, \Delta] \cap \tilde{I}_n^x = [0, \Delta] \cap [q_\alpha \hat{V}_n^x + M^x[\hat{C}_n], \infty) = \emptyset,$$

which is obviously equivalent to (3.13). As the same arguments apply to the measure $M_{\text{rel}}^x[C]$ these considerations lead to the following result.

Theorem 3.12. *Suppose that the assumptions of Theorem 3.11 hold and that $\Delta > 0$ ($\Delta_{\text{rel}} > 0$). The test (3.13), (3.14) is a consistent asymptotic-level- α test for the hypotheses (3.11), (3.12), respectively.*

Remark 3.13.

- (1) It follows from the proof of Theorem 3.11 that the probability of rejecting the null hypothesis in (3.11) by the test (3.13) converges to 0 if $M^x[C] < \Delta$, to α if $M^x[C] = \Delta$ and to 1 if $M^x[C] > \Delta$. A similar statement can be made for the hypothesis (3.12) and the test (3.14)
- (2) The methodology can easily be extended to test the hypotheses

$$H_0 : M^x[C] > \Delta \quad \text{vs.} \quad H_1 : M^x[C] \leq \Delta.$$

A consistent asymptotic level α -test is obtained by rejecting the null hypothesis, whenever

$$M^x[\hat{C}_n] + q_{1-\alpha} \hat{V}_n^x \leq \Delta,$$

Note that this formulation of the testing problem allows to decide for an approximately separable covariance structure at a controlled type-I error. Similarly, a consistent and asymptotic level α -test

for the hypotheses $H_0 : M_{\text{rel}}^x[C] > \Delta_{\text{rel}}$ vs. $H_1 : M_{\text{rel}}^x[C] \leq \Delta_{\text{rel}}$ is obtained by rejecting the null hypothesis whenever $M_{\text{rel}}^x[\hat{C}_n] + q_{1-\alpha} \hat{V}_{\text{rel},n} \leq \Delta_{\text{rel}}$. The proofs of these properties are omitted for the sake of brevity.

- (3) A careful inspection of the proofs in the Appendix shows that all results presented in this section are also valid without the assumption $\mathbb{E}X_n(s, t) \equiv 0$. In this case one has to replace each X_n in the definition of $\hat{C}_n(\lambda)$ in (3.3), by the empirically centred version $X_n - \bar{X}$, where \bar{X} is the average function of X_1, \dots, X_n .

4 Finite sample properties

4.1 Simulations

In this section we demonstrate the applicability of our approach by virtue of a simulation study. The measures μ_1 and μ_2 in (3.1) are chosen as uniform distributions on the sets $\{0, 1/(S-1), 2/(S-1), \dots, 1\}$ and $\{0, 1/(T-1), 2/(T-1), \dots, 1\}$, respectively. Similarly as in Constantinou et al. (2018), we generate synthetic data, according to the following moving average model

$$X_k(s, t) := \sum_{s'=1}^S \exp\{-b^2(s-s')^2\} [e_k(t, s') + e_{k-1}(t, s')] \quad k = 1, \dots, n, \quad (4.1)$$

where the spatio-temporal, random functions e_0, \dots, e_n are i.i.d. realizations of a centred Gaussian process $e \in H$, with covariance function

$$\sigma(s, s', t, t') := \frac{1}{(a|t-t'|+1)^{1/2}} \exp\left(-\frac{b^2|s-s'|^2}{(a|t-t'|+1)^c}\right) + c\left(1 - \frac{|s^2-s'^2|}{2} - (t-t')\right)_+, \quad (4.2)$$

and we choose $a = 10$ and $b = 5$. The choice of the covariance structure (4.2) is similarly motivated in Cressie and Huang (1999), who investigate classes of separable and non-separable functions. Related versions have been used for instance in Gneiting et al. (2006); Constantinou et al. (2018) and Bagchi and Dette (2020). As we can see by inspecting σ , the covariance function is separable for $c = 0$ and the corresponding covariance operator C of the data inherits this separability.

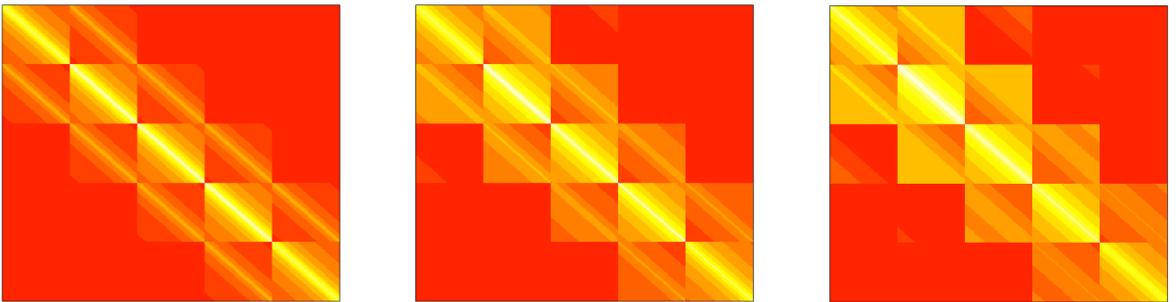


Figure 2: The covariance operator C in model (4.1) where $S = 5$ (25 spatial blocks) and $T=50$, for different choices of the separability parameter c (left $c = 0$, middle $c = 0.3$, right $c = 0.6$).

In Figure 2 we can visually inspect the increasing deviation from separability: The left operator ($c = 0$) is separable, i.e. all of the 25 spacial blocks are scaled versions of one single time matrix. For $c = 0.3$ and more strongly $c = 0.6$ this homogeneity disappears. The effect is most visible in the off-diagonal blocks which become asymmetric. In Figure 3 (left), we quantify the deviation from separability by virtue of our separability measures $M^x[C]$. Notice that for the measure of partial products we have chosen the time matrix to be $\text{Tr}_1[C]$ and optimize with respect to the space matrix (this corresponds to $\Delta_1 = \text{Id}$). In particular this implies that $M^{\text{tr}}[C] \geq M^{\text{prod}}[C]$. Finally we can assess separability by inspection of the singular values of the restacked operator $\Pi[C]$. Recall that if $\Pi[C]$ is rank one (just one positive singular value) C is separable. In Figure 3 (right) we display the first 15 singular values of $\Pi[C]$. The rapid decay after the first singular value indicates, that despite C 's observable inseparability, it is still reasonably close to the optimal, separable approximation C^{opt} .

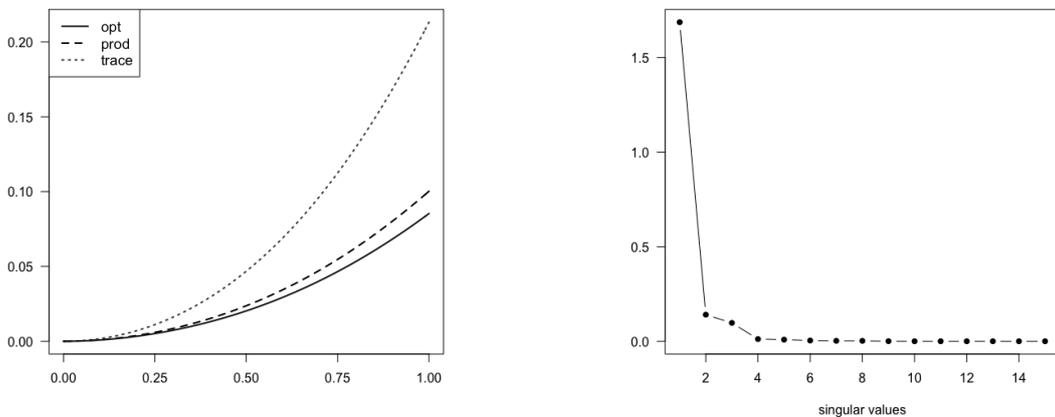


Figure 3: *Left: Comparison of the separability measures M^{tr} , M^{prod} , M^{opt} for $S = 5$ and $c \in [0, 1]$. Right: Singular values of $\Pi[C]$ for $S = 5$ and fixed separability parameter $c = 0.6$.*

In the following we investigate the coverage probabilities of the asymptotic confidence intervals defined in (3.9) and (3.10) for finite samples. For this purpose we simulate the data functions (4.1) for $T = 50$ and $S \in \{5, 10, 15, 20\}$. All simulations are based on 5000 simulation runs. The relative frequency of the events $M^x[C] \in \hat{I}_n^x$ and $M_{\text{rel}}^x[C] \in \hat{I}_{\text{rel},n}^x$ are displayed in Table 2.

We observe a reasonable approximation of the coverage probabilities in most cases. First, the results are relatively stable with respect to the number S of spatial locations. Second, the coverage probabilities for the absolute differences are slightly too large and a better approximation is attained for the relative differences (here in some case the coverage probabilities are slightly too small). For the absolute measures, the confidence intervals for partial traces approximate the prescribed $(1 - \alpha)$ -level better than those for partial products and optimal approximations. In particular we frequently observe that $\mathbb{P}\{M^{\text{opt}}[C] \in \hat{I}_n^{\text{opt}}\} \geq \mathbb{P}\{M^{\text{tr}}[C] \in \hat{I}_n^{\text{tr}}\}$. Despite this property we almost invariably observe that the interval \hat{I}_n^{opt} is narrower than \hat{I}_n^{tr} (these results are not displayed). This is due to the superior quality of optimal approximations compared to those by partial traces. This difference is visualized in the left panel of Figure 3, where we observe a sizable spread between the measures, already for $c = 0.6$. For larger c the gap between $M^{\text{tr}}[C]$ and $M^{\text{opt}}[C]$ grows so rapidly, that they are hardly comparable at all. In contrast

the measures $M^{\text{opt}}[C]$ and $M^{\text{prod}}[C]$ are very close in this example, the corresponding intervals are similar and level-approximation is also comparable. However, it should be pointed out that there exist covariance operators where the difference between these measures is larger, as it has been indicated in Example 2.3 (see Figure 1). It will also become clear in the subsequent data example, that this difference potentially translates into noticeably different confidence intervals.

		absolute				relative				
	n	α	$S = 5$	$S = 10$	$S = 15$	$S = 20$	$S = 5$	$S = 10$	$S = 15$	$S = 20$
tr	100	5%	0.969	0.971	0.971	0.971	0.930	0.945	0.941	0.941
		10%	0.922	0.924	0.926	0.924	0.858	0.876	0.868	0.876
	200	5%	0.963	0.966	0.967	0.963	0.955	0.957	0.953	0.964
		10%	0.909	0.917	0.918	0.921	0.888	0.899	0.906	0.900
prod	400	5%	0.957	0.958	0.960	0.959	0.941	0.961	0.964	0.958
		10%	0.899	0.905	0.900	0.897	0.876	0.902	0.900	0.904
	100	5%	0.979	0.980	0.981	0.981	0.932	0.944	0.944	0.940
		10%	0.948	0.934	0.946	0.944	0.856	0.870	0.877	0.875
opt	200	5%	0.975	0.977	0.984	0.985	0.955	0.938	0.932	0.931
		10%	0.931	0.937	0.947	0.952	0.894	0.908	0.919	0.917
	400	5%	0.970	0.976	0.979	0.982	0.940	0.962	0.969	0.979
		10%	0.917	0.937	0.940	0.942	0.875	0.906	0.917	0.930
tr	100	5%	0.981	0.979	0.980	0.981	0.928	0.940	0.939	0.938
		10%	0.939	0.937	0.944	0.946	0.852	0.874	0.861	0.871
	200	5%	0.975	0.966	0.985	0.985	0.952	0.959	0.966	0.966
		10%	0.932	0.937	0.951	0.954	0.888	0.902	0.916	0.916
400	5%	0.969	0.976	0.981	0.984	0.938	0.959	0.966	0.979	
	10%	0.918	0.935	0.944	0.945	0.870	0.900	0.915	0.931	

Table 2: Empirical coverage probabilities of the confidence intervals (3.9) and (3.10). The data is generated according to model (4.1), with separability parameter $c = 0.6$.

4.2 Data Example

In this section we employ the methodology developed in this paper to investigate the different measures of separability in a data example. The data consists of daily temperature averages, published by the national meteorological agency “Der Deutsche Wetterdienst” (https://www.dwd.de/DE/Home/home_node.html) at seven different stations in the cities of Bremen, Cottbus, Hohenpeissenberg, Karlsruhe, Magdeburg, Potsdam and Schwerin. We consider data from the years 1893 – 1941 and 1947 – 2008, where a middle period is left out because of incomplete measurements during and immediately after WWII.

The data is smoothed over a Fourier basis with 41 coefficients, which yields reasonable approximations of the temperature, while still reflecting general trends. Furthermore it has been empirically centered and a linear trend of temperature increase has been removed to account for climate change.

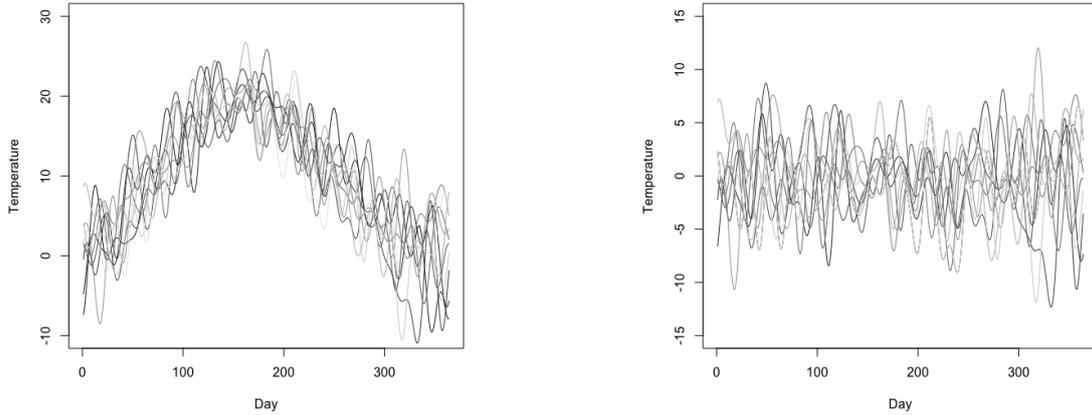


Figure 4: *Temperature curves in the first ten years of observation at Hohenpeissenberg. Left: Observations. Right: Observations after trend removal.*

In Figure 4 we illustrate the effect of centering and trend removal at the first ten observations at the location of Hohenpeissenberg (left before and right after detrending). We now turn to the investigation of the covariance operators of different collections of cities. Exemplarily we display in Figure 5 (left) the covariance operator corresponding to the locations Bremen, Hohenpeissenberg, Karlsruhe and Potsdam. Compared to the covariance model considered in Section 4, we observe a narrower concentration along the diagonals of the block matrices. This almost-bandedness is also present in other geostatistical applications such as wind data (see, for example, Figure 1 in Genton, 2007). In Figure 5 (right) we visualize the first fifteen singular values of the restacked operator $\Pi[\hat{C}_n]$, where the sharp decay after the first one already hints at approximate separability. This impression will be reinforced by our subsequent investigations.

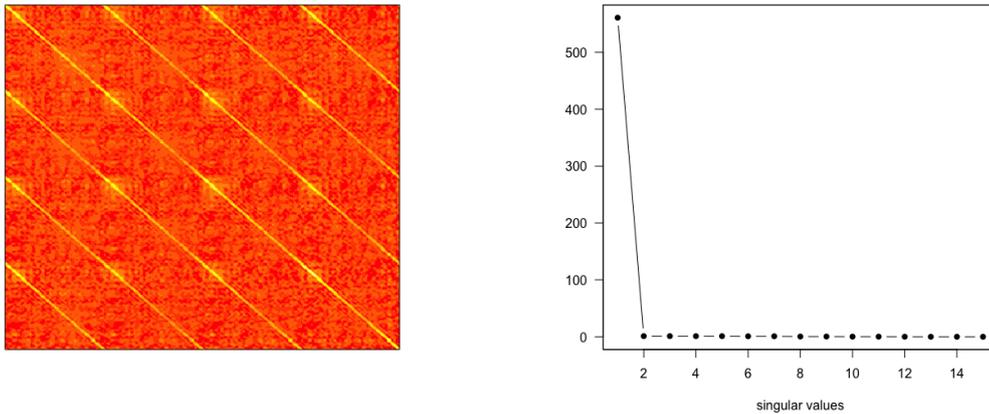


Figure 5: *Left: Empirical covariance operator \hat{C}_n for the cities Bremen, Hohenpeissenberg, Karlsruhe and Potsdam. Right: Singular values of $\Pi[\hat{C}_n]$. The fast decay after the first singular value indicates approximate separability.*

After these descriptive examinations we proceed to statistical inference. We consider different collections of cities and investigate the separability of the corresponding covariance operators by virtue of the empirical 95%-confidence intervals for the measures $M^x[C]$ ($x \in \{\text{tr}, \text{prod}, \text{opt}\}$). In particular we examine the differences between approximation methods.

To make the analysis as comprehensible as possible, we restrict ourselves to the measures of relative deviation, which are more easily interpretable. Moreover we do not display all possible combinations of cities, but confine ourselves to some illustrative examples. In the following the x -axis represents relative deviation in percent.

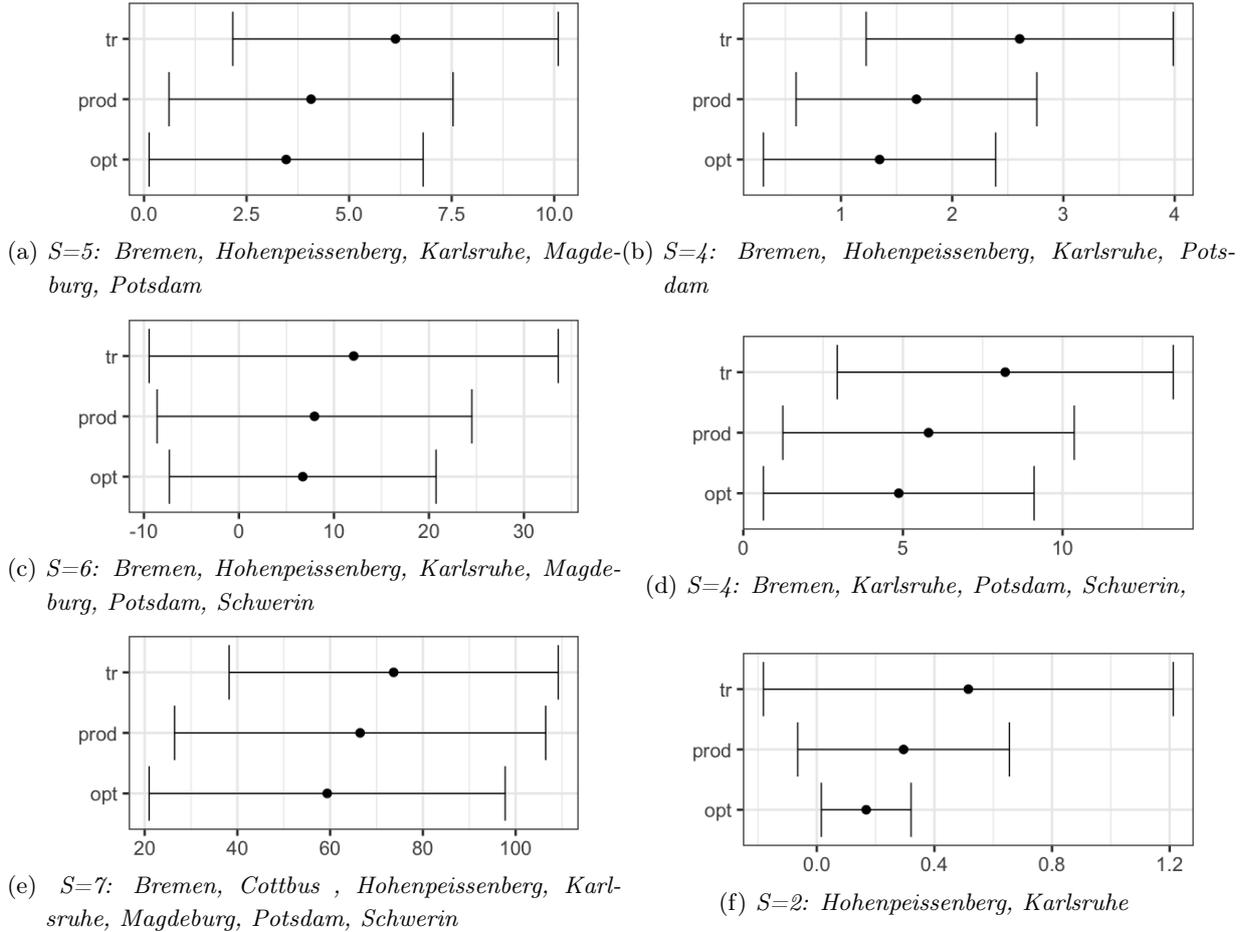


Figure 6: Empirical confidence intervals $\hat{I}_{\text{rel},n}^x$ for the relative separability measures $M_{\text{rel}}^x[C]$ ($x \in \{\text{tr}, \text{prod}, \text{opt}\}$). The x -axis represents the relative deviation in percent.

In the top two panels of Figure 6 (a) and (b), we observe deviation from separability but only to a small degree. In such cases (especially in (b)) it is reasonable to approximate the true covariance operator by a separable version, because the incurred imprecision is small. The memory space, which is saved by separable approximations for $S = 5$ amounts to $\approx 96\%$ and for $S = 4$ to $\approx 94\%$, which is quite substantial. When we compare the measures of separability we notice considerable differences. For example in (b) we

observe $M_{\text{rel}}^{\text{tr}}[\hat{C}_n] \notin \hat{I}_{\text{rel},n}^{\text{opt}}$, which underlines with statistical significance the differences of the measures. In the middle panels (c) and (d), we display examples of moderate deviations of separability. The confidence intervals for the relative measure of optimal approximations tend to be tighter than the other two (this is most visible in (c), where $\hat{I}_{\text{rel},n}^{\text{opt}} \subset \hat{I}_{\text{rel},n}^{\text{prod}} \subset \hat{I}_{\text{rel},n}^{\text{tr}}$). The bottom two examples are cases of either extreme inseparability (e) and extreme separability (f). In (e) we observe that all of the relative measures assume values $> 50\%$. Here it is not recommended to approximate the covariance by a separable operator. Even linear combinations of separable operators are not suitable in this case. In contrast, on the right we observe that a separable approximation of the covariance is reasonable. This may be useful despite the small number of spatial locations, as it still allows the reduction of memory space to a quarter.

Acknowledgements This work has been supported in part by the Collaborative Research Center “Statistical modeling of nonlinear dynamic processes” (SFB 823, Teilprojekt A1, C1) of the German Research Foundation (DFG). We are grateful to P. Kokoszka, M. Reimherr and S. Tavakoli for pointing out important references and helpful discussions.

References

- Aitchison, J. (1964). Confidence-region tests. *Journal of the Royal Statistical Society, Series B* 26, 462–476.
- Aston, J. A. D., D. Pigoli, and S. Tavakoli (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics* 45(4), 1431–1461.
- Bagchi, P. and H. Dette (2020). A test for separability in covariance operators of random surfaces. *Annals of Statistics*, To Appear. <https://arxiv.org/pdf/1710.08388.pdf>.
- Bel, L., A. Bar-Hen, R. Petit, and R. Cheddadi (2011). Spatio-temporal functional regression on paleoecological data. *Journal of Applied statistics* 38(4), 695–704.
- Berkes, I., L. Horváth, and G. Rice (2013). Weak invariance principles for sums of dependent random functions. *Stochastic Processes and their Applications* 123(2), 385 – 403.
- Bhatia, R. (2003). Partial traces and entropy inequalities. *Linear Algebra and its Applications* 370(1), 125–132.
- Billingsley, P. (1999). *Convergence of Probability Measures, 2nd Edition*. Wiley Series in Probability and Statistics. John Wiley & Sons.
- Bücher, A. and I. Kojadinovic (2016). A dependent multiplier bootstrap for the sequential empirical copula process under strong mixing. *Bernoulli* 22, 927–968.
- Cardot, H. and D. Degras (2018). Online principal component analysis in high dimension: Which algorithm to choose? *International Statistical Review* 86(1), 29–50.
- Chattopadhyay, A., P. Hassanzadeh, and S. Pasha (2020). Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific Reports* 10, 1317.
- Constantinou, P., P. Kokoszka, and M. Reimherr (2017). Testing separability of space-time functional processes. *Biometrika* 104(2), 425–437.

- Constantinou, P., P. Kokoszka, and M. Reimherr (2018). Testing separability of functional time series. *Journal of Time Series Analysis* 39(5), 731–747.
- Cressie, N. (2015). *Statistics for Spatial Data, Revised Edition*. Wiley Classics Literature. John Wiley & Sons.
- Cressie, N. and H.-C. Huang (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* 94(448), 1330–1339.
- Cressie, N. and C. K. Wikle (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.
- Crujeiras, R. M., Fernández-Casal, and W. González-Manteiga (2010). Nonparametric test for separability of spatio-temporal processes. *Environmetrics* 21, 382–399.
- Davies, E. B. (2007). *Linear Operators and their Spectra*. Cambridge University Press.
- Dehling, H. (1983). Limit theorems for sums of weakly dependent banach space valued random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 63(3), 393–432.
- Dehling, H., T. Mikosch, and M. Sørensen (2002). *Empirical process techniques for dependent data*. Birkhäuser.
- Dehling, H. and W. Philip (1982). Almost sure invariance principles for weakly dependent vector-valued random variables. *Annals of Probability* 10(3), 689–701.
- Dette, H., K. Kokot, and S. Volgushev (2020). Testing relevant hypotheses in functional time series via self-normalization. *Journal of the Royal Statistical Society, Series B*, To Appear. <https://arxiv.org/pdf/1809.06092.pdf>.
- Dunford, N. and J. T. Schwartz (1958). *Linear operators. Part I: General theory*. New York, Interscience Publishers.
- Fuentes, M. (2006). Testing for separability of spatial–temporal covariance functions. *Journal of Statistical Planning and Inference* 136(2), 447–466.
- Genton, M. (2007). Separable approximations of space-time covariance matrices. *Environmetrics* 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., M. G. Genton, and P. Guttorp (2006). *Geostatistical Space-Time Models, Stationarity, Separability and Full Symmetry* (1 ed.), Volume 107, Chapter 4, pp. 151–175. Chapman & Hall/CRC Press.
- Gohberg, I., S. Goldberg, and M. A. Kaashoek (1993). *Classes of Linear Operators*, Volume 63 of *Operator Theory: Advances and Applications*. Birkhäuser Basel.
- Gohberg, I., S. Goldberg, and M. A. Kaashoek (2003). *Basic Classes of Linear Operators* (1 ed.). Birkhäuser Basel.
- Golub, G. H. and C. F. Van Loan (2013). *Matrix Operations* (4 ed.), Volume 3 of *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins University Press.
- Gross, L. (1970). *Lectures in Modern Analysis and Applications II* (1 ed.), Volume 140 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg.
- Hasselmann, K. (1993). Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate* 6(10), 1957–1971.

- Hörmann, S. and P. Kokoszka (2010). Weakly dependent functional data. *The Annals of Statistics* 38(3), 1845–1884.
- Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. New York: Springer Series in Statistics.
- Huang, H.-C. and N.-J. Hsu (2004). Modeling transport effects on ground-level ozone using a non-stationary space-time model. *Environmetrics* 15(3), 251–268.
- Jun, M. and M. L. Stein (2004). Statistical comparison of observed and cmaq modeled daily sulfate levels. *Atmospheric Environment* 38(27), 4427–4436.
- Krall, J. R., H. H. Chang, S. Ebel-Sarnat, R. D. Peng, and L. Waller (2015). Current methods and challenges for epidemiological studies of the associations between chemical constituents of particulate matter and health. *Curr. Envir. Health Rep.* 2, 388–398.
- Kuelbs, J. and W. Philip (1980). Almost sure invariance principles for partial sums of mixing b-valued random variables. *The Annals of Probability* 8(6), 1003–1036.
- Kyriakidis, P. C. and A. G. Journel (1999). Geostatistical space–time models: A review. *Mathematical Geology* 31, 651–684.
- Laurini, M. P. (2019). A spatio-temporal approach to estimate patterns of climate change. *Environmetrics* 30(1), e2542.
- Li, B., M. G. Genton, and M. Sherman (2007). A nonparametric assessment of properties of space–time covariance functions. *Journal of the American Statistical Association* 102(478), 736–744.
- Lindquist, M. A. (2008). The statistical analysis of fmri data. *Statistical Science* 23(4), 439–464.
- Lu, N. and D. L. Zimmerman (2005). Testing for directional symmetry in spatial dependence using the periodogram. *Journal of Statistical Planning and Inference* 129(1–2), 369–385.
- Mas, A. (2006). A sufficient condition for the clt in the space of nuclear operators—application to covariance of random functions. *Statistics and Probability Letters* 76(14), 1503–1509.
- Matsuda, Y. and Y. Yajima (2004). On testing for separable correlations of multivariate time series. *Journal of Time Series Analysis* 25(4), 501–528.
- Mitchell, M. W., M. G. Genton, and M. L. Gumpertz (2005). Testing for separability of space–time covariances. *Environmetrics* 16(8), 819–831.
- Mitchell, M. W., M. G. Genton, and M. L. Gumpertz (2006). A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis* 97(5), 1025–1043.
- Rabiner, R. L. and R. W. Schafer (1978). *Digital Processing of Speech Signals*, Volume 100 of *Signal Processing Series*. Prentice-Hall.
- Samur, J. D. (1984). Convergence of sums of mixing triangular arrays of random vectors with stationary rows. *The Annals of Probability* 12(2), 390–426.
- Samur, J. D. (1987). On the invariance principle for stationary ϕ -mixing triangular arrays with infinitely divisible limits. *Probability Theory and Related Fields* 75(2), 245–259.
- Scaccia, L. and R. J. Martin (2005). Testing axial symmetry and separability of lattice processes. *Journal of Statistical Planning and Inference* 131(1), 19–39.
- Sherman, M. (2010). *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Wiley Series in Probability and Statistics. John Wiley & Sons.

- Sigrist, F., H. R. Künsch, and W. A. Stahel (2011). An autoregressive spatio-temporal precipitation model. *Procedia Environmental Sciences* 3, 2–7.
- Silverman, B. W. and J. A. Rice (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)* 53(1), 233–243.
- Skup, M. (2010). Longitudinal fmri analysis: A review of methods. *Statistics and its interface* 3(2), 235–252.
- Stein, M. L. (2005). Space–time covariance functions. *Journal of the American Statistical Association* 100(469), 310–321.
- Tingley, M. P. and P. Huybers (2013). Recent temperature extremes at high northern latitudes unprecedented in the past 600 years. *Nature* 496, 201–205.
- Tzeng, J. (2013). Split-and-combine singular value decomposition for large-scale matrix. *Journal of Applied Mathematics*.
- van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes. With applications to statistics*. New York: Springer Series in Statistics.
- Van Loan, C. F. and N. Pitsianis (1993). *Approximation with Kronecker Products*, Volume 232 of *NATO ASI Series*, pp. 293–314. Springer.
- Weidmann, J. (1980). *Linear Operators in Hilbert spaces*, Volume 68 of *Graduate Texts in Mathematics*. Berlin, New York, Springer.
- Worsley, K. J., S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, and A. C. Evans (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping* 4, 58–73.

A Proofs and technical details

A.1 Details on the calculation of the separability measures

In this section we provide some details concerning the example of (in-)separability of the matrix

$$C(q) := \left(\begin{array}{cc|cc} 2 & 0 & 1 & q \\ 0 & 2 & q & 1 \\ \hline 1 & q & 2 & q \\ q & 1 & q & 2 \end{array} \right) \in \mathcal{S}_2(\mathbb{R}^2 \otimes \mathbb{R}^2, \mathbb{R}^2 \otimes \mathbb{R}^2)$$

introduced in Section 2. We will show in detail how each of the separability measures is derived, starting with the partial traces, where a straightforward calculation yields

$$\mathrm{Tr}_1[C(q)] = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 2 & q \\ q & 2 \end{pmatrix} = \begin{pmatrix} 4 & q \\ q & 4 \end{pmatrix}, \quad \mathrm{Tr}_2[C(q)] := \begin{pmatrix} 2+2 & 1+1 \\ 1+1 & 2+2 \end{pmatrix} = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}.$$

The trace of $C(q)$ is 8. Hence we have

$$C(q)^{\mathrm{tr}} = \mathrm{Tr}_2[C(q)] \otimes \mathrm{Tr}_1[C(q)] / \mathrm{Tr}[C(q)] = \begin{pmatrix} 2 & q/2 & 1 & q/4 \\ q/2 & 2 & q/4 & 1 \\ \hline 1 & q/4 & 2 & q/2 \\ q/4 & 1 & q/2 & 2 \end{pmatrix}$$

It is now easy to see that $M^{\mathrm{tr}}[C(q)] = \|C(q) - C(q)^{\mathrm{tr}}\|_2^2 = 13/4 \cdot q^2$. For partial products and the parameter choice $\Delta_2 = \mathrm{Id}_2$ (the two-dimensional identity matrix) we have, that $P_2[C(q), \Delta_2] = \mathrm{Tr}_2[C(q)]$, which is calculated above. Notice that $C(q)$ can be decomposed as the sum of the separable matrices

$$C(q) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \otimes \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 2 & q \\ q & 2 \end{pmatrix}.$$

$P_1[C, P_2(C, \Delta_2)]$ can now be calculated using bilinearity

$$P_1[C(q), \mathrm{Tr}_2[C(q)]] = 4 \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + 4 \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} + 4 \begin{pmatrix} 2 & q \\ q & 2 \end{pmatrix} = \begin{pmatrix} 20 & 8q \\ 8q & 20 \end{pmatrix},$$

where we have used the defining property from (2.5) (the prefactors of the matrices are the Frobenius products of the space matrices and the partial trace $\mathrm{Tr}_2[C(q)]$). The squared Frobenius norm of $\mathrm{Tr}_2[C(q)]$ is 40. Together this yields

$$C^{\mathrm{prod}}(q) = \mathrm{Tr}_2[C(q)] \otimes P_1[C(q), \mathrm{Tr}_2[C(q)]] / \| \mathrm{Tr}_2[C(q)] \|_2^2 = \begin{pmatrix} 2 & 4/5q & 1 & 2/5q \\ 4/5q & 2 & 2/5q & 1 \\ \hline 1 & 2/5q & 2 & 4/5q \\ 2/5q & 1 & 4/5q & 2 \end{pmatrix}$$

and hence a simple calculation reveals that $M^{\mathrm{prod}}[C(q)] = \|C(q) - C(q)^{\mathrm{prod}}\|_2^2 = 70/25 \cdot q^2$. Finally we calculate the optimal approximation measure. First we notice that the squared Frobenius norm of $C(q)$

equals $20 + 6q^2$. Secondly, we calculate

$$\Pi[C(q)]\Pi[C(q)]^* = \begin{pmatrix} 8 & 4 & 4 & 8 \\ 4 & 2 + 2q^2 & 2 + 2q^2 & 4 + 2q^2 \\ 4 & 2 + 2q^2 & 2 + 2q^2 & 4 + 2q^2 \\ 8 & 2 + 2q^2 & 4 + 2q^2 & 8 + 2q^2 \end{pmatrix},$$

the largest eigenvalue of which is given by $\gamma_1^2 := 3q^2 + \sqrt{9q^4 + 4q^2 + 100} + 10$. As a consequence the measure of absolute deviation is given by

$$M^{\text{opt}}[C(q)] = \|C(q) - C(q)^{\text{opt}}\|_2^2 = \|C(q)\|_2^2 - \gamma_1^2 = 10 + 3q^2 - \sqrt{9q^4 + 4q^2 + 100}.$$

A.2 Proof of Theorem 3.3

The proof of this theorem consists of two steps. First we establish a functional CLT, to demonstrate weak convergence of $\sqrt{n}(\hat{C}_n - C)$ in the space of trace-class operators. Subsequently by an application of Theorem 3.2 in Samur (1987) we extend this result to an invariance principle. We note that the result of Samur (1987) concerns weak convergence in the Skorohod space equipped with its Skorohod metric. Now, for continuous functions the Skorohod metric reduces to the supremum metric, see Chapter 3 of Billingsley (1999). Since our random variables as well as the limiting Brownian motion have continuous sample paths, we obtain weak convergence in the space of continuous trace-class operators. We begin by stating the CLT.

Theorem A.1. *Let $(X_i)_{i \in \mathbb{Z}}$ be a sequence of random functions in H , defined in (3.1), satisfying Assumption 3.2. Then there exists a Gaussian random operator G in $\mathcal{S}_1(H, H)$ such that*

$$\sqrt{n}(\hat{C}_n - C) \xrightarrow{d} G.$$

The proof of this theorem consists in the application of a CLT for Banach space-valued random variables, namely Theorem 4.4 of Samur (1984). In the following we are going to verify its four conditions:

(C1) For any sequence of natural numbers $\{r_n\}_{n \in \mathbb{N}}$ with $r_n/n \rightarrow 0$ it follows that

$$\left\| \frac{1}{\sqrt{r_n}} \sum_{j=1}^{r_n} [X_j \otimes X_j - C] \right\|_1 \xrightarrow{\mathbb{P}} 0. \quad (\text{A.1})$$

(C2) For any $\varepsilon > 0$

$$r_n \mathbb{P} \left\{ \|X_1 \otimes X_1 - C\|_1 / \sqrt{r_n} > \varepsilon \right\} \rightarrow 0.$$

(C3) For all $L \in \mathcal{L}(H, H)$ the following limit exists

$$\Phi(L) := \lim_{n \rightarrow \infty} \mathbb{E} \text{Tr} \left(\sum_{j=1}^n [X_j \otimes X_j - C] \circ L \right)^2.$$

(C4) There exists an increasing sequence of finite dimensional subspaces $F_k \subset \mathcal{S}_1(H, H)$, such that

$$\lim_{k \rightarrow \infty} \sup_n \mathbb{E} \inf_n \left\{ \left\| \sum_{j=1}^n (X_j \otimes X_j - C) - f \right\|_1 : f \in F_k \right\} = 0,$$

where $\bigcup_{k \in \mathbb{N}} F_k$ is dense in $\mathcal{S}_1(H, H)$.

We will now proceed to verify each of the claims.

In order to show (C1), we prove convergence to 0 of the left side of (A.1) in expectation, which entails convergence in probability. By Lemma 32 on p. 1116 of Dunford and Schwartz (1958), we obtain for the expected trace norm

$$\mathbb{E} \left\| \frac{1}{\sqrt{n}} \sum_{j=1}^{r_n} (X_j \otimes X_j - C) \right\|_1 \leq \mathbb{E} \sum_{q \in \mathbb{N}} \frac{1}{\sqrt{n}} \left\| \sum_{j=1}^{r_n} Y_{j,q} \right\| =: E_n, \quad (\text{A.2})$$

where we have defined the centered random variables

$$Y_{j,q} := X_j \langle X_j, e_q \rangle - \mathbb{E} X_j \langle X_j, e_q \rangle \in H. \quad (\text{A.3})$$

An application of Jensen's inequality to the right hand side of (A.2) yields

$$E_n \leq \sqrt{\frac{r_n}{n}} \sum_{q \in \mathbb{N}} \left\{ \mathbb{E} \left\langle \sum_{j=1}^{r_n} Y_{j,q}, \sum_{k=1}^{r_n} Y_{k,q} \right\rangle \frac{1}{r_n} \right\}^{1/2}. \quad (\text{A.4})$$

Recalling the stationarity of the sequence $Y_{j,q}$, $j, q \geq 1$, inherited by X_j , $j \geq 1$, we can simplify this expression to

$$\sqrt{\frac{r_n}{n}} \sum_{q \in \mathbb{N}} \left\{ \mathbb{E} \sum_{|h| < r_n} \langle Y_{0,q}, Y_{|h|,q} \rangle \left(1 - \frac{|h|}{r_n} \right) \right\}^{1/2} \leq \sqrt{\frac{r_n}{n}} \sum_{q \in \mathbb{N}} \left\{ 2 \sum_{h \geq 0} |\mathbb{E} \langle Y_{0,q}, Y_{h,q} \rangle| \right\}^{1/2} \quad (\text{A.5})$$

We will now exploit the ϕ -mixing property of $Y_{h,q}$ to upper bound the covariance terms. We therefore use a special case of a general result found in Dehling et al. (2002) on page 24, which yields

$$|\mathbb{E} \langle Y_{0,q}, Y_{h,q} \rangle| \leq 2\sqrt{\phi(|h|)} \mathbb{E} \|Y_{0,q}\|^2. \quad (\text{A.6})$$

Plugging this bound in (A.5) and observing (A.4) we obtain

$$E_n \leq \sqrt{\frac{r_n}{n}} \sum_{q \in \mathbb{N}} \left\{ 2 \sum_{h \geq 0} 2\sqrt{\phi(h)} \mathbb{E} \|Y_{0,q}\|^2 \right\}^{1/2} = \sqrt{\frac{r_n}{n}} \sum_{q \in \mathbb{N}} \sqrt{\mathbb{E} \|Y_{0,q}\|^2} \left\{ 4 \sum_{h \geq 0} \sqrt{\phi(h)} \right\}^{1/2}.$$

Since by assumption $r_n/n \rightarrow 0$, it remains to prove the finiteness of both sums on the right to establish (C1). The second one is finite due to our summability assumption on $\sqrt{\phi(h)}$. To see the boundedness of the first one we recall the definition of $Y_{0,q}$ in (A.3). By the binomial formula in Hilbert spaces we get

$$\mathbb{E} \|Y_{0,q}\|^2 = \mathbb{E} \|X_0 \langle X_0, e_q \rangle - \mathbb{E} X_0 \langle X_0, e_q \rangle\|^2 \leq 2\mathbb{E} \|X_0\|^2 \langle X_0, e_q \rangle^2.$$

Applying the Cauchy–Schwarz inequality finally yields

$$\mathbb{E} \|X_0\|^2 \langle X_0, e_q \rangle^2 \leq \sqrt{\mathbb{E} \|X_0\|^4} \sqrt{\mathbb{E} \langle X_0, e_q \rangle^4}.$$

Noticing that by assumption

$$\sum_{q \in \mathbb{N}} \sqrt[4]{\mathbb{E} |\langle X_0, e_q \rangle|^4} < \infty,$$

the proof of condition (C1) is completed.

Verifying (C2) is straightforward. Indeed, it follows immediately by an application of Markov's inequality

and the dominated convergence theorem.

To establish condition (C3), let L be an arbitrary bounded operator. By definition of the trace we observe that $\Phi(L)$ equals

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} \left\{ \sum_{p \in \mathbb{N}} \left\langle \left[\sum_{i=1}^n X_i \otimes X_i - C \right] L[e_q], e_p \right\rangle \sum_{q \in \mathbb{N}} \left\langle \left[\sum_{j=1}^n X_j \otimes X_j - C \right] L[e_q], e_q \right\rangle \right\} \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i,j=1}^n \mathbb{E} \left\{ \sum_{p \in \mathbb{N}} \left\langle [X_i \otimes X_i - C] L[e_q], e_p \right\rangle \sum_{q \in \mathbb{N}} \left\langle [X_j \otimes X_j - C] L[e_q], e_q \right\rangle \right\}. \end{aligned}$$

We use the stationarity of the sequence X_j , $j \geq 1$, to transform the last expression into

$$\lim_{n \rightarrow \infty} \sum_{|h| < n} \left(1 - \frac{|h|}{n} \right) \mathbb{E} \left\{ \sum_{p \in \mathbb{N}} \left\langle [X_0 \otimes X_0 - C] L[e_p], e_p \right\rangle \sum_{q \in \mathbb{N}} \left\langle [X_{|h|} \otimes X_{|h|} - C] L[e_q], e_q \right\rangle \right\}.$$

To establish convergence for $n \rightarrow \infty$, we use the dominated convergence theorem. Dominated convergence is guaranteed by the following calculation:

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sum_{|h| < n} \left| \mathbb{E} \left\{ \sum_{p \in \mathbb{N}} \left\langle [X_0 \otimes X_0 - C] L[e_p], e_p \right\rangle \sum_{q \in \mathbb{N}} \left\langle [X_{|h|} \otimes X_{|h|} - C] L[e_q], e_q \right\rangle \right\} \right| \quad (\text{A.7}) \\ & \leq 2 \sum_{h \geq 0} \left| \mathbb{E} \left\{ \sum_{p \in \mathbb{N}} \left\langle [X_0 \otimes X_0 - C] L[e_p], e_p \right\rangle \sum_{q \in \mathbb{N}} \left\langle [X_h \otimes X_h - C] L[e_q], e_q \right\rangle \right\} \right| \\ & \leq \sum_{h \geq 0} 2 \sqrt{\phi(h)} \mathbb{E} \left\{ \left[\sum_{p \in \mathbb{N}} \left\langle [X_0 \otimes X_0 - C] L[e_p], e_p \right\rangle \right]^2 \right\}. \end{aligned}$$

Here we have again used the covariance inequality from (A.6). Given the summability of $\sqrt{\phi(h)}$ the first sum is finite. Turning to the expectation on the right, we observe that it can be expressed more compactly as

$$\mathbb{E} \text{Tr} \left(\sqrt{n} (\hat{C}_n - C) L \right)^2,$$

which is upper bounded by

$$\mathbb{E} \left\| \sqrt{n} (\hat{C}_n - C) L \right\|_1^2. \quad (\text{A.8})$$

Theorem 5.6.7 in Davies (2007) states that for a trace class operator A and a bounded operator B the inequality $\|AB\|_1 \leq \|A^*\|_1 \|B\|_{\mathcal{L}}$ holds. Applying this inequality to (A.8) yields

$$\mathbb{E} \left\| \left[\sqrt{n} (\hat{C}_n - C) \right] \right\|_1^2 \|L\|_{\mathcal{L}}^2,$$

which is finite as can be deduced by similar reasoning as in (C1). Thus (A.7) is indeed finite and the dominated convergence theorem ensures that $\Phi(L)$ is well defined.

In order to show condition (C4), we define

$$F_k := \text{span}\{e_p \otimes e_q : 1 \leq p, q \leq k\} \subset \mathcal{S}_1(H, H),$$

which is a k^2 -dimensional subspace. We then conclude that

$$\begin{aligned}
& \limsup_{k \rightarrow \infty} \sup_n \mathbb{E} \inf_{f \in F_k} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n [X_i \otimes X_i - C] - f \right\|_1^2 \\
& \leq \limsup_{k \rightarrow \infty} \sup_n \frac{1}{n} \mathbb{E} \left\| \sum_{p,q>k} e_p \otimes e_q \sum_{i=1}^n \langle X_i, e_p \rangle \langle X_i, e_q \rangle - \mathbb{E} \langle X_i, e_p \rangle \langle X_i, e_q \rangle \right\|_1^2 \\
& \leq 2 \limsup_{k \rightarrow \infty} \sup_n \frac{1}{n} \mathbb{E} \left\| \sum_{p,q>k} e_p \otimes e_q \sum_{i=1}^n \langle X_i, e_p \rangle \langle X_i, e_q \rangle \right\|_1^2 \\
& + 2 \limsup_{k \rightarrow \infty} \sup_n \frac{1}{n} \mathbb{E} \left\| \sum_{p,q>k} e_p \otimes e_q \sum_{i=1}^n \mathbb{E} \langle X_i, e_p \rangle \langle X_i, e_q \rangle \right\|_1^2.
\end{aligned}$$

The first term on the right is according to Jensen's inequality upper bounding the second one. It will thus suffice to show that the former converges to 0. Notice that $\sum_{p,q>k} e_p \otimes e_q \sum_{i=1}^n \langle X_i, e_p \rangle \langle X_i, e_q \rangle$ is a symmetric, positive definite operator. Consequently its trace norm equals its trace.

$$\begin{aligned}
& \limsup_{k \rightarrow \infty} \sup_n \frac{1}{n} \mathbb{E} \left\| \sum_{p,q>k} e_p \otimes e_q \sum_{i=1}^n \langle X_i, e_p \rangle \langle X_i, e_q \rangle \right\|_1^2 \\
& = \limsup_{k \rightarrow \infty} \sup_n \frac{1}{n} \mathbb{E} \text{Tr} \left(\sum_{p,q>k} e_p \otimes e_q \sum_{i=1}^n \langle X_i, e_p \rangle \langle X_i, e_q \rangle \right)^2 \\
& = \limsup_{k \rightarrow \infty} \sup_n \frac{1}{n} \mathbb{E} \left(\sum_{p>k} \sum_{i=1}^n \langle X_i, e_p \rangle^2 \right)^2 \\
& = \limsup_{k \rightarrow \infty} \sup_n \frac{1}{n} \mathbb{E} \sum_{i,j=1}^n \sum_{p>k} \langle X_i, e_p \rangle^2 \sum_{q>k} \langle X_j, e_q \rangle^2 \\
& = \limsup_{k \rightarrow \infty} \sup_n \sum_{|h|<n} \left(1 - \frac{|h|}{n}\right) \sqrt{\phi(h)} \mathbb{E} \sum_{p>k} \langle X_0, e_p \rangle^2 = 0
\end{aligned}$$

In the last step we have again used the summability of the $\phi(h)$ as well as the fact that

$$\lim_{k \rightarrow \infty} \mathbb{E} \sum_{p>k} \langle X_0, e_p \rangle^2 = 0.$$

This concludes the proof of Theorem A.1. Theorem 3.3 now follows immediately by Theorem 3.2 in Samur (1987) (Condition I). □

A.3 Proof of Theorem 3.4

The proof of Fréchet differentiability in $\{\lambda \mapsto \lambda C\}$ is similar for all three maps. We have to demonstrate that each of them is the composition of Fréchet differentiable functions, which yields differentiability of the composition by the chain rule (see for instance Section 9 of Chapter 3 in van der Vaart and Wellner, 1996). For notational ease we subsequently do not distinguish between a function f and its evaluation $f(\lambda)$. In particular we write λC instead of $\{\lambda \mapsto \lambda C\}$.

To show Fréchet differentiability of \mathbf{F}^{tr} we notice that \mathbf{F}^{tr} is the composition of the two maps

$$\mathbf{A}^{(\text{tr})} : \mathcal{C}(I, \mathcal{S}_1(H, H)) \rightarrow \mathcal{C}(I, \mathcal{S}_1(H_1, H_1) \times \mathcal{S}_1(H_2, H_2) \times \mathcal{S}_1(H, H))$$

and

$$\mathbf{B}^{(\text{tr})} : \mathcal{C}(I, \mathcal{S}_1(H_1, H_1) \times \mathcal{S}_1(H_2, H_2) \times \mathcal{S}_1(H, H)) \rightarrow \mathcal{C}(I, \mathcal{S}_1(H, H))$$

pointwise defined as

$$\mathbf{A}^{(\text{tr})} : L(\lambda) \mapsto \begin{pmatrix} \text{Tr}_2[L(\lambda)]/\text{Tr}[L(\lambda)] \\ \text{Tr}_1[L(\lambda)] \\ L(\lambda) \end{pmatrix}$$

and

$$\mathbf{B}^{(\text{tr})} : \begin{pmatrix} L_1(\lambda) \\ L_2(\lambda) \\ L_3(\lambda) \end{pmatrix} \mapsto L_1(\lambda) \otimes L_2(\lambda) - L_3(\lambda).$$

Beginning with $\mathbf{A}^{(\text{tr})}$, we have to show Fréchet differentiability of each component. By definition (see Section 3.9 in van der Vaart and Wellner, 1996) the Fréchet derivative has to be a bounded, linear map. In the following we will use the continuity of the partial traces to find the desired Fréchet derivatives of the maps \mathbf{A}_i^{tr} , $i = 1, 2, 3$.

Due to linearity and continuity with respect to the maximum norm (defined in (3.4)) of the partial trace and the identity, the second and the third component of $\mathbf{A}^{(\text{tr})}$ are trivially Fréchet differentiable, where each map is its own derivative. For the first component $\mathbf{A}_1^{(\text{tr})}$ a straightforward calculation yields that for any function $h \in \mathcal{C}(I, \mathcal{S}_1(H, H))$ the following equation holds pointwise, for any $\lambda \in I$

$$\mathbf{A}_1^{(\text{tr})}[\lambda C + h(\lambda)] - \mathbf{A}_1^{(\text{tr})}[\lambda C] = \frac{\text{Tr}_2[C\lambda + h(\lambda)]}{\text{Tr}[C\lambda + h(\lambda)]} - \frac{\text{Tr}_2[C\lambda]}{\text{Tr}[C\lambda]} = \frac{\text{Tr}_2[h(\lambda)] \text{Tr}[C] - \text{Tr}_2[C] \text{Tr}[h(\lambda)]}{\text{Tr}[C\lambda + h(\lambda)] \text{Tr}[C]}.$$

We hence define the derivative $D\mathbf{A}_1^{(\text{tr})}(\lambda C)$ of $\mathbf{A}_1^{(\text{tr})}$ in λC as follows

$$D\mathbf{A}_1^{(\text{tr})}(\lambda C) : h(\lambda) \mapsto \frac{\text{Tr}_2[h(\lambda)] \text{Tr}[C] - \text{Tr}_2[C] \text{Tr}[h(\lambda)]}{\lambda \text{Tr}[C]^2}.$$

This function is linear and bounded, since λ is bounded away from 0. We now have to prove that this function is indeed the Fréchet derivative. Let $h \in \mathcal{C}(I, \mathcal{S}_1(H, H))$ be a function. Some simple calculations show that

$$\begin{aligned} & \mathbf{A}_1^{(\text{tr})}[\lambda C + h(\lambda)] - \mathbf{A}_1^{(\text{tr})}[\lambda C] - D\mathbf{A}_1^{(\text{tr})}(\lambda C)[h(\lambda)] \\ &= \frac{\text{Tr}_2[h(\lambda)] \text{Tr}[C] - \text{Tr}_2[C] \text{Tr}[h(\lambda)]}{\text{Tr}[C\lambda + h(\lambda)] \text{Tr}[C]} - \frac{\text{Tr}_2[h(\lambda)] \text{Tr}[C] - \text{Tr}_2[C] \text{Tr}[h(\lambda)]}{\lambda \text{Tr}[C]^2} \\ &= \frac{-\text{Tr}_2[h(\lambda)] \text{Tr}[h(\lambda)]}{\lambda \text{Tr}[C\lambda + h(\lambda)] \text{Tr}[C]} + \frac{\text{Tr}[h(\lambda)]^2 \text{Tr}_2[C]}{\lambda \text{Tr}[C\lambda + h(\lambda)] \text{Tr}[C]^2} \end{aligned}$$

Noting that for $\lambda \in I$, both $1/\lambda$ and $1/\|\text{Tr}[C\lambda]\|_1$ are bounded uniformly in λ by a constant and recalling that the trace and the partial traces are Lipschitz continuous with respect to the trace norm, it follows

that the above expression is of the order $\mathcal{O}(\|h\|_\infty^2)$. Hence $D\mathbf{A}_1^{(\text{tr})}(\lambda C)$ indeed is the Fréchet derivative of $\mathbf{A}_1^{(\text{tr})}$ in λC . Summarizing we have that

$$D\mathbf{A}^{(\text{tr})}(\lambda C) : L(\lambda) \mapsto \begin{pmatrix} \text{Tr}_2[L(\lambda)]/\text{Tr}[\lambda C] - \text{Tr}[L(\lambda)]\text{Tr}_2[C]/(\lambda\text{Tr}[C]^2) \\ \text{Tr}_1[L(\lambda)] \\ L(\lambda) \end{pmatrix}.$$

The map $\mathbf{B}^{(\text{tr})}$ is differentiable in $\mathbf{A}^{(\text{tr})}[C\lambda] = (C_1(\lambda), C_2(\lambda), \lambda C)$, with derivative, given as

$$D\mathbf{B}^{(\text{tr})}(C_1(\lambda), C_2(\lambda), \lambda C) : \begin{pmatrix} L_1(\lambda) \\ L_2(\lambda) \\ L_3(\lambda) \end{pmatrix} \mapsto C_1(\lambda) \otimes L_2(\lambda) + L_1(\lambda) \otimes C_2(\lambda) - L_3(\lambda).$$

The proof is similar, but easier than the one for the differential of $\mathbf{B}^{(\text{prod})}$ and hence omitted. By the above arguments $\mathbf{F}^{(\text{tr})}$ is differentiable in λC .

We now turn to the differentiability of the map $\mathbf{F}^{(\text{prod})}$. We can again decompose this map into two simpler ones

$$\mathbf{A}^{(\text{prod})} : \mathcal{C}(I, \mathcal{S}_2(H, H)) \rightarrow \mathcal{C}(I, \mathcal{S}_2(H_1, H_1) \times \mathcal{S}_2(H_2, H_2) \times \mathcal{S}_2(H, H))$$

and

$$\mathbf{B}^{(\text{prod})} : \mathcal{C}(I, \mathcal{S}_2(H_1, H_1) \times \mathcal{S}_2(H_2, H_2) \times \mathcal{S}_2(H, H)) \rightarrow \mathcal{C}(I, \mathcal{S}_2(H, H))$$

pointwise defined as

$$\mathbf{A}^{(\text{prod})} : L(\lambda) \mapsto \begin{pmatrix} P_2(L(\lambda), \Delta_2) \\ P_1(L(\lambda), P_2(L(\lambda), \Delta_2)) \\ L(\lambda) \end{pmatrix}$$

and

$$\mathbf{B}^{(\text{prod})} : \begin{pmatrix} L_1(\lambda) \\ L_2(\lambda) \\ L_3(\lambda) \end{pmatrix} \mapsto \frac{L_1(\lambda) \otimes L_2(\lambda)}{\|L_1(\lambda)\|_2^2} - L_3(\lambda).$$

We begin with the differentiability of $\mathbf{A}^{(\text{prod})}$. As in the previous case one sees that the first and the last component are obviously differentiable due to linearity and boundedness. For the component map $\mathbf{A}_2^{(\text{prod})}$, by bilinearity of the maps P_1, P_2 , the difference $\mathbf{A}_2^{(\text{prod})}[\lambda C + h(\lambda)] - \mathbf{A}_2^{(\text{prod})}[\lambda C]$ equals

$$P_1(\lambda C, P_2(h(\lambda), \Delta_2)) + P_1(h(\lambda), P_2(\lambda C, \Delta_2)) + P_1(h(\lambda), P_2(h(\lambda), \Delta_2)).$$

Due to the bilinearity of the partial products, the first and the second term are linear in $h(\lambda)$. The third term is obviously of order $\mathcal{O}(\|h\|_\infty^2)$. Hence the derivative of $\mathbf{A}_2^{(\text{prod})}$ in λC is pointwise defined as

$$D\mathbf{A}_2^{(\text{prod})}(\lambda C)[h(\lambda)] := P_1(\lambda C, P_2(h(\lambda), \Delta_2)) + P_1(h(\lambda), P_2(\lambda C, \Delta_2)).$$

We claim that the derivative of $\mathbf{B}^{(\text{prod})}$ in any function $T(\lambda) := (T_1(\lambda), T_2(\lambda), T_3(\lambda))$ is given by the bounded linear map

$$D\mathbf{B}^{(\text{prod})}(T(\lambda)) : \begin{pmatrix} L_1(\lambda) \\ L_2(\lambda) \\ L_3(\lambda) \end{pmatrix} \mapsto \frac{L_1(\lambda) \otimes T_2(\lambda) + T_1(\lambda) \otimes L_2(\lambda)}{\|T_1(\lambda)\|_2^2} + \frac{-2\langle L_1(\lambda), T_1(\lambda) \rangle (T_1(\lambda) \otimes T_2(\lambda))}{\|T_1(\lambda)\|_2^4} - L_3(\lambda).$$

We sketch the arguments of the proof. First note that

$$\begin{aligned} & \mathbf{B}^{(\text{prod})}[(L_1(\lambda), L_2(\lambda), L_3(\lambda)) + (h_1(\lambda), h_2(\lambda), h_3(\lambda))] - \mathbf{B}^{(\text{prod})}[(L_1(\lambda), L_2(\lambda), L_3(\lambda))] \\ &= -h_3(\lambda) + \frac{h_1(\lambda) \otimes L_2(\lambda) + L_1(\lambda) \otimes h_2(\lambda)}{\|L_1(\lambda) + h_1(\lambda)\|_2^2} - \frac{2\langle h_1(\lambda), L_1(\lambda) \rangle L_1(\lambda) \otimes L_2(\lambda)}{\|L_1(\lambda) + h_1(\lambda)\|_2^2 \|L_1(\lambda)\|_2^2} + \mathcal{O}(\|h\|_\infty^2). \end{aligned}$$

Subtracting $D\mathbf{B}^{(\text{prod})}(T(\lambda))[(h_1(\lambda), h_2(\lambda), h_3(\lambda))]$ from the above yields a remainder, which is of order $\mathcal{O}(\|h\|_\infty^2)$, as a short calculation reveals. Hence $D\mathbf{B}^{(\text{prod})}(T(\lambda))$ must be the Fréchet differential of $\mathbf{B}^{(\text{prod})}$ in $T(\lambda)$.

Finally we turn to the differentiability of $\mathbf{F}^{(\text{opt})}$. Let $\{u_i\}_{i \geq 1}$ be a basis of the space H_1 and $\{v_j\}_{j \geq 1}$ a basis of H_2 . Then $\{u_i \otimes v_j\}_{i,j \geq 1}$ is a tensor basis of H and $\{u_i \otimes v_j \otimes u_k \otimes v_l\}_{i,j,k,l \geq 1}$ of $H \otimes H$. Recall that $H \otimes H \cong \mathcal{S}_2(H, H)$, i.e. we may identify any operator $L \in \mathcal{S}_2(H, H)$, with the tensor basis expansion

$$L = \sum_{i,j,k,l} \alpha_{i,j,k,l} u_i \otimes v_j \otimes u_k \otimes v_l,$$

where $\alpha_{i,j,k,l} \in \mathbb{R}$. Similarly any sequential operator $L(\lambda)$ in the space $\mathcal{C}(I, \mathcal{S}_2(H, H))$ can be expressed by a tensor expansion, where accordingly the coefficients $\alpha_{i,j,k,l}(\lambda)$ depend on the sequential parameter. We can hence define the map

$$\mathbf{A}^{(\text{opt})} : \begin{cases} \mathcal{C}(I, \mathcal{S}_2(H, H)) \rightarrow \mathcal{C}(I, \mathcal{S}_2(H_1 \otimes H_1, H_2 \otimes H_2)) \\ \sum_{i,j,k,l} \alpha_{i,j,k,l}(\lambda) u_i \otimes v_j \otimes u_k \otimes v_l \mapsto \sum_{i,j,k,l} \alpha_{i,j,k,l}(\lambda) u_i \otimes u_k \otimes v_j \otimes v_l \end{cases}$$

which is bijective, linear and norm preserving (in particular bounded). Notice that $\mathbf{A}^{(\text{opt})}$ is a version of the map Π , introduced in Section 2.2.3 for sequential operators. Secondly, denote by $L(\lambda)^*$ the adjoint and consider the map

$$\mathbf{B}^{(\text{opt})} : \begin{cases} \mathcal{C}(I, \mathcal{S}_2(H_1 \otimes H_1, H_2 \otimes H_2)) \rightarrow \mathcal{C}(I, \mathcal{S}_2^{\text{sym}}(H_1 \otimes H_1, H_1 \otimes H_1) \times \mathcal{S}_2^{\text{sym}}(H_2 \otimes H_2, H_2 \otimes H_2)) \\ L(\lambda) \mapsto (L(\lambda)^* L(\lambda), L(\lambda) L(\lambda)^*) \end{cases}.$$

Here the space $\mathcal{S}_2^{\text{sym}}(H_i \otimes H_i, H_i \otimes H_i)$ is defined as the vector space of all Hilbert–Schmidt operators in $\mathcal{S}_2(H_i \otimes H_i, H_i \otimes H_i)$, which are symmetric. The inner product on this space is induced by that on $\mathcal{S}_2(H_i \otimes H_i, H_i \otimes H_i)$. The map $\mathbf{B}^{(\text{opt})}$ is Fréchet differentiable in the sequential operator $A_0(\lambda) := \mathbf{A}^{(\text{opt})}[\lambda C]$ and its derivative

$$D\mathbf{B}^{(\text{opt})}(A_0(\lambda)) : L(\lambda) \mapsto (L^*(\lambda)L_0(\lambda) + L_0^*(\lambda)L(\lambda), L(\lambda)L_0^*(\lambda) + L_0(\lambda)L^*(\lambda)).$$

Note that the image of $D\mathbf{B}^{(\text{opt})}(A_0(\lambda))$ consists of tuples of pointwise symmetric operators. Recall that for symmetric operators both eigenvectors and eigenvalues are well defined objects. We hence define for $i = 1, 2$ the maps

$$\mathbf{C}_i^{(\text{opt})} : \begin{cases} \mathcal{C}(I, \mathcal{S}_2^{\text{sym}}(H_i \otimes H_i, H_i \otimes H_i)) \rightarrow \mathcal{C}(I, H_i \otimes H_i) \\ L(\lambda) \mapsto v_1^{L(\lambda)}, \end{cases}$$

where $v_1^{L(\lambda)}$ denotes the normalized eigenvector of $L(\lambda)$ corresponding to the largest eigenvalue. Let $B_0(\lambda) := (\mathbf{B}^{(\text{opt})} \circ \mathbf{A}^{(\text{opt})})[\lambda C]$, which has unequal first two eigenvalues for all $\lambda \in I$. This assertion holds

because of Assumption 3.2 (4), which implies $\gamma_1 > \gamma_2$ for the first two singular values of $\Pi[C] = \mathbf{A}^{(\text{opt})}[C]$. \mathbf{C}_i is differentiable in $B_0(\lambda)$ with derivative

$$D\mathbf{C}_i^{(\text{opt})}(B_0(\lambda)) : h(\lambda) \mapsto \sum_{k>1} \frac{\langle h(\lambda), v_1^{B_0(\lambda)} \otimes v_k^{B_0(\lambda)} \rangle v_k^{B_0(\lambda)}}{\gamma_1^{B_0(\lambda)} - \gamma_k^{B_0(\lambda)}},$$

where $\gamma_k^{L(\lambda)}$ denotes the k -th largest eigenvalue of the operator $L(\lambda)$ and $v_k^{L(\lambda)}$ the corresponding eigenvector. Analogously we observe that the map

$$\mathbf{C}_3^{(\text{opt})} : \begin{cases} \mathcal{C}(I, \mathcal{S}_2^{\text{sym}}(H_1 \otimes H_1, H_1 \otimes H_1)) \rightarrow \mathcal{C}(I, \mathbb{R}) \\ L(\lambda) \mapsto \gamma_1^{L(\lambda)}, \end{cases}$$

is differentiable in $B_0(\lambda)$ with derivative

$$D\mathbf{C}_3^{(\text{opt})}(B_0(\lambda)) : h(\lambda) \mapsto \langle h(\lambda), v_1^{B_0(\lambda)} \otimes v_1^{B_0(\lambda)} \rangle.$$

The proof of the differentiability of these maps is because of its technical nature deferred to the end of this proof. By differentiability of each component, the combined map

$$\mathbf{C}^{(\text{opt})} : \begin{cases} \mathcal{C}(I, \mathcal{S}_2^{\text{sym}}(H_1 \otimes H_1, H_1 \otimes H_1) \times \mathcal{S}_2^{\text{sym}}(H_2 \otimes H_2, H_2 \otimes H_2)) \rightarrow \mathcal{C}(I, (H_1 \otimes H_1) \times (H_2 \otimes H_2) \times \mathbb{R}) \\ (L_1(\lambda), L_2(\lambda)) \mapsto (\mathbf{C}_1^{(\text{opt})}(L_1(\lambda)), \mathbf{C}_2^{(\text{opt})}(L_2(\lambda)), \mathbf{C}_3^{(\text{opt})}(L_1(\lambda))) \end{cases}$$

is differentiable. Next, the map

$$\mathbf{D}^{(\text{opt})} : \begin{cases} \mathcal{C}(I, (H_1 \otimes H_1) \times (H_2 \otimes H_2) \times \mathbb{R}) \rightarrow \mathcal{C}(I, \mathcal{S}_2(H_1 \otimes H_1, H_2 \otimes H_2)) \\ (V(\lambda), W(\lambda), r(\lambda)) \mapsto r(\lambda)\lambda[V(\lambda) \otimes W(\lambda)] \end{cases}$$

is Fréchet differentiable in any 3-tuple $(V_0(\lambda), W_0(\lambda), r_0(\lambda))$ with derivative $D\mathbf{D}^{(\text{opt})}(V_0(\lambda), W_0(\lambda), r_0(\lambda))$ given as

$$(V(\lambda), W(\lambda), r(\lambda)) \mapsto \lambda \{r_0(\lambda)[V_0(\lambda) \otimes W(\lambda)] + r(\lambda)[V_0(\lambda) \otimes W_0(\lambda)] + r_0(\lambda)[V(\lambda) \otimes W_0(\lambda)]\},$$

which follows by elementary calculations. Finally $(\mathbf{A}^{(\text{opt})})^{-1}$ is bounded, linear and thus differentiable. We now conclude that the map

$$\mathbf{E}^{(\text{opt})} := (\mathbf{A}^{(\text{opt})})^{-1} \circ \mathbf{D}^{(\text{opt})} \circ \mathbf{C}^{(\text{opt})} \circ \mathbf{B}^{(\text{opt})} \circ \mathbf{A}^{(\text{opt})} : \mathcal{C}(I, \mathcal{S}_2(H, H)) \rightarrow \mathcal{C}(I, \mathcal{S}_2(H, H)),$$

which maps a sequential operator $L(\lambda)$ to its optimal separable approximation $L^{\text{opt}}(\lambda)$ is differentiable in $\lambda \mapsto \lambda C$. Hence it follows immediately that $\mathbf{F}^{(\text{opt})}$ also is.

We conclude this section by a proof of the differentiability of the component maps $\mathbf{C}_1^{(\text{opt})}$, $\mathbf{C}_2^{(\text{opt})}$ and $\mathbf{C}_3^{(\text{opt})}$. We first make an observation, that will be used below. For an operator $A \in \mathcal{C}(I, \mathcal{S}_2^{\text{sym}}(H_i \otimes H_i, H_i \otimes H_i))$, $i = 1, 2$, by the spectral theorem we have the representation

$$A(\lambda) = \sum_{k \geq 1} \gamma_k^{A(\lambda)} v_k^{A(\lambda)} \otimes v_k^{A(\lambda)},$$

where $\gamma_k^{A(\lambda)} \in C(I, \mathbb{R})$, $v_k^{A(\lambda)} \in \mathcal{C}(I, H_i \otimes H_i)$, $i = 1, 2$, are the eigenvalues and respective normalized eigenvectors of $A(\lambda)$. Combing this identity with $\sum_{j \geq 1} \langle v_j^{A(\lambda)}, x \rangle v_j^{A(\lambda)} = x$, for $x \in H_i \otimes H_i$ we see that

$$(\gamma_1^{A(\lambda)} \text{Id} - A(\lambda))[x] = \sum_{i \geq 2} (\gamma_1^{A(\lambda)} - \gamma_i^{A(\lambda)}) \langle v_i^{A(\lambda)}, x \rangle v_i^{A(\lambda)}.$$

We further define the symmetric operator $B \in \mathcal{C}(I, \mathcal{S}_2^{\text{sym}}(H_i \otimes H_i, H_i \otimes H_i))$ as

$$B(\lambda) : y \mapsto \sum_{j \geq 2} \frac{\langle v_j^{A(\lambda)}, y \rangle}{\gamma_1^{A(\lambda)} - \gamma_j^{A(\lambda)}} v_j^{A(\lambda)}.$$

Now a simple calculation shows that $B(\lambda)[(\gamma_1^{A(\lambda)} \text{Id} - A(\lambda))[x]] = x$, for $x \in \text{Ker}(\gamma_1^{A(\lambda)} \text{Id} - A(\lambda))^\perp$. Hence $B(\lambda)$ is the left inverse of $(\gamma_1^{A(\lambda)} \text{Id} - A(\lambda))$ on this space and by abuse of notation we will denote it by $(\gamma_1^{A(\lambda)} \text{Id} - A(\lambda))^{-1}$.

We first consider the eigenvectors and only give the proof for $\mathbf{C}_1^{(\text{opt})}$, as the case of $\mathbf{C}_2^{(\text{opt})}$ is analogous. Let $A, h \in \mathcal{C}(I, \mathcal{S}_2^{\text{sym}}(H_1 \otimes H_1, H_1 \otimes H_1))$ and $\|h\|_\infty \rightarrow 0$, then

$$\begin{aligned} v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)} &= \frac{1}{\gamma_1^{(A+h)(\lambda)}} \left((A+h)(\lambda)[v_1^{(A+h)(\lambda)}] - \gamma_1^{(A+h)(\lambda)} v_1^{A(\lambda)} \right) \\ &= \frac{1}{\gamma_1^{(A+h)(\lambda)}} \left(A(\lambda)[v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)}] + h(\lambda)[v_1^{(A+h)(\lambda)}] + v_1^{A(\lambda)} \left(\gamma_1^{A(\lambda)} - \gamma_1^{(A+h)(\lambda)} \right) \right) \end{aligned}$$

where we used the properties of eigenvalues and eigenvectors. We further simplify by multiplying both sides by $\gamma_1^{(A+h)(\lambda)}$ and rearranging terms, which gives

$$\left(\gamma_1^{(A+h)(\lambda)} \text{Id} - A(\lambda) \right) \left[v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)} \right] = h(\lambda)[v_1^{(A+h)(\lambda)}] + v_1^{A(\lambda)} \left(\gamma_1^{A(\lambda)} - \gamma_1^{(A+h)(\lambda)} \right)$$

Define the function

$$DC_1^{(\text{opt})}(A(\lambda))[h(\lambda)] := (\gamma_1^{A(\lambda)} \text{Id} - A(\lambda))^{-1} \left[h(\lambda)[v_1^{A(\lambda)}] \right]$$

A calculation shows that,

$$\begin{aligned} & \left(v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)} \right) / \|h\|_\infty \\ &= \left(\gamma_1^{(A+h)(\lambda)} \text{Id} - A(\lambda) \right)^{-1} \left[\frac{h(\lambda)}{\|h\|_\infty} \left[v_1^{(A+h)(\lambda)} \right] + \frac{v_1^{A(\lambda)}}{\|h\|_\infty} \left(\gamma_1^{A(\lambda)} - \gamma_1^{(A+h)(\lambda)} \right) \right] \end{aligned}$$

Note from the above observation that $(\gamma_1^{A(\lambda)} \text{Id} - A(\lambda))^{-1}$ is bounded and $(\gamma_1^{A(\lambda)} \text{Id} - A(\lambda))^{-1}[v_1^{A(\lambda)}] = 0$. Using these facts and subtracting $DC_1^{(\text{opt})}(A(\lambda))[h(\lambda)]/\|h\|_\infty$ from the above expression yields a term, which has norm of order

$$\begin{aligned} & \mathcal{O} \left(\left\| \left(\gamma_1^{A+h} \text{Id} - (A+h) \right)^{-1} \right\|_\infty \|v_1^{A+h} - v_1^A\| \right. \\ & \left. + 2 \left\| \left(\gamma_1^{A+h} \text{Id} - (A+h) \right)^{-1} - \left(\gamma_1^A \text{Id} - A \right)^{-1} \right\|_\infty \right), \end{aligned}$$

where we used the inequality $|\gamma_1^A(\lambda) - \gamma_1^{A+h}(\lambda)| \leq \|(A+h)(\lambda) - A(\lambda)\|_2$ for every λ . Now both terms converge to 0, by continuity of the eigenvalues and eigenvectors (see Lemmas 2.2. and 2.3 in Horváth and Kokoszka (2012)).

Next, we derive the Fréchet derivative of the eigenvalue map. Let again h be a symmetric operator, converging to 0. Then, we have the following:

$$\begin{aligned}\gamma_1^{(A+h)(\lambda)} - \gamma_1^{A(\lambda)} &= \langle (A+h)(\lambda)[v_1^{(A+h)(\lambda)}], v_1^{(A+h)(\lambda)} \rangle - \langle A(\lambda)[v_1^{A(\lambda)}], v_1^{A(\lambda)}(\lambda) \rangle \\ &= \langle A(\lambda)[v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)}], v_1^{(A+h)(\lambda)} + v_1^{A(\lambda)} \rangle \\ &\quad + \langle h(\lambda)[v_1^{(A+h)(\lambda)}], v_1^{(A+h)(\lambda)} \rangle\end{aligned}$$

where $v_1^L(\lambda)$ denotes the normalized eigenvector belonging to the largest eigenvalue of the symmetric operator $L(\lambda)$. Using the symmetry of $A(\lambda)$ and subtracting $\langle h(\lambda)[v_1^{A(\lambda)}], v_1^{A(\lambda)} \rangle$ we see the above expression equals

$$\langle v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)}, A(\lambda)[v_1^{(A+h)(\lambda)}] \rangle + \gamma_1^{A(\lambda)} \langle v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)}, v_1^{A(\lambda)} \rangle. \quad (\text{A.9})$$

Notice that according to the parallelogram law for two vectors v, w of unit length the following identity holds:

$$\langle v - w, v \rangle = -\frac{1}{2}\|v - w\|^2.$$

As a consequence the second term in (A.9) equals $-\gamma_1^{A(\lambda)}\|v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)}\|^2/2 = \mathcal{O}(\|h\|_\infty^2)$. For the first term a simple calculation yields

$$\begin{aligned}&\langle v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)}, A(\lambda)[v_1^{(A+h)(\lambda)}] \rangle \\ &= \langle v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)}, A(\lambda)[v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)}] \rangle + \gamma_1^{A(\lambda)} \langle v_1^{(A+h)(\lambda)} - v_1^{A(\lambda)}, v_1^{A(\lambda)} \rangle.\end{aligned}$$

Both terms on the right are of order $\mathcal{O}(\|h\|_\infty^2)$, which can be seen for the first one, by an application of Cauchy-Schwarz and for the second by using the parallelogram law.

A.4 Proof of Corollary 3.5

Let $x \in \{\text{tr}, \text{prod}, \text{opt}\}$. We define

$$H_n^x(\lambda) = M^x[\hat{C}_n(\lambda)] - M^x[\lambda C].$$

In order to prove the weak convergence

$$\{\sqrt{n}H_n(\lambda)\}_{\lambda \in I} \xrightarrow{d} \{\sigma^x \lambda \mathbb{B}(\lambda)\}_{\lambda \in I}, \quad (\text{A.10})$$

we rewrite $H_n^x(\lambda)$ as follows:

$$\begin{aligned}H_n(\lambda) &= \|\hat{C}_n(\lambda) - \hat{C}_n^x(\lambda)\|_2^2 - \lambda^2 \|C - C^x\|_2^2 \\ &= \left\langle \hat{C}_n(\lambda) - \hat{C}_n^x(\lambda) - \lambda C + \lambda C^x, \hat{C}_n(\lambda) - \hat{C}_n^x(\lambda) + \lambda C - \lambda C^x \right\rangle_2 \\ &= \left\langle \mathbf{F}^x[\hat{C}_n(\lambda)] - \mathbf{F}^x[C\lambda], \mathbf{F}^x[\hat{C}_n(\lambda)] + \mathbf{F}^x[C\lambda] \right\rangle_2.\end{aligned}$$

We recall Theorems 3.3 and 3.4. These imply that

$$\mathbf{F}^x[\hat{C}_n(\lambda)] + \mathbf{F}^x[C\lambda] \xrightarrow{\mathbb{P}} 2\lambda \mathbf{F}^x[C]$$

and by application of the functional- Δ -method

$$\sqrt{n} \left(\mathbf{F}^x[\hat{C}_n(\lambda)] - \mathbf{F}^x[C\lambda] \right) \xrightarrow{d} \mathbb{B}_G(\lambda),$$

where \mathbb{B}_G is a Brownian motion corresponding to the Gaussian process G , from Theorem 3.3. We hence observe that the weak convergence

$$\{\sqrt{n}H_n(\lambda)\}_{\lambda \in I} \xrightarrow{d} \{2\lambda \langle \mathbf{F}^x[C], \mathbb{B}_G(\lambda) \rangle_2\}_{\lambda \in I} \stackrel{d}{=} \{\sigma^x \lambda \mathbb{B}(\lambda)\}_{\lambda \in I},$$

holds, where

$$(\sigma^x)^2 := 4\mathbb{E}[\langle \mathbf{F}^x[C], \mathbb{B}_G(\lambda) \rangle_2^2] \quad (\text{A.11})$$

□

A.5 Proof of Theorem 3.8 and Theorem 3.10

Both results are proved by the same arguments and for the sake of brevity we restrict ourselves to the proof of Proof of Theorem 3.8 . According to Corollary 3.5 for $x \in \{\text{tr}, \text{prod}, \text{opt}\}$ it holds that

$$\left\{ \sqrt{N} (M^x[\hat{C}_n(\lambda)] - M^x[\lambda C]) \right\}_{\lambda \in I} \xrightarrow{d} \{\sigma^x \lambda \mathbb{B}(\lambda)\}_{\lambda \in I}.$$

Recalling the definition of \hat{V}_n^x in (3.6) we observe that

$$\begin{aligned} & \left[\int_I (M^x[\hat{C}_n(\lambda)] - \lambda^2 M^x[\hat{C}_n])^2 d\nu(\lambda) \right]^{1/2} \\ &= \left[\int_I (\{M^x[\hat{C}_n(\lambda)] - M^x[\lambda C]\} + \{\lambda^2 M^x[C] - \lambda^2 M^x[\hat{C}_n]\})^2 d\nu(\lambda) \right]^{1/2} \\ & \xrightarrow{d} \sigma^x \left[\int_I \lambda^2 (\mathbb{B}(\lambda) - \lambda \mathbb{B}(1))^2 d\nu(\lambda) \right]^{1/2}. \end{aligned}$$

By the continuous mapping theorem weak convergence of the joint vector follows

$$\sqrt{N} \left(M^x[\hat{C}_n] - M^x[C], \hat{V}_n^x \right) \xrightarrow{d} \sigma^x \left(\mathbb{B}(1), \left[\int_I \lambda^2 (\mathbb{B}(\lambda) - \lambda \mathbb{B}(1))^2 d\nu(\lambda) \right]^{1/2} \right),$$

which entails the weak convergence in (3.7) of the self-normalized statistic and proves the result. □

A.6 Proof of Theorem 3.11

We will restrict ourselves to a proof of the result in (3.7) for the quantity M^x . The corresponding statement for the relative deviation M_{rel}^x in (3.8) follows by the same arguments. Recalling the definition of \hat{I}_n^x in (3.9) we obtain by an elementary calculation that

$$\mathbb{P}\{M^x[C] \in \hat{I}_n^x\} = \mathbb{P}\{q_{\alpha/2} \hat{V}_n^x \leq \hat{M}^x[\hat{C}_n] - M^x[C] \leq q_{1-\alpha/2} \hat{V}_n^x\}.$$

As $\sigma^x > 0$, it follows from Theorem 3.8 that

$$\lim_{n \rightarrow \infty} \mathbb{P}\{M^x[C] \in \hat{I}_n^x\} = \mathbb{P}\{q_{\alpha/2} \leq W \leq q_{1-\alpha/2}\} = 1 - \alpha.$$

□